MARKOV CHAINS AND MARKOV DECISION THEORY

ARINDRIMA DATTA

Abstract. In this paper, we begin with a formal introduction to probability and explain the concept of random variables and stochastic processes. After this, we present some characteristics of a finite-state Markov Chain, which is a discrete-time stochastic process. Later, we introduce the concept of Markov Chains with rewards and the Markov Decision Theory, and apply them through various examples. The paper ends with the description of the Dynamic Programming Algorithm, a set of rules that maximizes the aggregate rewrd over a given number of trials in a multi-trial Markov Chain with rewards.

Contents

1. Preliminaries and Definitions

In this section, we provide a formal treatment of various concepts of statistics like the notion of events and probability mapping. We also define a random variable which is the building block of any stochastic process, of which the Markov process is one.

Axiom of events Given a sample space Ω, the class of subsets F of Ω that constitute the set of events satisfies the following axioms:

1.Ω is an event.

2.For every sequence of events $A_1, A_2, ...,$ the union $\cup_n A_n$ is an event

3.For every event A, the complement A^c is an event.

F is called the event space.

Axiom of Probability Given any sample space Ω and any class of event spaces F, a probability rule is a function $\mathbb{P}\{\}$ mapping each event $A \in F$ to a (finite) real

number in such a way that the following three probability axioms hold:

1.$\mathbb{P}\{\Omega\} = 1$.

2.For every event $A, \mathbb{P}\{A\} \geq 0$.

3.The probability of the union of any sequence $A_1, A_2, ...$ of disjoint events is given by the sum of the individual probabilities

$$\mathbb{P}\{\cup_{n=1}^{\infty} A_n\} = \sum_{n=1}^{\infty} \mathbb{P}\{A_n\} \tag{1.1}$$

With this definition of probability mapping of an event, we will now characterize a random variable, which in itself, is a very important concept.

Definition 1.2. A random variable is a function X from the sample space Ω of a probability model to the set of real numbers $\mathbb{R}$, denoted by $X(\omega)$ for $\omega \in \Omega$ where the mapping $X(\omega)$ must have the property that $\{\omega \in \Omega : X(\omega) \leq x\}$ is an event for each $x \in \mathbb{R}$.

Thus, random variables can be looked upon as real-valued functions from a set of possible outcomes (sample space Ω), only if a probability distribution, defined as $F_X(x) = \mathbb{P}\{\omega \in \Omega : X(\omega) \leq x\}$ exists. Or in other words, random variables are the real-valued functions, only if they turn the sample space to a probability space.

A **stochastic process** (or random process) is an infinite collection of random variables. Any such stochastic process is usually indexed by a real number, often interpreted as time, so that each sample point maps to a function of time giving rise to a sample path. These sample paths might vary continuously with time or might vary only at discrete times. In this paper, we will be working with stochastic processes that are discrete in time variation.

2. Finite State Markov Chains

A class of stochastic processes that are defined only at integer values of time are called integer-time processes, of which a finite state Markov Chain is an example. Thus, at each integer time $n \geq 0$, there is an integer-valued random variable X_n, called the state at time n and a Markov Chain is the collection of these random variables $\{X_n; n \geq 0\}$. In addition to being an integer-time process, what really makes a Markov Chain special is that it must also satisfy the following Markov property.

Definition 2.1. Markov property of an integer-time process $\{X_n, n \geq 0\}$, is the property by which the sample values for random variable, such as X_n, $n \geq 1$, lie in a countable set S, and depend on the past only through the most recent random variable X_{n-1}. More specifically, for all positive integers n, and for all $i, j, k, ..., m$ in S

$$\mathbb{P}(X_n = j | X_{n-1} = i; X_{n-2} = k, ..., X_0 = m) = \mathbb{P}(X_n = j | X_{n-1} = i) \tag{2.2}$$

Definition 2.3. A homogeneous Markov Chain has the property that $\mathbb{P}\{X_n = j | X_{n-1} = i\}$ depends only on i and j and not on n, and is denoted by

$$\mathbb{P}\{X_n = j | X_{n-1} = i\} = P_{ij} \tag{2.4}$$

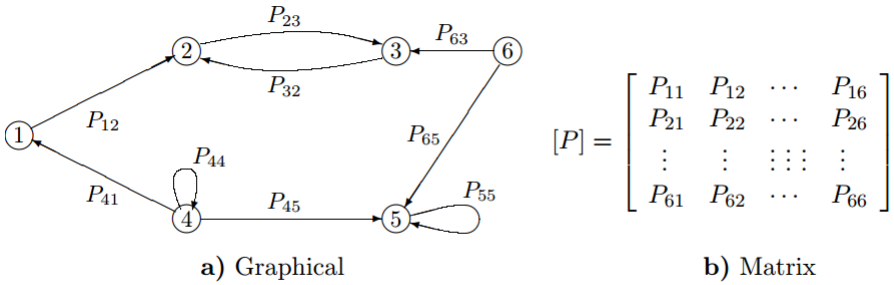

FIGURE 1. Graphical and Matrix Representation of a 6 state Markov Chain [1]

The initial state X_0 can have an arbitrary probability distribution. A finite-state Markov chain is a Markov chain in which S is finite.

Markov chains are often described by a directed graph as in Figure 1a . In this graphical representation, there is one node for each state and a directed arc for each non-zero transition probability. If $P_{ij} = 0$, then the arc from node i to node j is omitted. A finite-state Markov chain is also often described by a matrix $[P]$ as in Figure 1b. If the chain has M states, then $[P]$ is an M x M matrix with elements P_{ij}.

3. Classification of States of a Markov Chain

An (n-step) **walk** is an ordered string of nodes, $(i_0, i_1, ...i_n), n \geq 1$, in which there is a directed arc from i_{m-1} to i_m for each m, $1 \leq m \leq n$. A **path** is a walk in which no nodes are repeated. A **cycle** is a walk in which the first and last nodes are the same and no other node is repeated

Definition 3.1. A state j is accessible from i (abbreviated as $i \to j$) if there is a walk in the graph from i to j

For example, in Figure 1(a), there is a walk from node 1 to node 3 (passing through node 2), so state 3 is accessible from 1.

Remark 3.2. We see that $i \to j$ if and only if $\mathbb{P}\{X_n = j | X_0 = i\} > 0$ for some $n \geq 1$. We denote $\mathbb{P}\{X_n = j | X_0 = i\}$ by $\mathbf{P_{ij}^n}$. Thus $i \to j$ if and only if $P_{ij}^n > 0$ for some $n \geq 1$. For example, in Figure 1(a), $P_{13}^2 = P_{12}P_{23} > 0$.

Two distinct states i and j communicate (denoted by $i \leftrightarrow j$) if i is accessible from j and j is accessible from i.

Definition 3.3. For finite-state Markov chains, a recurrent state is a state i which is accessible from all the states that are, in turn, accessible from the state i . Thus i is recurrent if and only if $i \to j \implies j \to i$. A transient state is a state that is not recurrent.

A transient state i , therefore, has the property that if we start in state i, there is a non-zero probability that we will never return to i.

Definition 3.4. A class C of states is a non-empty set of states such that each $i \in C$ communicates with every other state $j \in C$ and communicates with no $j \notin C$.

With this definition of a class, we now specify some characteristics of states belonging to the same class.

Theorem 3.5. *For finite-state Markov chains, either all states in a class are transient or all are recurrent.*

Proof. : Let C be a class, and i, $m \in C$ are states of Markov chains in the same class (i.e., $i \leftrightarrow m$). Assume for contradiction that state i is transient (i.e., for some state $j \in C$, $i \to j$ but $j \nrightarrow i$). Then since, $m \to i$ and $i \to j$, so $m \to j$. Now if $j \to m$, then the walk from j to m could be extended to i which would make the state i recurrent and would be a contradiction. Therefore there can be no walk from j to m, which makes the state m transient. Since we have just shown that all states in a class are transient if any one state in the class is, it follows that the states in a class are either all recurrent or all transient. □

Definition 3.6. The period of a state i, denoted $d(i)$, is the greatest common divisor (gcd) of those values of n for which $P_{ii}^n > 0$. If the period is 1, the state is called aperiodic.

Theorem 3.7. *For any Markov chain , all states in a class have the same period.*

Proof. : Let i and j be any distinct pair of states in a class C. Then $i \leftrightarrow j$ and there is some r such that $P_{ij}^r > 0$ and some s such that $P_{ji}^s > 0$. Since there is a walk of length $r + s$ from i to j and back to i, $r + s$ must be divisible by $d(i)$. Let t be any integer such that $P_{jj}^t > 0$. Since there is a walk of length $r + t + s$ from i to j, then back to j, and then to i, $r + t + s$ is divisible by $d(i)$, and thus t is divisible by $d(i)$. Since this is true for any t such that $P_{jj}^t > 0$, $d(j)$ is divisible by $d(i)$. Reversing the roles of i and j, $d(i)$ is divisible by $d(j)$, so $d(i) = d(j)$. □

Since the states in a class C all have the same period and are either all recurrent or all transient, we refer to the class C itself as having the period of its states and as being recurrent or transient.

Definition 3.8. For a finite-state Markov chain, an ergodic class of states is a class that is both recurrent and aperiodic. A Markov chain consisting entirely of one ergodic class is called an ergodic chain.

Definition 3.9. A unichain is a finite-state Markov chain that contains a single recurrent class and possibly, some transient states.

Thus, an ergodic unichain is a Markov chain which solely consists of a single aperiodic recurrent class.

4. Matrix Representation and the Steady state, $[P^n]$ for large n

The matrix $[P]$ of transition probabilities of a Markov chain is called a stochastic matrix; that is, a square matrix of nonnegative terms in which the elements in each row sum to 1. We first consider the n step transition probabilities P_{ij}^n in terms of

$[P]$. The probability, of reaching state j from state i, in two steps is the sum over k of the probability of transition from i first to k and then to j. Thus

$$P_{ij}^2 = \sum_{i=1}^{M} P_{ik} P_{kj} \tag{4.1}$$

Noticeably, this is just the i, j term of the product of the matrix $[P]$ with itself. If we denote $[P][P]$ as $[P^2]$,this means that P_{ij}^2 is the (i, j) element of the matrix $[P^2]$. Similarly, it can be shown that $[P]^n = [P^n]$ and $[P^{m+n}] = [P^m][P^n]$. The last equality can be written explicitly in terms of an equation, known as the Chapman-Kolmogorov equation.

$$P_{ij}^{m+n} = \sum_{k=1}^{M} P_{ik}^m P_{kj}^n \tag{4.2}$$

Theorem 4.3. *For an aperiodic Markov Chain, there exists an $N < \infty$ such that $P_{ii}^n > 0$ for all $i \in \{1, .., k\}$ and all $n \geq N$.*

Lemma 4.4. *Let $A = \{a_1, a_2...\}$ be a set of positive integers which are (i) relatively prime and (ii) closed under addition. Then there is some $N < \infty$ such that for any $n \geq N$, $n \in A$.*

Proof. : The proof of this lemma can be found in "Olle Haggstrom. Finite Markov Chains and Algorithmic Applications. Cambridge University Press, 2002" and because it is a technical number theory lemma, we will not reproduce the proof here.

Proof. (Theorem): Let $A_i = \{n \geq 1 | P_{ii}^n > 0\}$ be the set of return times to state i starting from state i. By the aperiodicity of the Markov chain, A_i has a greatest common factor of 1, satisfying part (i) of Lemma 4.4.

Next, let a_1 and $a_2 \in A_i$, then $P_{ii}^{a_1} > 0$ and $P_{ii}^{a_2} > 0 \implies P_{ii}^{a_1+a_2} = \sum_{k=1}^{M} P_{ik}^{a_1} P_{ki}^{a_2} > 0$, which in turn implies that $a_1 + a_2 \in A_i$. Hence A_i is closed under addition and satisfies part (ii) of Lemma 4.4. The theorem then follows from Lemma 4.4. □.

Corollary 4.5. *For ergodic Markov Chains there exists an $M < \infty$ such that $P_{ij}^n > 0$ for all $i, j \in \{1, .., k\}$ and all $n \geq M$.*

Proof. : Using the aperiodicity of ergodic Markov chains, and applying Theorem 4.4, we are able to find an integer $N < \infty$ such that $P_{ii}^n > 0$ for all $i \in \{1, .., k\}$ and all $n \geq N$.

Next, we pick two arbitrary states i and j. Since an ergodic Markov chain consists of a single recurrent class, states i and j must belong to the same class and hence communicate with each other. Thus, there is some $n_{i,j}$ such that $P_{ij}^{n_{i,j}} > 0$. Let $M_{i.j} = N + n_{i.j}$.

Then, for any $m \geq M_{i,j}$ we have

$\mathbb{P}(X_m = j | X_0 = i)$
$\geq \mathbb{P}(X_m = j, X_{m-n_{i,j}} = i | X_0 = i)$ as the event is a subset of the event in the previous line)
$= \mathbb{P}(X_{m-n_{i,j}} = i | X_0 = i) P(X_m = j | X_{m-n_{i,j}} = i)$ by the independence of the events
> 0

Therefore $P_{ij}^m > 0$ for all $m \geq M_{i,j}$. Repeating this process for all combinations of two arbitrary states i and j we get $\{M_{1,1}..., M_{1,k}, M_{2,1}, ...M_{k,k}\}$. Now we set $M = \max\{M_{1,1}, ..., M_{k,k}\}$ and this M satisfies the required property as stated in the corollary. □

The transition matrix : The matrix $[P^n]$ is very important as the i, j element of this matrix is $P_{ij}^n = \mathbb{P}\{X_n = j | X_0 = i\}$. Due to the Markov property, each state in a Markov chain remembers only the most recent history. Thus, we would expect the memory of the past to be lost with increasing n, and the dependence of P_{ij}^n on both n and i to disappear as $n \to \infty$. This has two implications: first, $[P^n]$ should converge to a limit as $n \to \infty$, and, second, for each column j, the elements in that column namely, $P_{1j}^n, P_{2j}^n, ..., P_{Mj}^n$ should all tend toward the same value. We call this converging limit, π_j .

And if $P_{ij}^n \to \pi_j$, each row of the limiting matrix converge to $(\pi_1, ..., \pi_M)$, i.e., each row becomes same as every other row. We will now prove this convergence property for an ergodic finite-state Markov Chain.

Theorem 4.6. *Let $[P]$ be the matrix of an ergodic finite-state Markov chain. Then there is a unique steady-state vector π, which is positive and satisfies*

$$\lim_{n\to\infty} P_{ij}^n = \pi_j \text{ for each } i, j \tag{4.7}$$

or in a compact notation

$$\lim_{n\to\infty} [P^n] = e\pi \text{ where } e = (1, 1, ..., 1)^T \tag{4.8}$$

Proof. For each i, j, k and n, we use the Chapman-Kolmogorov equation, along with $P_{kj}^n \leq \max_l P_{lj}^n$ and $\sum_k P_{ik} = 1$. This gives us

$$P_{ij}^{n+1} = \sum_k P_{ik} P_{kj}^n \leq \sum_k P_{ik} \max_l P_{lj}^n = \max_l P_{lj}^n$$

Similarly, we also have

$$P_{ij}^{n+1} = \sum_k P_{ik} P_{kj}^n \geq \sum_k P_{ik} \min_l P_{lj}^n = \min_l P_{lj}^n$$

Now, let $\alpha = \min_{i,j} P_{ij}$ and l_{min} be the value of l that minimizes P_{lj}^n. Then

$$\begin{aligned}
P_{ij}^{n+1} &= \sum_k P_{ik} P_{kj}^n \\
&= \sum_{k \neq l_{min}} P_{ik} P_{kj}^n + P_{il_{min}} \min_l P_{lj}^n \\
&\leq \sum_{k \neq l_{min}} P_{ik} \max_l P_{lj}^n + P_{il_{min}} \min_l P_{lj}^n \\
&= \max_l P_{lj}^n - P_{il_{min}} (\max_l P_{lj}^n - \min_l P_{lj}^n) \\
&\leq \max_l P_{lj}^n - \alpha (\max_l P_{lj}^n - \min_l P_{lj}^n)
\end{aligned}$$

which would further imply that $\max_i P_{ij}^{n+1} \leq \max_l P_{lj}^n - \alpha(\max_l P_{lj}^n - \min_l P_{lj}^n)$.

By a similar set of inequalities , we have $\min_i P_{ij}^{n+1} \geq \min_l P_{lj}^n + \alpha(\max_l P_{lj}^n - \min_l P_{lj}^n)$.

Next, we subtract the two equations to obtain

$$\max_i P_{ij}^{n+1} - \min_i P_{ij}^{n+1} \leq (\max_l P_{lj}^n - \min_l P_{lj}^n)(1-2\alpha)$$

Then, using induction on n, we obtain from the above equation that $\max_i P_{ij}^n - \min_i P_{ij}^n \leq (1-2\alpha)^n$
Now, if we assume that $P_{ij} > 0$, $\forall i, j$ then $\alpha > 0$ and since, $(1-2\alpha) < 1$, in the limit $n \to \infty$ we would have $\max_i P_{ij}^n - \min_i P_{ij}^n \to 0$ or

$$\lim_{n\to\infty} \max_l P_{lj}^n = \lim_{n\to\infty} \min_l P_{lj}^n > 0 \tag{4.9}$$

But, α might not always be positive. However, due to the ergodicity of our Markov chain and Corollary 4.6, we know that there exists some integer $h > 0$ such that $P_{ij}^h > 0$. Carrying out a similar process as before and replacing α by $\min_{i,j} P_{ij}^h$, which is now positive, we obtain the equation $\max_i P_{ij}^n - \min_i P_{ij}^n \leq (1-2\alpha)^{n/h}$ from which we get the same limit as equation 4.9.

Now, define the vector $\pi > 0$ as $\pi_j = \lim_{n\to\infty} \max_l P_{lj}^n = \lim_{n\to\infty} \min_l P_{lj}^n > 0$. Since π_j lies between the minimum and the maximum of P_{lj}^n, in this limit, $\pi_j = \lim_{n\to\infty} P_{lj}^n > 0$. This can be represented in a more compact notation as

$$\lim_{n\to\infty} [P^n] = e\pi \text{ where } e = (1, ..., 1)^T$$

which proves the existence of the limit in the theorem.

To complete the rest of the proof, we also need to show that π as defined above is the unique steady-state vector. Let μ be any steady state vector, i.e., any probability vector solution to $\mu[P] = \mu$. Then μ must satisfy $\mu = \mu[P^n]$ for all $n > 1$. In the limit $n \to \infty$,

$$\mu = \mu \lim_{n\to\infty} [P^n] = \mu e\pi = (\mu e)\pi = e\pi = \pi.$$

Thus, π is the steady state vector and is unique. □

5. Markov Chains with rewards

In this section, we look into a more interesting problem, namely the Markov Chain with Rewards. Now, we associate each state i of a Markov chain with a reward, r_i. The reward r_i associated with a state could also be viewed as a cost or some real-valued function of the state. The concept of a reward in each state is very important for modelling corporate profits or portfolio performance. It is also useful for studying queuing delay, the time until some given state is entered, and similar interesting phenomena.
It is clear from the setup, the sequence of rewards associated with the transitions between the states of the Markov chain is not independent, but is related by the statistics of the Markov chain.

The **gain** is defined to be the steady-state expected reward per unit time, assuming a single recurrent class of states and is denoted by $\mathbf{g} = \sum_{\mathbf{i}} \pi_{\mathbf{i}} \mathbf{r_i}$ where π_i is the steady-state probability of being in state i.

Let us now explore the concept of Markov Chains with Rewards with an Example.

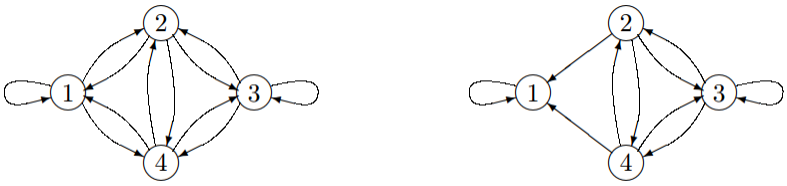

FIGURE 2. The conversion of a recurrent Markov chain with M = 4 into a chain for which state 1 is a trapping state [1]

Example 5.1. The first-passage time of a set A with respect to a stochastic process is the time until the stochastic process first enters A. Thus, it is an interesting concept, for we might be interested in knowing the average number of steps it takes to go from one given state, say i, to a fixed state, say 1 in a Markov chain. Here we calculate the expected value of the first-passage-time.

Since the first-passage time to a given state (say state 1) is independent of the transitions made after the first entry into that step, we can modify any given Markov chain to convert this required state into a trapping state so that there is no exit from that step. Which means, we modify P_{11} to 1 and P_{1j} to 0 for all $j \neq 1$. We leave P_{ij} unchanged for all $i \neq 1$ and all j. We show such a modification in Figure 2. This modification does not change the probability of any sequence of states up to the point that state 1 is first entered and so the essential behavior of the Markov chain is preserved.

Let us call v_i the expected number of steps to first reach state 1 starting in state $i \neq 1$. This is our required expected first passage time to state 1. v_i can be computed considering the first step and then adding the remaining steps to reach state 1 from the state that is entered next. For example, for the chain in figure 2, we have the equations

$$v_2 = 1 + P_{23}v_3 + P_{24}v_4.$$
$$v_3 = 1 + P_{32}v_2 + P_{33}v_3 + P_{34}v_4.$$
$$v_4 = 1 + P_{42}v_2 + P_{43}v_3.$$

Similarly, for an arbitrary chain of M states where 1 is a trapping state and all other states are transient, this set of equations becomes

$$v_i = 1 + \sum_{j \neq 1} P_{ij}v_j \text{ where } i \neq 1 \tag{5.2}$$

We can now define $r_i = 1$ for $i \neq 1$ and $r_i = 0$ for $i = 1$, to be the unit reward obtained for entering the trapping state from state i. This makes intuitive sense because in a real-life situation, we would expect the reward to cease to exist once the trapping state is entered.

With this definition of r_i, v_i becomes the expected aggregate reward before entering the trapping state or the expected transient reward. If we take v_1 to be 0 (i.e 0 reward in recurrent state) , Equation 5.2 along with $v_1 = 0$, has the vector form

$$v = r + [P]v. \tag{5.3}$$

In the next two subsections we will explore more general cases of expected aggregate rewards in a Markov Chain with rewards.

5.1. **The expected aggregate reward over multiple transitions.** In the general case, we let X_m be the state at time m and $R_m = R(X_m)$ the reward at that time m, which, in the context of the previous example, would imply that if the sample value of X_m is i, then r_i is the sample value of R_m. Taking $X_m = i$, the aggregate expected reward $v_i(n)$ over n trials from X_m to X_{m+n-1} is

$$\begin{aligned} v_i(n) &= \mathbb{E}[R(X_m) + R(X_{m+1}) + \ + R(X_{m+n-1})|X_m = i] \\ &= r_i + \sum_j P_{ij} r_j + ... + \sum_j P_{ij}^{n-1} r_j. \end{aligned}$$

In case of a homogeneous Markov Chain, this expression does not depend on the starting time m. Considering the expected reward for each initial state i, this expression can be compactly written in the following vector notation.

$$v(n) = r + [P]r + ... + [P^{n1}]r = \sum_{h=0}^{n-1} [P^h]r \tag{5.4}$$

where $v(n) = (v_1(n), v_2(n), ..., v_M(n))^T$, $r = (r_1, ..., r_M)^T$ and P^0 is the identity matrix. Now if we take the case where the Markov chain is an ergodic unichain, we have $\lim_{n\to\infty}[P]^n = e\pi$. Multiplying both sides of the limit with the vector r, we obtain $\lim_{n\to\infty}[P]^n r = e\pi r = ge$ where g is the steady-state reward per unit time. And by definition, g is equal to πr.
If $g \neq 0$, then from equation 5.4, we can say that $v(n)$ changes by approximately ge for each unit increase in n. Thus, $v(n)$ does not have a limit as $n \to \infty$. However, as shown below, $v(n) - nge$ does have a limit, given by

$$\begin{aligned} &\lim_{n\to\infty}[v(n) - nge] \\ &= \lim_{n\to\infty} \sum_{h=0}^{n-1} [P^h - e\pi]r. \text{ since } e\pi r = ge \end{aligned}$$

For an ergodic unichain, the limit exits or the infinite sum converges because it can be shown that $|P_{ij}^n - \pi_j| < o(\exp(-n\varepsilon))$ for very small ε and for all i, j, n [1]$p.126$. Thus, $\sum_{h=n}^{\infty} (P_{ij}^h - \pi_j) < o(\exp(-n\varepsilon))$.

This limit is a vector over the states of the Markov chain, which gives the asymptotic relative expected advantage of starting the chain in one state relative to another. It is also called the **relative gain vector** and denoted by $\boldsymbol{w}$.

Theorem 5.5. *Let $[P]$ be the transition matrix for an ergodic unichain. Then the relative gain vector w satisfies the following linear vector equation.*

$$w + ge = [P]w + r \tag{5.6}$$

Proof. : Multiplying $[P]$ on the left of both the sides of equation in the definition of w , we get

$$\begin{aligned}
[P]w &= \lim_{n\to\infty} \sum_{h=0}^{n-1} ([P^{h+1} - e\pi]r) \text{ since } e\pi = [P]e\pi \\
&= \lim_{n\to\infty} \sum_{h=1}^{n} ([P^{h} - e\pi]r) \\
&= \lim_{n\to\infty} \sum_{h=0}^{n} ([P^{h} - e\pi]r) - [P^0 - e\pi]r \\
&= w - [P^0]r + e\pi r \\
&= w - r + ge.
\end{aligned}$$

Rearranging the terms, we get the required result.

5.2. **The expected aggregate reward with an additional final reward.** A variation to the previous situation might be the case when an added final reward is assigned to the final state. We can view this final reward, say u_i, as a function of the final state i. For example, it might be particularly advantageous to end in one particular state versus the other.

As before, we set $R(X_{m+h})$ to be the reward at time $m + h$, for $0 \leq h \leq n - 1$ and define $U(X_{m+n})$ to be the final reward at time $m + n$, where $U(X) = u_i$ for $X = i$. Let $v_i(n, u)$ be the expected reward from time m to $m+n$, using the reward r from time m to $m + n - 1$ and using the final reward u at time $m + n$. Then the expected reward is obtained by modifying Equation (5.4):

$$v(n, u) = r + [P]r + \cdots + [P^{n-1}]r + [P^n]u = \sum_{h=0}^{n-1} [P^h]r + [P^n]u. \tag{5.7}$$

This simplifies if u is taken to be the relative-gain vector w.

Theorem 5.8. *Let $[P]$ be the transition matrix of a unichain and let w be the corresponding relative-gain vector. For each $n \geq 1$, if $u = w$, then*

$$v(n, w) = nge + w \tag{5.9}$$

For an arbitrary final reward vector u,

$$v(n, u) = nge + w + [P^n](u - w) \tag{5.10}$$

Proof. : We use induction to prove the theorem.

For $n = 1$, we obtain from (5.7) and theorem 5.5 that

$$v(1, w) = r + [P]w = ge + w \tag{5.11}$$

so the induction hypothesis is satisfied for $n = 1$.

For $n > 1$,

$$\begin{aligned} v(n,w) &= \sum_{h=0}^{n-1}[P^h]r + [P^n]w \\ &= \sum_{h=0}^{n-2}[P^h]r + [P^{n-1}]r + [P^n]w \\ &= \sum_{h=0}^{n-2}[P^h]r + [P^{n-1}](r + [P]w) \\ &= \sum_{h=0}^{n-2}[P^h]r + [P^{n-1}](ge + w) \\ &= (\sum_{h=0}^{n-2}[P^h]r + [P^{n-1}]w) + [P^{n-1}]ge \\ &= v(n-1,w) + ge. \text{ since } ge = e\pi r \text{ and } e\pi = [P^{n-1}]e\pi \end{aligned}$$

Using induction on n, we obtain (5.9). To establish (5.10), note from (5.7) that

$$v(n,u) - v(n,w) = [P^n](u-w) \tag{5.12}$$

Then (5.10) follows by using (5.9) for the value of $v(n,w)$.

6. Markov decision theory

Till now, we have only analyzed the behavior of a Markov chain with rewards. In this section, we consider a much intricate situation where a decision maker can choose among various possible rewards and transition probabilities. At each time m, the decision maker, given $X_m = i$, selects one of the K_i possible choices for state i and each choice k is associated with a reward $r^{(k)}$ and a set of transition probabilities $P_{ij}^{(k)}, \forall j$. We also assume that if decision k is selected at time m, the probability of entering state j at time $m+1$ is $P_{ij}^{(k)}$, independent of earlier states and decisions.

Example 6.1. An example is given in Figure 3, in which the decision maker has a choice between two possible decisions in state 2 ($K_2 = 2$), and has a single choice in state 1 ($K_1 = 1$). Such a situation might arise when we know that there is a trade off between instant gain (alternative 2) and long term gain (alternative 1). We see that decision 2 is the best choice in state 2 at the nth of n trials for a large n because of the huge reward associated with this decision. However, at an earlier step, it is less obvious what to do. We will address this question in the next section, when we derive the algorithm to choose the right decision at each trial to maximize the aggregate reward

The set of rules used by the decision maker in selecting an alternative at each time is called a **policy**. We might be interested in calculating the expected aggregate reward over n steps of the Markov chain as a function of the policy used by the decision maker.

A familiar situation would be where for each state i, the policy uses the same decision, say k_i, at each occurrence of i. Such a policy is called a stationary policy. Since both rewards and transition probabilities in a stationary policy depend only

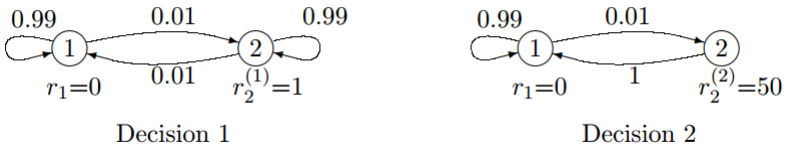

FIGURE 3. A Markov decision problem with two alternatives in state 2 [1]

on the state and the corresponding decision, and not on time, such a policy corresponds to a homogeneous Markov chain with transition probabilities $P_{ij}^{(k_i)}$. We denote the resulting transition probability matrix of the Markov Chain as $[P^k]$, where $k = (k_1, ..., k_M)$. The aggregate gain for any such policy was found in the previous section.

7. Dynamic programming algorithm

In a more general case, where, the choice of a policy at any given point in time varies as a function of time, we might want to derive an algorithm to choose the optimal policy for maximizing expected aggregate reward over an arbitrary number n of trials from times m to $m + n - 1$. It turns out that the problem is further simplified if we include a final reward $\{u_i | 1 \leq i \leq M\}$ at time $m + n$. This final reward u is chosen as a fixed vector, rather than as part of the choice of policy.

The optimized strategy, as a function of the number of steps n and the final reward u, is called an **optimal dynamic policy** for that u. This policy is found from the dynamic programming algorithm.

First let us consider the optimal decision with $n = 1$. Given $X_m = i$, a decision k is made with immediate reward $r_i^{(k)}$. If the next state X_{m+1} is state j, then the transition probability is $P_{ij}^{(k)}$ and the final reward is then u_j. The expected aggregate reward over times m and $m + 1$, maximized over the decision k, is then

$$v_i^*(1, u) = \max_k \{r_i^{(k)} + \sum_j P_{ij}^{(k)} u_j\}. \tag{7.1}$$

Next, we look at $v_i^*(2, u)$, i.e., the maximal expected aggregate reward starting at $X_m = i$ with decisions made at times m and $m + 1$ and a final reward at time $m + 2$.

The key to dynamic programming is that an optimal decision at time $m+1$ can be selected based only on the state j at time $m + 1$. That the decision is optimal independent of the decision at time m can be shown using the following argument. Regardless of what the decision is made at time m, the maximal expected reward at times $m+1$ (given $X_{m+1} = j$) , is $\max_k(r_j^{(k)} + \sum_l P_{lj}^{(k)} u_l)$. Ths is equal to $v_j^*(1, u)$, as found in (7.1).

Using this optimized decision at time $m+1$, it is seen that if $X_m = i$ and decision k is made at time m, then the sum of expected rewards at times $m + 1$ and $m + 2$

is $\sum_j P_{ij}^{(k)} v_j^*(1,u)$. Adding the expected reward at time m and maximizing over decisions at time m

$$v_i^*(2,u) = \max\{r_i^{(k)} + \sum_j P_{ij}^{(k)} v_i^*(1,u)\}. \tag{7.2}$$

Continuing this way, we find, after n steps, that

$$v_i^*(n,u) = \max\{r_i + \sum_j P_{ij} v_i^*(n-1,u)\}. \tag{7.3}$$

Noteworthy is the fact that the algorithm is independent of the starting time m. The parameter n, usually referred to as **stage** n, is the number of decisions over which the aggregate gain is being optimized. So we obtain the optimal dynamic policy for any fixed final reward vector u and any given number of trials.

Example 7.4. The dynamic program algorithmn can be elaborated with a short example. We reconsider the case in Example 6.1 with final reward $u = 0$. Since $r_1 = 0$ and $u_1 = u_2 = 0$, the aggregate gain in state 1 at stage 1 is

$$v_1^*(1,u) = r_1 + \sum_j P_{1j} u_j = 0.$$

Similarly, since policy 1 has an immediate reward $r_2^{(1)} = 1$, and policy 2 has an immediate reward $r_2^{(2)} = 50$ in stage 2,

$$v_2^*(1,u) = \max\{[r_2^{(1)} + \sum_j P_{2j}^{(1)} u_j], [r_2^{(2)} + \sum_j P_{2j}^{(2)} u_j]\} = \max\{1, 50\} = 50$$

To go on to the stage 2, we use the results above for $v_j(1,u)$.

$$v_1^*(2,u) = r_1 + P_{11} v_1^*(1,u) + P_{12} v_2^*(1,u) = P_{12} v_2^*(1,u) = 0.5$$

$$v_2^*(1,u) = \max\{[r_2^{(1)} + \sum_j P_{2j}^{(1)} v_j^*(1,u)], [r_2^{(2)} + P_{21}^{(2)} v_1^*(2,u)]\}$$

$$= \max\{(1 + P_{22}^{(2)} v_2^*(1,u)), 50\} = \max\{50.5, 50\} = 50.5$$

Thus for a two-trial situation like this, decision 1 is optimal in state 2 for the first trial (stage 2), and decision 2 is optimal in state 2 for the second trial (stage 1). This is because, the choice of decision 2 at stage 1 has made it very profitable to be in state 2 at stage 1. Thus if the chain is in state 2 at stage 2, it is preferable to choose decision 1 (i.e., the small unit gain) at stage 2 with the corresponding high probability of remaining in state 2 at stage 1.

For larger n, however, $v_1^*(n,u) = n/2$ and $v_2^*(n,u) = 50 + n/2$. The optimum dynamic policy (for $u = 0$) would then be decision 2 for stage 1 (i.e., for the last decision to be made) and decision 1 for all stages $n > 1$ (i.e., for all decisions before the last).

From this example we also see that the maximization of expected gain is not always what is most desirable in all applications. For instance, someone who is risk-averse might well prefer decision 2 at the next to final decision (stage 2), as this guarantees a reward of 50, rather than taking a small chance of losing that reward.

Acknowledgments. I would like to thank Peter May for organizing the REU and my mentor, Yan Zhang, for painstakingly reviewing my paper. This paper has only been possible because of their help, for which I am indebted to them.

References

[1] Robert Gallager Course Notes, MIT OCW $http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-262-discrete-stochastic-processes-spring-2011/course-notes/MIT6_262S11_chap03.pdf$

[2] Olle Haggstrom. Finite Markov Chains and Algorithmic Applications. Cambridge University Press, 2002