These bound the error if g^B is estimated from the n^{th} stage of the Bellman algorithm. The lower bound is non-decreasing in n, the upper bound is non-increasing, and both converge to g^B with increasing n.

In comparing the use of the Bellman and Howard algorithms for calculating g^B, it should be noted that the iterations are fairly similar (at least if **A** is chosen to maximize each component of the inequality in step 4 of the Howard algorithm). The difference is that the Howard algorithm calculates a new relative gain vector $\mathbf{w}^B$ at each iteration, whereas the Bellman algorithm uses the previous value of expected aggregate gain in place of the relative gain. The Howard algorithm requires on the order of J^3 steps to find $\mathbf{w}^B$ (i.e., the computation required to solve J simultaneous linear equations) and requires on the order of $J(\sum_i K_i)$ steps for the check in step 2 and the maximization in step 3. The Bellman algorithm requires on the order of $J(\sum_i K_i)$ steps for the entire iteration. Thus if there are many decision alternatives and few states, the computation per iteration of the two algorithms is similar, whereas with many states and few alternatives, the Howard algorithm requires much more computation per iteration. Naturally, this does not help in seeing how many iterations are required with each algorithm, and there seems to be no easy way to answer this question.

Another related question is how the Howard algorithm compares with the brute force method of calculating g^A for every policy **A**. Finding g^A for a policy requires finding the steady state probability vector π^A for $[P^A]$, which is of comparable complexity to finding $\mathbf{w}^B$. The number of different policies is $K_1K_2 \ldots K_J$, which is the number of times that a steady state probability must be calculated in the brute force method. In the Howard algorithm, on the other hand, each iteration yields a better policy than the one before. If one assumes (with little real justification) that the improved policy at each iteration is a random equiprobable choice among all possible improved algorithms, then it turns out (see [Ros83] section 4.6) that the expected number of required iterations is approximately equal to the natural log of the total number of policies.

4.7 SUMMARY

This chapter has developed the basic results about finite state Markov chains from a primarily algebraic standpoint. It was shown that the states of any finite state chain can be partitioned into classes, where each class is either transient or recurrent, and each class is periodic or aperiodic. If the entire chain is one recurrent class, then the Frobenius theorem, with all its corollaries, shows that $\lambda=1$ is an eigenvalue of largest magnitude and has positive right and left eigenvectors, unique within a scale factor. The left eigenvector (scaled to be a probability vector) is the steady state probability vector. If the chain is also aperiodic, then the eigenvalue $\lambda=1$ is the only eigenvalue of magnitude 1, and all rows of $[P]^n$ converge geometrically in n to the steady state vector. This same analysis can be applied to each aperiodic recurrent class of a general Markov chain, given that the chain ever enters that class.

For a periodic recurrent chain of period d, there are d-1 other eigenvalues of magnitude 1, with all d eigenvalues uniformly placed around the unit circle in the complex plane. Exercise 4.13 shows how to interpret these eigenvectors, and shows that $[P]^{nd}$ converges geometrically as $n\to\infty$.

For an arbitrary finite state Markov chain, if the initial state is transient, then the Markov chain will eventually enter a recurrent state, and the probability that this takes more than n steps approaches zero geometrically in n; exercise 4.10 shows how to find the probability that each recurrent class is entered. Given an entry into a particular recurrent class, then the results above can be used to analyze the behavior within that class.

The results about Markov chains were extended to Markov chains with rewards. As with renewal processes, the use of reward functions provides a systematic way to approach a large class of problems ranging from first passage times to dynamic programming. The key result here is theorem 5, which provides both an exact expression and an asymptotic expression for the expected aggregate reward over n stages.

Finally, the results on Markov chains with rewards were used to approach Markov decision theory. We developed the Bellman dynamic programming algorithm, and also investigated the optimal stationary policy. Theorem 9 demonstrated the relationship between the optimal dynamic policy and the optimal stationary policy. This section provided only an introduction to dynamic programming. We omitted all discussion of the relation between optimal stationary and dynamic policies when the stationary chains contain transients and multiple recurrent classes; it appears that these situations are best treated on a case by case basis. Also we omitted discounting (in which future gain is considered worth less than present gain because of interest rates), and we omitted infinite state spaces.

For an introduction to vectors, matrices, and linear algebra, see any introductory text on linear algebra such as Strang [Str88]. Gantmacher [Gan59] has a particularly complete treatment of non-negative matrices and Perron-Frobenius theory. For further reading on Markov decision theory and dynamic programming, see Bertsekas, [Ber87]. Howard, [How60] and Bellman [Bel57] are of historic interest and provide very accessible introductory material.

EXERCISES

4.1) a) Prove that, for a finite state Markov chain, if $P_{ii}>0$ for some i in a recurrent class A, then class A is aperiodic.

b) Show that every finite state Markov chain contains at least one recurrent set of states. Hint: Construct a directed graph in which the states are nodes and an edge goes from i to j if $i\to j$ but i is not accessible from j. Show that this graph contains no cycles, and thus contains one or more nodes with no outgoing edges. Show that each such node is in a recurrent class. Note: this result is not true for Markov chains with countably infinite state spaces.

4.2) A transition probability matrix P is said to be doubly stochastic if

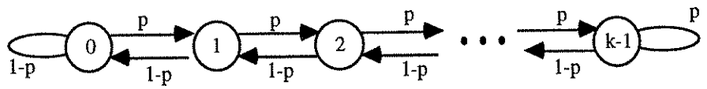
$$\sum_j P_{ij} = 1 \text{ for all } i; \quad \sum_i P_{ij} = 1 \text{ for all } j$$

That is, both the row and the column sums each equal 1. If a doubly stochastic chain has J states and is ergodic (i.e., has a single class of states and is aperiodic), calculate its steady state probabilities.

4.3) a) Find the steady state probabilities $\pi_0, \ldots, \pi_{k-1}$ for the Markov chain below. Express your answer in terms of the ratio $\rho = p/q$. Pay particular attention to the special case $\rho=1$.

b) Sketch $\pi_0, \ldots, \pi_{k-1}$. Give one sketch for $\rho=1/2$, one for $\rho=1$, and one for $\rho=2$.

c) Find the limit of π_0 as k approaches ∞; give separate answers for $\rho<1$, $\rho=1$, and $\rho>1$. Find limiting values of π_{k-1} for the same cases.



4.4) Answer each of the following questions for each of the following non-negative matrices [A]

$$\text{i)} \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \qquad \text{ii)} \begin{bmatrix} 1 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \end{bmatrix}$$

a) Find $[A]^n$ in closed form for arbitrary $n>1$.

b) Find all eigenvalues and all right eigenvectors of [A].

c) Use (b) to show that there is no diagonal matrix [Λ] and no invertible matrix [Q] for which $[A][Q] = [Q][\Lambda]$.

d) Rederive the result of part (c) using the result of (a) rather than (b).

4.5) a) Find the steady state probabilities for each of the Markov chains in figure 4.2 of section 4.1. Assume that all clockwise probabilities in the first graph are the same, say p, and assume that $P_{4,5} = P_{4,1}$ in the second graph.

b) Find the matrices $[P]^2$ for the same chains. Draw the graphs for the Markov chains represented by $[P]^2$, i.e., the graph of two step transitions for the original chains. Find the steady state probabilities for these two step chains. Explain why your steady state probabilities are not unique.

c) Find $\lim_{n\to\infty}[P]^{2n}$ for each of the chains.

4.6) Show that the graph for an ergodic Markov chain of J states must contain at least one cycle with $t \le J-1$ nodes. Show that, for any $J \ge 3$, there is an ergodic Markov chain for which the graph consists of exactly one cycle of length J and one cycle of length J-1. Show that, for this chain, $P_{ij}^n = 0$ for some i, j, and for $n = (J-1)^2$. The point of this problem is to show that theorem 3 is relatively tight.

4.7) a) Show that if x_1 and x_2 are real or complex numbers, then $|x_1+x_2| = |x_1|+|x_2|$ implies that for some β, βx_1, and βx_2 are both real and non-negative.

b) Show from this that if the inequality in (17) is satisfied with equality, then there is some β for which $\beta x_i = |x_i|$ for all i.

4.8) a) Let λ be an eigenvalue of a matrix [A], and let ν and π be right and left eigenvectors respectively of λ, normalized so that $\pi\nu = 1$. Show that

$$[[A] - \lambda\nu\pi]^2 = [A]^2 - \lambda^2\nu\pi.$$

b) Show that $[[A]^n - \lambda^n\nu\pi]\,[[A] - \lambda\nu\pi] = [A]^{n+1} - \lambda^{n+1}\nu\pi$.

c) Use induction to show that $[[A] - \lambda\nu\pi]^n = [A]^n - \lambda^n\nu\pi$.

4.9) Let [P] be the transition matrix for a Markov chain with one recurrent class of states and one or more transient classes. Suppose there are J recurrent states, numbered 1 to J, and K transient states, J+1 to J+K. Thus [P] can be partitioned as $[P] = \begin{bmatrix} P_r & 0 \\ P_{tr} & P_{tt} \end{bmatrix}$.

a) Show that $[P]^n$ can be partitioned as $[P]^n = \begin{bmatrix} [P_r]^n & [0] \\ [P_{ij}^n] & [P_{tt}]^n \end{bmatrix}$. That is, the blocks on the diagonal are simply products of the corresponding blocks of [P], and the lower left block is whatever it turns out to be.

b) Let Q_i be the probability that the chain will be in a recurrent state after K transitions, starting from state i, i.e., $Q_i = \sum_{j\le J} P_{ij}^K$. Show that $Q_i > 0$ for all transient i.

c) Let Q be the minimum Q_i over all transient i and show that $P_{ij}^{nK} \le (1-Q)^n$ for all transient i,j (i.e., show that $[P_{tt}]^n$ approaches the all zero matrix [0] with increasing n).

d) Let $\pi = (\pi_r, \pi_t)$ be a left eigenvector of [P] of eigenvalue 1 (if one exists). Show that $\pi_t = \mathbf{0}$ and show that π_r must be positive and be a left eigenvector of $[P_r]$. Thus show that π exists and is unique (within a scale factor).

e) Show that **e** is the unique right eigenvector of [P] of eigenvalue 1 (within a scale factor).

4.10) Generalize exercise 4.9 to the case of a Markov chain [P] with r recurrent classes and one or more transient classes. In particular,

a) Show that [P] has exactly r linearly independent left eigenvectors, $\pi^{(1)}, \pi^{(2)}, \ldots, \pi^{(r)}$ of eigenvalue 1, and that the i^{th} can be taken as a probability vector that is positive on the i^{th} recurrent class and zero elsewhere.

b) Show that [P] has exactly r linearly independent right eigenvectors, $\nu^{(1)}, \nu^{(2)}, \ldots, \nu^{(r)}$ of eigenvalue 1, and that the i^{th} can be taken as a vector with $\nu_j^{(i)}$ equal to the probability that recurrent class i will ever be entered starting from state j.

4.11) Prove theorem 6A. Hint: Use theorem 6 and the results of exercise 4.9.

4.12) Generalize exercise 4.11 to the case of a Markov chain [P] with r aperiodic recurrent classes and one or more transient classes. In particular, using the left and right eigenvectors $\pi^{(1)}, \pi^{(2)}, \ldots, \pi^{(r)}$ and $\nu^{(1)},\ldots, \nu^{(r)}$ of exercise 4.10, show that

$\lim_{n\to\infty}[P]^n = \sum_i \nu^{(i)}\pi^{(i)}$.

4.13) Suppose a Markov chain with matrix [P] is irreducible and periodic with period d and let T_i, $1\le i\le d$, be the i^{th} subclass in the sense of theorem 2. Assume the states are numbered so that the first J_1 states are in T_1, the next J_2 are in T_2, and so forth. Thus [P] has the block form given by

$$[P] = \begin{bmatrix} 0 & [P_1] & \ddots & 0 \\ 0 & 0 & [P_2] & \ddots & \ddots \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & \ddots & \ddots & [P_{d-1}] \\ [P_d] & 0 & \ddots & \ddots & 0 \end{bmatrix}$$

where $[P_i]$ has dimension J_i by $J_{(i+1)}$ for i<d and J_d by J_1 for i=d.

a) Show that $[P]^d$ has the form

$$[P]^d = \begin{bmatrix} [Q_1] & 0 & \ddots & 0 \\ 0 & [Q_2] & \ddots & \ddots \\ \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & \ddots & [Q_d] \end{bmatrix}$$

where $[Q_i] = [P_i][P_{i+1}]\ldots[P_d][P_1]\ldots[P_{i-1}]$.

b) Show that $[Q_i]$ is the matrix of an ergodic Markov chain, so that with the eigenvectors defined in exercises 4.10 and 4.12, $\lim_{n\to\infty}[P]^{nd} = \sum_i \nu^{(i)}\pi^{(i)}$.

c) Show that $\hat{\pi}^{(i)}$, the left eigenvector of $[Q_i]$ of eigenvalue 1 satisfies $\hat{\pi}^{(i)}[P_i] = \hat{\pi}^{(i+1)}[P_{i+1}]$ for i<d and $\hat{\pi}^{(d)}[P_d] = \hat{\pi}^{(1)}[P_1]$.

d) Let $\alpha = \frac{2\pi\sqrt{-1}}{d}$ and let $\pi^{(k)} = (\hat{\pi}^{(1)}, \hat{\pi}^{(2)}e^{\alpha k}, \hat{\pi}^{(3)}e^{2\alpha k}, \ldots, \hat{\pi}^{(d)}e^{(d-1)\alpha k})$. Show that $\pi^{(k)}$ is a left eigenvector of [P] of eigenvalue $e^{-\alpha k}$.

4.14) Assume a friend has developed an excellent program for finding the steady state probabilities for finite state Markov chains. More precisely, given the transition matrix [P], the program returns $\lim_{n\to\infty} P_{ii}^n$ for each i. Assume all chains are aperiodic.

a) You want to find the expected time to first reach a given state k starting from a given state m for a Markov chain with transition matrix [P]. You modify the matrix to [P'] where $P'_{km} = 1$, $P'_{kj} = 0$ for $j\ne m$, and $P'_{ij} = P_{ij}$ otherwise. How do you find the desired first passage time from the program output given [P'] as an input? Hint: The times at which a Markov chain enters any given state can be considered as renewals in a (perhaps delayed) renewal process.

b) Using the same [P'] as the program input, how can you find the expected number of returns to state m before the first passage to state k?

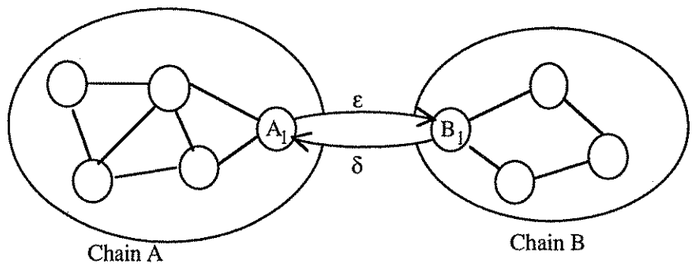
c) Suppose, for the same Markov chain [P] and the same starting state m, you want to find the probability of reaching some given state n before the first passage to k. Modify [P] to some [P''] so that the above program with P'' as an input allows you to easily find the desired probability.

d) Let $P(X(0)=i) = Q_i$, $1\le i\le J$ be an arbitrary set of initial probabilities for the same Markov chain [P] as above. Show how to modify [P] to some [P'''] for which the steady state probabilities allow you to easily find the expected time of the first passage to state k.

4.15) Suppose A and B are each ergodic Markov chains with transition probabilities $\{P_{A_i,A_j}\}$ and $P_{B_i,B_j}\}$ respectively. Denote the steady state probabilities of A and B by $\{\pi_{A_i}\}$ and $\{\pi_{B_i}\}$ respectively. The chains are now connected and modified as shown below. In particular, states A_1 and B_1 are now connected and the new transition probabilities P' for the combined chain are given by

$$P'_{A_1,B_1} = \varepsilon, \quad P'_{A_1,A_j} = (1-\varepsilon)P_{A_1,A_j} \text{ for all } A_j$$
$$P'_{B_1,A_1} = \delta, \quad P'_{B_1,B_j} = (1-\delta)P_{B_1,B_j} \text{ for all } B_j$$

All other transition probabilities remain the same. Think intuitively of ε and δ as being small, but do not make any approximations in what follows. Give your answers to the following questions as functions of ε, δ, $\{\pi_{A_i}\}$ and $\{\pi_{B_i}\}$.



a) Assume that $\varepsilon>0$, $\delta=0$ (i.e., that A is a set of transient states in the combined chain). Starting in state A_1, find the conditional expected time to return to A_1 given that the first transition is to some state in chain A.

b) Assume that $\varepsilon>0$, $\delta=0$. Find $T_{A,B}$, the expected time to first reach state B_1 starting from state A_1. Your answer should be a function of ε and the *original* steady state probabilities $\{\pi_{A_i}\}$ in chain A.

c) Assume $\varepsilon>0$, $\delta>0$. find $T_{B,A}$, the expected time to first reach state A_1, starting in state B_1. Your answer should depend only on δ and $\{\pi_{B_j}\}$.

d) Assume $\varepsilon>0$ and $\delta>0$. Find $P'(A)$, the steady state probability that the combined chain is in one of the states $\{A_j\}$ of the original chain A.

e) Assume $\varepsilon>0$, $\delta=0$. For each state $A_j \neq A_1$ in A, find v_{A_j}, the expected number of visits to state A_j, starting in state A_1, before reaching state B_1. Your answer should depend only on ε and $\{\pi_{A_j}\}$.

f) Assume $\varepsilon>0$, $\delta>0$. For each state A_j in A, find π'_{A_j}, the steady state probability of being in state A_j in the combined chain. Hint: Be careful in your treatment of state A_1.

4.16) For the Markov chain with rewards in figure 4.5:

a) Find the steady state reward per stage, g, using the steady state probability vector π.

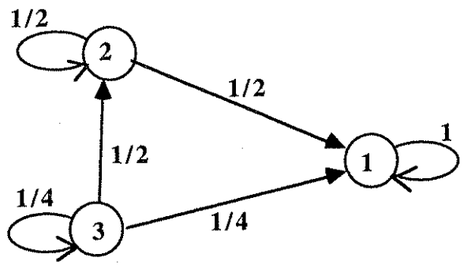
b) Let $w_1 = 0$ and use (34) to find w_2.

c) Assume g in (34) is an unknown. Again, let $w_1 = 0$ and solve (34) for w_2 and g.

d) Now let w_1 be an arbitrary real number y. Again use (34) to solve for w_2 and g. How does your value of g compare to the values found in (a) and (c). Explain your answer.

e) Let $w_1=0$, but take P_{12} as an arbitrary probability. Find g and w_2 again and give an intuitive explanation of why P_{12} effects the asymptotic relative gain of state 2.

4.17) Consider the Markov chain below:



a) Suppose the chain is started in state i and goes through n transitions; let $v_i(n)$ be the expected number of transitions (out of the total of n) until the chain enters the trapping state, state 1. Find an expression for $\mathbf{v}(n) = (v_1(n), v_2(n), v_3(n))$ in terms of $\mathbf{v}(n-1)$ (take $v_1(n) = 0$ for all n). Hint: view the system as a Markov reward system; what is the value of $\mathbf{r}$?

b) Solve numerically for $\lim_{n\to\infty} \mathbf{v}(n)$. Interpret the meaning of the elements v_i in the solution of Eq. (25).

c) Give a direct argument why (25) provides the solution directly to the expected time from each state to enter the trapping state.

4.18) Prove lemma 4; for the stationary policy result, use either induction on (29) or use (31) directly. For the dynamic policy, use induction on (41).

4.19) George drives his car to the theater, which is at the end of a one-way street. There are parking places along the side of the street and a parking garage that costs \$5 at the theater. Each parking place is independently occupied or unoccupied with probability 1/2. If George parks n parking places away from the theater, it costs him n cents (in time and shoe leather) to walk the rest of the way. George is myopic and can only see the parking place he is currently passing.

If George has not already parked by the time he reaches the n^{th} place, he first decides whether or not he will park if the place is unoccupied, and then observes the place and acts according to his decision. George can never go back and must park in the parking garage if he has not parked before.

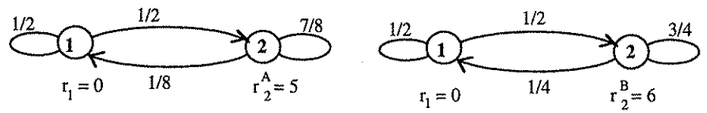
a) Model the above problem as a 2 state Markov decision problem. In the "driving" state, state 2, there are two possible decisions: park if the current place is unoccupied or drive on whether or not the current place is unoccupied.

b) Find $v^*_i(n)$, the *minimum* expected aggregate *cost* for n stages (i.e., immediately before observation of the n^{th} parking place) starting in state i = 1 or 2; it is sufficient to express $v^*_i(n)$ in times of $v^*_i(n-1)$. The final costs, in cents, at stage 0 should be $v_2(0) = 500$, $v_1(0) = 0$.

c) For what values of n is the optimal decision the decision to drive on?

d) What is the probability that George will park in the garage, assuming that he follows the optimal policy?

4.20) Consider a dynamic programming problem with two states and two possible policies, denoted **A** and **B**, in state 2; there is no choice of policies in state 1;



a) Find the steady state gain per stage, g^A and g^B, for stationary policies **A** and **B**.

b) Find the relative gain vectors, $\mathbf{w}^A$ and $\mathbf{w}^B$, for stationary policies **A** and **B**.

c) Suppose the final reward, at stage 0, is $v_1(0) = 0$, $v_2(0) = v$. For what range of v does the dynamic programming algorithm use decision A in state 2 at stage 1?

d) For what range of v does the dynamic programming algorithm use decision A in state 2 at stage 2? at stage n?

e) Find the optimal gain $v^*_2(n)$ and $v^*_1(n)$ as a function of stage n assuming v = 10.