Markov Decision Processes

Lecture notes for the course "Games on Graphs"

B. Srivathsan

Chennai Mathematical Institute, India

1 Markov Chains

We will define Markov chains in a manner that will be useful to study simple stochastic games. The usual definition of Markov chains is more general.

Definition 1 (Markov Chains) A *Markov Chain* G is a graph (V_{avg}, E). The vertex set V_{avg} is of the form $\{1, 2, \ldots, n-1, n\}$. Vertex $n-1$ is called the 0-sink and vertex n is called the 1-sink. From every vertex, there are two outgoing edges each with probability $\frac{1}{2}$. Note that both the outgoing edges can lead to the same vertex. Only edges from $n-1$ and n are self loops.

A Markov Chain of n vertices can be specified using its $n \times n$ transition matrix. Each entry of this matrix is either 0, $\frac{1}{2}$ or 1. A Markov Chain is said to be *stopping* if from every vertex, there is a path to a sink vertex.

Definition 2 (Reachability probabilities) For a stopping Markov Chain G, let $\overline{v} := \langle v_1, v_2, \ldots, v_n \rangle$ denote the probabilities to reach the 1-sink n from each vertex. Then, $\overline{v}$ satisfies the following constraints:

$$\begin{aligned} v_n &= 1 \\ v_{n-1} &= 0 \\ \text{for each } 1 \leq i \leq n-2 \quad v_i &= \frac{1}{2}v_j + \frac{1}{2}v_k \quad \text{where } j \text{ and } j \text{ are children of } i \end{aligned}$$

Let us restrict $\overline{v}$ to the first $n-2$ coordinates. If $\overline{v}$ is seen as a column vector, then the above system of equations can be written as $\overline{v} = Q\overline{v} + \overline{b}$, where Q is the $n-2 \times n-2$ transition matrix of G (not containing sink states $n-1$, n) and $\overline{b}$ is a column vector containing either 0, $\frac{1}{2}$ or 1 (giving the 1 step probability to reach 1-sink). A solution to this equation gives the reachability probabilities.

Here are some results about Markov chains.

Lemma 3 Let Q be a transition matrix of a stopping Markov chain. Then, $\lim_{n\to\infty} Q^n = \mathbf{0}$.

Proof
See Theorem 11.3 (page 417) of [GS] □

Lemma 4 Let Q be a transition matrix of a stopping Markov chain. Then $I - Q$ is invertible. Moreover, $(I - Q)^{-1} = I + Q + Q^2 + \ldots$.

Proof
See Theorem 11.4 (page 418) of [GS] □

Theorem 5 *For a stopping Markov chain G, the system of equations $\overline{v} = Q\overline{v} + b$ in Definition 2 has a unique solution, given by $\overline{v} = (I - Q)^{-1}b$.*

Proof
Follows from Lemma 4. □

2 Markov Decision Processes

Definition 6 (Markov Decision Process) A *Markov Decision Process (MDP) G* is a graph $(V_{avg} \sqcup V_{max}, E)$. The vertex set is of the form $\{1, 2, \ldots, n-1, n\}$. From every vertex, there are two outgoing edges. Note that both the outgoing edges can lead to the same vertex. Each outgoing edge from vertices in V_{avg} is marked with probability $\frac{1}{2}$. Vertex $n-1$ is called the 0-sink and vertex n is called the 1-sink. Only edges from $n-1$ and n are self loops.

For convenience, we will call vertices in V_{max} as *max vertices* and vertices in V_{avg} as *average vertices*.

Definition 7 (Strategy/Policy) A *strategy* σ in an MDP is a function $\sigma : V_{max} \mapsto V_{max} \sqcup V_{avg}$ which chooses an edge for every max vertex. Strategies are also called *policies*.

Observe that we have restricted to positional strategies in the above definition. Other kinds of strategies for this model are out of scope of this course.

Each strategy σ for G gives a Markov Chain G_σ. An MDP is said to be stopping if for every σ, the Markov Chain G_σ is stopping. Corresponding to G_σ is the vector denoting reachability probabilities $\overline{\sigma}$ given in Definition 2.

Definition 8 (Value vector) For an MDP, we define its value vector as:

$$\overline{v} := \begin{bmatrix} \max_\sigma v_\sigma(1) \\ \max_\sigma v_\sigma(2) \\ \vdots \\ \max_\sigma v_\sigma(n) \end{bmatrix}$$

In the following, we will give different methods to compute the value vector of an MDP.

2.1 Charactering value vector using constraints

Similar to the constraints for Markov chain given in Definition 2, we will now give constraints for MDPs. Given and MDP G, consider the following set of constraints over the variables $\langle w_1, w_2, \ldots, w_n \rangle$:

$$\begin{aligned} 0 \leq w_i \leq 1 \quad & \text{for every } i \\ w_n &= 1 \\ w_{n-1} &= 0 \\ \text{for each } i \leq n-2 \text{ in } V_{max} \quad & w_i = \max(w_j, w_k) \quad \text{where } j \text{ and } k \text{ are children of } i \\ \text{for each } i \leq n-2 \text{ in } V_{avg} \quad & w_i = \frac{1}{2}w_j + \frac{1}{2}w_k \quad \text{where } j \text{ and } k \text{ are children of } i \end{aligned} \tag{1.1}$$

Theorem 9 *For a stopping MDP, its value vector $\overline{v}$ is the unique solution to (1.1).*

Before proving the above theorem, let us mention that just from the definition of $\overline{v}$, each of its components need not be related, that is, $\overline{v}(1)$ and $\overline{v}(2)$ can arise out of different strategies. But Theorem 9 relates all these values. Therefore, in order to prove the above theorem, it will be convenient if we can get a link between the components of $\overline{v}$: we will show that the whole of $\overline{v}$ can be obtained using special types of strategies.

Definition 10 (Optimal strategies) A strategy σ for an MDP is said to be optimal if for every max vertex i, the value $v_\sigma(i)$ equals $\max(v_\sigma(j), v_\sigma(k))$ where j and k are its children.

Lemma 11 For every optimal strategy ρ, the vector $\overline{v}_\rho$ equals the value vector $\overline{v}$ of the MDP.

Proof
Clearly, $\overline{v}_\rho \leq \overline{v}$. We will now show that $\overline{v}_\sigma \leq \overline{v}_\rho$ for every strategy σ. This will prove the lemma.

Consider an arbitrary strategy σ. The value vectore $\overline{v}_\sigma$ is obtained as the solution to some equation: $\overline{v}_\sigma = Q_\sigma \overline{v}_\sigma + b_\sigma$. Therefore, $\overline{v}_\sigma = (I - Q)^{-1} b_\sigma$. We will first show that $\overline{v}_\rho \geq Q_\sigma \overline{v}_\rho + b_\sigma$. For an average vertex i, the left hand side gives $\overline{v}_\rho(i)$ and the right hand sige gives $\overline{v}_\rho(j) + \overline{v}_\rho(k)$. Hence both are equal. For a max vertex i, the left hand side gives $\max(\overline{v}_\rho(j), \overline{v}_\rho(k))$ and the right hand side gives either $\overline{v}_\rho(j)$ or $\overline{v}_\rho(k)$ (assuming that j and k are the children of i). This shows that $\overline{v}_\rho \geq Q_\sigma \overline{v}_\rho + b_\sigma$. Rearranging, we get that $\overline{v}_\rho \geq (I - Q)^{-1} b_\sigma$, which gives $\overline{v}_\rho \geq \overline{v}_\sigma$. □

Proof of Theorem 9

Each solution to (1.1) corresponds to an optimal strategy. From Lemma 11, every optimal strategy gives the unique value vector of the MDP. This shows that (1.1) cannot have multiple solutions. To show that there is a solution, it is sufficient to prove existence of an optimal strategy. This is shown in Theorem 13 in the next section.

Algorithm 1.1: Strategy improvement algorithm for MDPs, also known as policy iteration

```
algorithm strategy − improvement(G)
  σ ← an arbitrary positional strategy
  v_σ ← probabilities to reach 1−sink in G_σ
  repeat
      pick a node i ∈ V_max s.t. v_σ(i) < max(v_σ(j), v_σ(k)) where j,k are its children
      σ'(i) ← arg max{v_σ(j), v_σ(k)}
      σ ← σ'
      v_σ ← probabilities to reach 1−sink in G_σ
  until σ is optimal
```

2.2 Strategy improvement/policy iteration

We will now give an algorithm to compute an optimal strategy for an MDP. It starts with an arbitrary initial strategy and over repeated iterations, it converges to an optimal one (refer to Algorithm 1.1).

It is clear that if the algorithm terminates, it computes an optimal strategy. We will now show termination.

Let us denote the strategy obtained after the p^{th} iteration by σ_p and the value corresponding to this as $\overline{v}_p$.

Lemma 12 The values satisfy $\overline{v}_p \leq \overline{v}_{p+1}$ and there is at least one vertex where the inequality is strict.

Proof
Strategy σ_p induces a Markov Chain G_p. The values $\overline{v}_p$ are obtained as a solution to some equations: $\overline{v}_p = Q_p\overline{v} + \overline{b}_p$ as given in Definition 2. Similarly, $\overline{v}_{p+1} = Q_{p+1}\overline{v}_{p+1} + \overline{b}_{p+1}$.

Note that σ_{p+1} is obtained from σ_p by modifying the strategy at a single vertex. Therefore, the matrices Q_p and Q_{p+1} can differ at most at one row i. Wlog, let us say that $Q_p(i,j) = 1$ and $Q_{p+1}(i,k) = 1$. The rest of the entries are the same. Moreover, $\overline{v}_p(k) - \overline{v}_p(j) > 0$. The case where the change of strategy leads to modification of $\overline{b}$ vector can be handled similarly. Let us now look at $\overline{v}_{p+1} - \overline{v}_p$.

$$\begin{aligned}
\overline{v}_{p+1} - \overline{v}_p &= Q_{p+1}\overline{v}_{p+1} - Q_p\overline{v}_p \qquad \text{as } \overline{b}_p = \overline{b}_{p+1} \\
\overline{v}_{p+1} - \overline{v}_p &= Q_{p+1}\overline{v}_{p+1} - Q_{p+1}\overline{v}_p + Q_{p+1}\overline{v}_p - Q_p\overline{v}_p \\
(I - Q_{p+1})(\overline{v}_{p+1} - \overline{v}_p) &= (Q_{p+1} - Q_p)\overline{v}_p \\
(I - Q_{p+1})(\overline{v}_{p+1} - \overline{v}_p) &= (0, \ldots, 0, \overline{v}_p(k) - \overline{v}_p(j), 0, \ldots, 0)^T
\end{aligned}$$

In the last line, the quantity $\overline{v}_p(k) - \overline{v}_p(j)$ appears in the i^{th} coordinate. This, we know is strictly bigger than 0. Since G_{p+1} is stopping, from Lemma 4 we get that $I - Q_{p+1}$ is invertible. Therefore, $v_{p+1} - v_p = (I - Q_{p+1})^{-1}(0, \ldots, 0, \overline{v}_p(k) - \overline{v}_p(j), 0, \ldots, 0)^T$. From Lemma 4, we also know that the inverse has all non-negative entries, and the diagonal entries are strictly positive. This shows that $\overline{v}_p \leq \overline{v}_{p+1}$, where the inequality is strict in at least one component. □

Theorem 13 *The strategy improvement algorithm terminates with an optimal strategy.*

Proof
From Lemma 12, we can infer that after each iteration, a strategy which has not been seen before is obtained. Since the number of strategies is finite, the algorithm terminates. Moreover, the termination condition says that the final obtained strategy is optimal. □

2.3 Linear Programming

Definition 14 (Linear Program for an MDP) For an MDP G, we associate the following linear program:

$$\text{Minimize} \quad \Sigma_i \, x_i \quad \text{subject to the constraints}$$

$$\begin{aligned} 0 \leq x_i \leq 1 \quad & \text{for every } i \\ x_n = 1 & \\ x_{n-1} = 0 & \\ \text{for each } i \leq n-2 \text{ in } V_{max} \qquad x_i \geq x_j, x_k \quad & \text{where } j \text{ and } k \text{ are children of } i \\ \text{for each } i \leq n-2 \text{ in } V_{avg} \qquad x_i = \frac{1}{2}x_j + \frac{1}{2}x_k \quad & \text{where } j \text{ and } k \text{ are children of } i \end{aligned}$$

Theorem 15 *For a stopping MDP, the linear program of Definition 14 has a unique solution, and this solution gives the value vector $\overline{v}$ of the MDP.*

Proof
From 9, the value vector $\overline{v}$ is one solution to constraints of the above LP. We will now show that for any other feasible solution $\overline{x}$ of the LP, we will have $\overline{v} \leq \overline{x}$. This will imply that $\overline{v}$ is the solution which minimizes the sum, and hence will be the unique solution to the LP.

Let $\overline{x}$ be a feasible solution. Suppose there are some vertices where $\overline{v}$ is strictly bigger than $\overline{x}$. Let:

$$U = \{\ i \mid 1 \leq i \leq n,\ v_i - x_i > 0 \text{ and } |v_i - x_i| \text{ is maximum } \}.$$

The above set U gives the set of vertices where the difference between $\overline{v}$ and $\overline{x}$ is maximum. We will now show a property about this set U.

Suppose $i \in V_{max}$ belongs to U. Since $\overline{v}$ is tight (that is, it satisfies (1.1)), we have $v_i = v_j$ for some child j of i. Also, as $\overline{x}$ is a feasible solution, $x_i \geq x_j$. As i belongs to U, we have $v_i - x_i \geq v_j - x_j$ and substituting $v_i = v_j$ gives $x_i \leq x_j$. Combining this with the previous argument gives $x_i = x_j$ and hence $v_i - x_i = v_j - x_j$. This shows that j belongs to U as well. Similarly, one can show that if $i \in V_{avg}$ belongs to U, both its children j and k belong to U. Since G is stopping, this property entails that if U is non-empty, then either 0-sink or 1-sink belongs to U. But for the sink vertices both $\overline{v}$ and $\overline{x}$ assign the same value. This gives a contradiction to the assumption that there are some vertices where $\overline{v}$ gives a strictly bigger value than $\overline{x}$. □

Algorithm 1.2: Value iteration algorithm, also known as Successive approximation algorithm

```
algorithm value-iteration(G)
  let ū be the vector assigning 1 to 1-sink and 0 to everything else
  repeat
     define ū' as follows:
       u'(i) = max(u(j), u(k)), if i is a max node and j,k are its children
       u'(i) = ½u(j) + ½u(k), if i is an average node and j,k are its children
       u'(n-1) = 0
       u'(n) = 1
     let u ← u'
  until ū satisfies (1.1)
```

2.4 Value iteration

Algorithm 1.2 gives yet another method to compute the value vector. This method need not necessarily terminate. It converges to the value vector in the limit.

Let $\overline{v}_n$ be the vector obtained after n iterations. The procedure maintains the following invariant:

$$\begin{aligned} \overline{v}_m(i) &= \max(\overline{v}_{m-1}(j), \overline{v}_{m-1}(k)) \quad \text{if } i \in V_{max} \text{ with children } j,\ k \\ \overline{v}_m(i) &= \frac{1}{2}\overline{v}_{m-1}(j) + \frac{1}{2}\overline{v}_{m-1}(k) \quad \text{if } i \in V_{avg} \text{ with children } j,\ k \end{aligned} \tag{1.2}$$

Theorem 16 *For a stopping MDP, the sequence $\overline{v}_m$ converges to the value vector.*[1]

Proof
We first show that the sequence $\overline{v}_m$ converges. We can prove by induction that $\overline{v}_m$ is monotonically increasing (at each vertex). It is also bounded above by 1. Therefore $\overline{v}_m$ converges, say to v^*.

From (1.2), we get:

$$\begin{aligned} \lim_{m\to\infty} \overline{v}_m(i) &= \lim_{m\to\infty} \max(\overline{v}_{m-1}(j), \overline{v}_{m-1}(k)) \quad \text{if } i \in V_{max} \text{ with children } j,\ k \\ \lim_{m\to\infty} \overline{v}_m(i) &= \lim_{m\to\infty} (\frac{1}{2}\overline{v}_{m-1}(j) + \frac{1}{2}\overline{v}_{m-1}(k)) \quad \text{if } i \in V_{avg} \text{ with children } j,\ k \end{aligned} \tag{1.3}$$

From the fact that sequence $\overline{v}_n$ converges to v^* and from (1.3) , we get:

$$\begin{aligned} v^*(i) &= \max(v^*(j), v^*(k)) \quad \text{if } i \in V_{max} \text{ with children } j,\ k \\ v^*(i) &= \frac{1}{2}v^*(j) + \frac{1}{2}v^*(k) \quad \text{if } i \in V_{avg} \text{ with children } j,\ k \end{aligned}$$

Moreover every $\overline{v}_m$ assigns 0 to the 0-sink and 1 to the 1-sink, giving us that v^* is 0 and 1 for the 0 and 1-sink respectively. Hence v^* is a solution to the system of constraints given in (1.1). From Theorem 9, v^* is the value vector for the MDP. □

[1]The proof of this theorem written in this document was given by one of the students J. Kishor (B. Sc Maths and C.S. - 3^{rd} year)

Intuitively, we can think of the values in $\overline{v}_m$ as the maximum probability to reach the 1-sink in at most m steps (using any strategy, which need not necessarily be positional). In the limit, this sequence converges to the probability to reach 1-sink. The above theorem says that this value equals the value vector of the MDP as in Definition 8, which we know can be obtained using positional strategies. This therefore gives an insight as to why for stopping MDPs, positional strategies are as powerful as general (deterministic) strategies.

References

[GS] Charles M. Grinstead and Laurie J. Snell. *Introduction to Probability.*