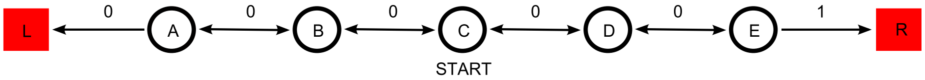
Applying LSTD on Example 6.2 of Sutton and Barto (1998): probability and linear algebra perspective.

Parts of these notes and most definitions and theorems are borrowed from Chapter 11 of Grinstead and Snell (1997)

1 Environment



All episodes start in the center state, C, and proceed either left or right by one state on each step, with equal probability. Episodes terminate either on the extreme left or the extreme right. When an episode terminates on the right a reward of +1 occurs; all other rewards are zero.

1.1 Specifying a Markov Chain

We have a set of *states*, $S = \{s_1, s_2, \ldots, s_r\}$. The process starts in one of these states and moves successively from one state to another. If the chain is currently in state s_i, then it moves to state s_j at the next step with a probability denoted by p_{ij}, and this probability does not depend upon which states the chain was in before the current state. For the example above $S = \{L, A, B, C, D, E, R\}$

1.2 Transition Matrix

From the above information we determine the transition probabilities. These are most conveniently represented in a square array as

$$\mathbf{P} = \begin{array}{c|ccccccc} & L & A & B & C & D & E & R \\ \hline L & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & .5 & 0 & .5 & 0 & 0 & 0 & 0 \\ B & 0 & .5 & 0 & .5 & 0 & 0 & 0 \\ C & 0 & 0 & .5 & 0 & .5 & 0 & 0 \\ D & 0 & 0 & 0 & .5 & 0 & .5 & 0 \\ E & 0 & 0 & 0 & 0 & .5 & 0 & .5 \\ R & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

The entries in the second row of the matrix $\mathbf{P}$ represent the probabilities for being in various states following being in a state A.

We consider the question of determining the probability that, given the chain is in state i today, it will be in state j two steps from now. We denote this probability by $p_{ij}^{(2)}$. In the example 6.2 we see that if you start in state A then the event that we end up in state C in two steps is the disjoint union of seven events: 1) we transition into state L in the first step and state C in the second step, 2) we stay in state A during the first step and transition into state C during the second step, 3) we transition into state B during first state and transition from B into C during second step, ... For example the conditional probability that we transition into state B during the first step given that we start in state A times the conditional probability that we will transition into state C during the second step given that we will start in state B. Using the transition matrix $\mathbf{P}$ is a product $p_{23}p_{34}$. Thus, we have

$$p_{24}^{(2)} = p_{21}p_{14} + p_{22}p_{24} + p_{23}p_{34} \cdots + p_{27}p_{74} = .25$$

In general, if a Markov chain has r states, then

$$p_{ij}^{(2)} = \sum_{k=1}^{r} p_{ik}p_{kj}$$

Theorem 1. *Let* $\mathbf{P}$ *be the transition matrix of a Markov chain. The* ij*th entry* $p_{ij}^{(n)}$ *of the matrix* $\mathbf{P}^n$ *gives the probability that the Markov chain, starting in state* s_i*, will be in state* s_j *after* n *steps.*

We now consider the long-term behavior of a Markov chain when it starts in a state chosen by a probability distribution on the set of states. If $\mathbf{u}$ is a probability vector which represent the initial state of a Markov chain, then we think of ith component of $\mathbf{u}$ as representing the probability that the chain starts in state s_i.

Theorem 2. *Let* $\mathbf{P}$ *be the transition matrix of a Markov chain and let* $\mathbf{u}$ *be the probability vector which represents the starting distribution. Then the probability that the chain is in state* s_i *after* n *steps is the* i*th entry in the vector*

$$\mathbf{u}^{(n)} = \mathbf{u}\mathbf{P}^n$$

Note that if we want to examine the behavior of the chain under the assumption that it starts in a certain state s_i, we simply choose $\mathbf{u}$ to be the probability vector with ith entry equal to 1 and all other entries equal to 0.

1.3 Absorbing Markov Chains

One special type of Markov chains are *absorbing Markov chains.*

Definition 3. A state s_i of a Markov chain is called *absorbing* if it is impossible to leave it (i.e. $p_{ii} = 1$). A Markov chain is *absorbing* if it has at least one absorbing state, and from every state it is possible to go to an absorbing state (not necessarily in one step).

In our example states L and R are absorbing.

Definition 4. In an absorbing Markov chain, a state which is not absorbing is called *transient.*

In our example the states *A, B, C, D,* and E are transient states, and from any of these it is possible to reach the absorbing states L and R.

Questions of interest: *What is the probability that the process will eventually reach an absorbing state? What is the probability that the process will end up in a given absorbing state? On the average, how many times will the process be in each transient state?*

1.3.1 Canonical Form

Consider an arbitrary absorbing Markov chain. Renumber the states so that the transient states come first. If there are r absorbing states and t transient states, the transition matrix will have the following *canonical form*

$$\mathbf{P} = \begin{array}{cc} & \begin{array}{cc} \mathit{TR.} & \mathit{ABS.} \end{array} \\ \begin{array}{c} \mathit{TR.} \\ \mathit{ABS.} \end{array} & \begin{pmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \end{array}$$

Here $\mathbf{I}$ is an r-by-r identity matrix, $\mathbf{0}$ is an r-by-t zero matrix, $\mathbf{R}$ is a nonzero t-by-r matrix, and $\mathbf{Q}$ is an t-by-t matrix. The first t states are transient and the last r states are absorbing.

In 1.2, we saw that the entry $p_{ij}^{(n)}$ of the matrix $\mathbf{P^n}$ is the probability of being in state s_j after n steps, when the chain is started in state s_i. A standard matrix algebra argument shows that $\mathbf{P}^n$ is of the form

$$\mathbf{P}^n = \begin{array}{cc} & \begin{array}{cc} \mathit{TR.} & \mathit{ABS.} \end{array} \\ \begin{array}{c} \mathit{TR.} \\ \mathit{ABS.} \end{array} & \begin{pmatrix} \mathbf{Q^n} & * \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \end{array}$$

where the asterisk $*$ stands for the t-by-r matrix in the upper right-hand corner of $\mathbf{P}^n$. (This sub-matrix can be written in terms of $\mathbf{Q}$ and $\mathbf{R}$, but the expression is complicated and is not needed at this time.) The form of $\mathbf{P}^n$ shows that the entries of $\mathbf{Q}^n$ give the probabilities for being in each of the transient states after n steps for each possible transient starting state.

1.3.2 Probability of Absorption

Theorem 5. *In an absorbing Markov chain, the probability that the process will be absorbed is 1 (i.e.,* $\mathbf{Q}^n \to \mathbf{0}$ *as* $n \to \infty$*).*

Proof. Proof relies on the fact that from each non absorbing state it is possible to reach an absorbing state in finitely many steps, i.e. there exists m such that starting from s_j the probability that the process will not reach an absorbing state is $p_j < 1$ hence as the number of steps increases probabilities tend to 0. □

1.3.3 The Fundamental Matrix

Theorem 6. *For an absorbing Markov chain the matrix* $\mathbf{I} - \mathbf{Q}$ *has an inverse* $\mathbf{N}$ *and* $\mathbf{N} = \mathbf{I} + \mathbf{Q} + \mathbf{Q^2} + \dots$ *. The ij-entry* n_{ij} *of the matrix* $\mathbf{N}$ *is the expected number of times the chain is in state* s_j*, given that it starts in state* s_i*. The initial sate is counted if* $i = j$*.*

Proof. Let $(\mathbf{I}-\mathbf{Q})\mathbf{x}=\mathbf{0}$; that is $\mathbf{x}=\mathbf{Q}\mathbf{x}$. Then, iterating this we see that $\mathbf{x}=\mathbf{Q^n}\mathbf{x}$. Since $\mathbf{Q^n}\to\mathbf{0}$, we have $\mathbf{x}\mathbf{Q^n}\to\mathbf{0}$, so $\mathbf{x}=\mathbf{0}$. Thus $(\mathbf{I}-\mathbf{Q})^{\mathbf{-1}}=\mathbf{N}$ exists. Note that

$$(\mathbf{I}-\mathbf{Q})(\mathbf{I}+\mathbf{Q}+\mathbf{Q^2}+\cdots+\mathbf{Q^n})=\mathbf{I}-\mathbf{Q^{n+1}}$$

Multiplying both sides by $\mathbf{N}$ gives

$$(\mathbf{I}+\mathbf{Q}+\mathbf{Q^2}+\cdots+\mathbf{Q^n})=\mathbf{N}(\mathbf{I}-\mathbf{Q^{n+1}})$$

Letting n tend to infinity we have

$$\mathbf{N}=\mathbf{I}+\mathbf{Q}+\mathbf{Q^2}+\dots$$

Let $X^{(k)}$ be a random variable which equals 1 if the chain is in state s_j after k steps when starting from state s_i, and equals 0 otherwise. Then we have

$$P(X^{(k)}=1)=q_{ij}^{(k)}$$

and

$$P(X^{(k)}=0)=1-q_{ij}^{(k)}$$

The expected number of times the chain is in state s_j in the first n steps, given that it starts in state s_i is

$$E(X^{(0)}+X^{(1)}+\cdots+X^{(n)})=q_{ij}^{(0)}+q_{ij}^{(1)}+\cdots+q_{ij}^{(n)}$$

Letting n tend to infinity we have

$$E(X^{(0)}+X^{(1)}+\cdots+X^{(n)})=q_{ij}^{(0)}+q_{ij}^{(1)}+\cdots=n_{ij}$$

□

Definition 7. For an absorbing Markov chain $\mathbf{P}$, the matrix $\mathbf{N}=(\mathbf{I}-\mathbf{Q})^{\mathbf{-1}}$ is called the *fundamental matrix* for $\mathbf{P}$. The entry n_{ij} of $\mathbf{N}$ gives the expected number of times that the process is in the transient state s_j if it is started in the transient state s_i

Continuing without our example the transition matrix in canonical form is

$$\mathbf{P}=\begin{array}{c|ccccccc} & E & A & B & C & D & L & R \\ \hline E & 0 & 0 & 0 & 0 & .5 & 0 & .5 \\ A & 0 & 0 & .5 & 0 & 0 & .5 & 0 \\ B & 0 & .5 & 0 & .5 & 0 & 0 & 0 \\ C & 0 & 0 & .5 & 0 & .5 & 0 & 0 \\ D & .5 & 0 & 0 & .5 & 0 & 0 & 0 \\ L & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ R & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

From this we see that the matrix $\mathbf{Q}$ is

$$\mathbf{Q} = \begin{array}{cc} & \begin{array}{ccccc} E & A & B & C & D \end{array} \\ \begin{array}{c} E \\ A \\ B \\ C \\ D \end{array} & \begin{pmatrix} 0 & 0 & 0 & 0 & .5 \\ 0 & 0 & .5 & 0 & 0 \\ 0 & .5 & 0 & .5 & 0 \\ 0 & 0 & .5 & 0 & .5 \\ .5 & 0 & 0 & .5 & 0 \end{pmatrix} \end{array}$$

and

$$\mathbf{I} - \mathbf{Q} = \begin{array}{cc} & \begin{array}{ccccc} E & A & B & C & D \end{array} \\ \begin{array}{c} E \\ A \\ B \\ C \\ D \end{array} & \begin{pmatrix} 1 & 0 & 0 & 0 & -.5 \\ 0 & 1 & -.5 & 0 & 0 \\ 0 & -.5 & 1 & -.5 & 0 \\ 0 & 0 & -.5 & 1 & -.5 \\ -.5 & 0 & 0 & -.5 & 1 \end{pmatrix} \end{array}$$

Computing $(\mathbf{I} - \mathbf{Q})^{\mathbf{-1}}$, we find

$$\mathbf{N} = \begin{array}{cc} & \begin{array}{ccccc} E & A & B & C & D \end{array} \\ \begin{array}{c} E \\ A \\ B \\ C \\ D \end{array} & \begin{pmatrix} 1\frac{2}{3} & \frac{1}{3} & \frac{2}{3} & 1 & 1\frac{1}{3} \\ \frac{1}{3} & 1\frac{2}{3} & 1\frac{1}{3} & 1 & \frac{2}{3} \\ \frac{2}{3} & 1\frac{1}{3} & 2\frac{2}{3} & 2 & 1\frac{1}{3} \\ 1 & 1 & 2 & 3 & 2 \\ 1\frac{1}{3} & \frac{2}{3} & 1\frac{1}{3} & 2 & 2\frac{2}{3} \end{pmatrix} \end{array}$$

From the forth row of $\mathbf{N}$ we see that if we start in state C then the expected number of visits to states E, A, B, and D are 1, 1, 2, and 2.

1.3.4 Time to Absorption

We now consider the question: Given that the chain starts in state s_i, what is the expected number of steps before the chain is absorbed?

Theorem 8. *Let t_i be the expected number of steps before the chain is absorbed, given that the chain starts in state s_i, and let $\mathbf{t}$ be the column vector whose ith entry is t_i. Then*

$$\mathbf{t} = \mathbf{Nc}$$

where $\mathbf{c}$ is a column vector whose entries are 1.

Proof. If we add all the entries in the ith row of $\mathbf{N}$, we will have the expected number of times in any of the transient states for a given starting state s_i, that is, the expected time required before being absorbed. Thus t_i, is the sum of the entries in the ith row of N. If we write this statement in matrix form, we obtain the theorem □

1.3.5 Absorption Probabilities

Theorem 9. *Let b_{ij} be the probability that an absorbing chain will be absorbed in the absorbing state s_j if it starts in the transient state s_i. Let $\mathbf{B}$ be the matrix with entries b_{ij}. Then $\mathbf{B}$ is a t-by-r matrix, and*

$$\mathbf{B} = \mathbf{NR}$$

where $\mathbf{N}$ is the fundamental matrix and $\mathbf{R}$ is as in the canonical form.

Proof. We have

$$\mathbf{B_{ij}} = \sum_n \sum_k q_{ik}^{(n)} r_{kj} = \sum_k \sum_n q_{ik}^{(n)} r_{kj} = \sum_k n_{ik} r_{kj} = (\mathbf{NR})_{ij}$$

In the example 6.2 we found that □

$$\mathbf{N} = \begin{pmatrix} 1\frac{2}{3} & \frac{1}{3} & \frac{2}{3} & 1 & 1\frac{1}{3} \\ \frac{1}{3} & 1\frac{2}{3} & 1\frac{1}{3} & 1 & \frac{2}{3} \\ \frac{2}{3} & 1\frac{1}{3} & 2\frac{2}{3} & 2 & 1\frac{1}{3} \\ 1 & 1 & 2 & 3 & 2 \\ 1\frac{1}{3} & \frac{2}{3} & 1\frac{1}{3} & 2 & 2\frac{2}{3} \end{pmatrix}$$

Hence,

$$\mathbf{t} = \mathbf{Nc} = \begin{pmatrix} 1\frac{2}{3} & \frac{1}{3} & \frac{2}{3} & 1 & 1\frac{1}{3} \\ \frac{1}{3} & 1\frac{2}{3} & 1\frac{1}{3} & 1 & \frac{2}{3} \\ \frac{2}{3} & 1\frac{1}{3} & 2\frac{2}{3} & 2 & 1\frac{1}{3} \\ 1 & 1 & 2 & 3 & 2 \\ 1\frac{1}{3} & \frac{2}{3} & 1\frac{1}{3} & 2 & 2\frac{2}{3} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 5 \\ 5 \\ 8 \\ 9 \\ 8 \end{pmatrix}$$

Thus starting in states E, A, B, C, and D the expected times to absorption are 5, 5, 8, 9, and 8, respectively

From the canonical form,

$$\mathbf{R} = \begin{array}{c} \\ E \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cc} L & R \end{array} \\ \begin{pmatrix} 0 & .5 \\ .5 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \end{array}$$

Hence,

$$\mathbf{B} = \mathbf{NR} = \begin{pmatrix} 1\frac{2}{3} & \frac{1}{3} & \frac{2}{3} & 1 & 1\frac{1}{3} \\ \frac{1}{3} & 1\frac{2}{3} & 1\frac{1}{3} & 1 & \frac{2}{3} \\ \frac{2}{3} & 1\frac{1}{3} & 2\frac{2}{3} & 2 & 1\frac{1}{3} \\ 1 & 1 & 2 & 3 & 2 \\ 1\frac{1}{3} & \frac{2}{3} & 1\frac{1}{3} & 2 & 2\frac{2}{3} \end{pmatrix} \begin{pmatrix} 0 & .5 \\ .5 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} = \begin{array}{c} \\ E \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cc} L & R \end{array} \\ \begin{pmatrix} .167 & .833 \\ .833 & .167 \\ .667 & .333 \\ .500 & .500 \\ .333 & .667 \end{pmatrix} \end{array}$$

Notice that matrix **B** can be though of as expected number of times states L and R are visited when starting from the other states.

2 Rewards

We are now ready to introduce a reward structure on top of the Markov chain. In particular example 6.2 specifies that when an episode terminates on the right a reward of +1 occurs. Let **U** denote rewards for transitioning into a state:

$$\mathbf{U} = \begin{array}{c} E \\ A \\ B \\ C \\ D \\ L \\ R \end{array} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

Also, because episodes terminate upon reaching L or R we can rewrite **P** without affecting any prior derivations:

$$\mathbf{P} = \begin{array}{c|ccccccc} & E & A & B & C & D & L & R \\ \hline E & 0 & 0 & 0 & 0 & .5 & 0 & .5 \\ A & 0 & 0 & .5 & 0 & 0 & .5 & 0 \\ B & 0 & .5 & 0 & .5 & 0 & 0 & 0 \\ C & 0 & 0 & .5 & 0 & .5 & 0 & 0 \\ D & .5 & 0 & 0 & .5 & 0 & 0 & 0 \\ L & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ R & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}$$

From **B** and **N** we can construct a matrix that represents expected number of times each state is visited when starting in any state:

$$\mathbf{T} = \begin{pmatrix} \mathbf{N} & \mathbf{B} \\ \mathbf{0} & \mathbf{I_0} \end{pmatrix}$$

where $\mathbf{I_0}$ is the identity matrix of size equal to the number of absorbing states (i.e. you only count visiting absorbing state once and then episode is terminated). Then value of each state (which is the expected return from all returns) can be computed as follows:

$$\mathbf{V} = \mathbf{TU}$$

which for our example is

$$\mathbf{V} = \begin{array}{c} E \\ A \\ B \\ C \\ D \\ L \\ R \end{array} \begin{pmatrix} .833 \\ .167 \\ .333 \\ .500 \\ .667 \\ 0 \\ 1.00 \end{pmatrix}$$

If any rewards were to be assigned in transient states **V** would would be adjusted appropriately as long as there is no discounting. For example, suppose that additionally a reward of 1 is received when visiting state C. Then

$$\mathbf{U_1} = \begin{array}{c} E \\ A \\ B \\ C \\ D \\ L \\ R \end{array} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

and

$$\mathbf{V_1} = \begin{array}{c} E \\ A \\ B \\ C \\ D \\ L \\ R \end{array} \begin{pmatrix} 1.833 \\ 1.167 \\ 2.333 \\ 3.500 \\ 2.667 \\ 0 \\ 1.000 \end{pmatrix}$$

which is exactly **V** plus how many time you expect to visit state C starting from a given state.

3 LSTD as model-based learning

If we know the model of the environment, that is we know true rewards **U** for each state and transition probabilities **P** then we can figure out values of each state by solving a system of Bellman equations as in equation (3) of Boyan (2002), as follows:

$$\beta = (\mathbf{I} - \mathbf{P})^{\mathbf{-1}}\mathbf{U}$$

Let us examine $(\mathbf{I} - \mathbf{P})^{\mathbf{-1}}$ in more detail:

$$\mathbf{P} = \quad \begin{pmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \Longrightarrow$$

$$(\mathbf{I} - \mathbf{P})^{\mathbf{-1}} = \begin{bmatrix} (\mathbf{I_Q} - \mathbf{Q}) & -\mathbf{R} \\ \mathbf{0} & \mathbf{I_0} \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{I_Q} - \mathbf{Q})^{-1} & -(\mathbf{I_Q} - \mathbf{Q})^{\mathbf{-1}}(-\mathbf{R})\mathbf{I_0^{-1}} \\ \mathbf{0} & \mathbf{I_0^{-1}} \end{bmatrix}$$

where $\mathbf{I_Q}$ is an identity matrix of the same size as $\mathbf{Q}$. From derivations above we have $(\mathbf{I_Q} - \mathbf{Q})^{\mathbf{-1}} = \mathbf{N}$ and $\mathbf{NR} = \mathbf{B}$. And since that $\mathbf{I_0^{-1}} = \mathbf{I_0}$ we have

$$(\mathbf{I} - \mathbf{P})^{\mathbf{-1}} = \begin{bmatrix} \mathbf{N} & \mathbf{B} \\ \mathbf{0} & \mathbf{I_0} \end{bmatrix} = \mathbf{T}$$

That is by computing $(\mathbf{I} - \mathbf{P})^{\mathbf{-1}}$ we are implicitly computing expected number of visits to each transient state and expected probabilities of being absorbed in each absorbing state. Thus

$$\beta = (\mathbf{I} - \mathbf{P})^{\mathbf{-1}}\mathbf{U} = \mathbf{TU} = \mathbf{V} \quad !$$

References

Boyan, J. A. (2002): “Technical update: Least-squares temporal difference learning,” *MACHINE LEARNING*, 49, 233—246.

Grinstead, C. M., and J. L. Snell (1997): *Introduction to Probability*. American Mathematical Society, 2 revised edn.

Sutton, R. S., and A. G. Barto (1998): *Reinforcement Learning: An Introduction*. The MIT Press.