STATS 701 – Theory of Reinforcement Learning
Markov Reward Processes

Ambuj Tewari

Associate Professor, Department of Statistics, University of Michigan
tewaria@umich.edu
https://ambujtewari.github.io/stats701-winter2021/

Slide Credits: Prof. M. Vidyasagar @ IIT Hyderabad, India

Winter 2021

Outline

1. Markov Processes
   - Markov Processes: Basics
   - Markov Processes with Absorbing States
2. Markov Reward Processes

Outline

1. Markov Processes
   - Markov Processes: Basics
   - Markov Processes with Absorbing States

2. Markov Reward Processes

Outline

Markov Processes: Definition

Suppose $\mathcal{X} = \{x_1, \cdots, x_n\}$ is a finite set, and $\{X_t\}_{t \geq 0}$ is a stochastic process, that is, a sequence of random variables, where each X_t assumes values in $\mathcal{X}$.

Definition

The stochastic process $\{X_t\}$ is a Markov process if

$$\Pr\{X_{t+1} | X_0^t\} = \Pr\{X_{t+1} | X_t\},$$

where X_0^t denotes $X_0, \cdots, X_t$.

Elaboration of Definition

The defining equation is a shorthand for the following statement: Suppose $u \in \mathcal{X}$ and $(y_0, \cdots, y_t) \in \mathcal{X}^{t+1}$ are arbitrary. Then

$$\Pr\{X_{t+1} = u | X_0^t = (y_0, \cdots, y_t)\} = \Pr\{X_{t+1} = u | X_t = y_t\}.$$

In other words, the conditional probability of the state X_{t+1} depends only on the value of X_t. Adding information about the values of X_τ for $\tau < t$ does not change the conditional probability.

One can also say that X_{t+1} is independent of X_0^{t-1} given X_t. (The future is conditionally independent of the past given the present.)

State Transition Matrix

A Markov process is completely characterized by the initial state distribution and the state transition matrix A, where

$$a_{ij} := \Pr\{X_{t+1} = x_j | X_t = x_i\}.$$

Thus in a_{ij}, i denotes the current state and j the future state.

Note: Some authors interchange the roles of i and j in the above.

If the transition probability does not depend on t, then the Markov process is said to be stationary; otherwise it is said to be nonstationary.

Row-Stochasticity of the State Transition Matrix

The matrix A is row-stochastic. Note that X_{t+1} must be one of $\{x_1, \cdots, x_n\}$, no matter what X_t is. Therefore

$$\sum_{j=1}^{n} a_{ij} = 1, i = 1, \ldots n.$$

The above equation can be expressed compactly as

$$A\mathbf{1}_n = \mathbf{1}_n,$$

where $\mathbf{1}_n$ denotes the column vector consisting of n ones. So 1 is an eigenvalue of A with eigenvector $\mathbf{1}_n$.

Stationary Distribution

Hence A must have a row eigenvector π corresponding to the eigenvalue 1.

Theorem

Every row-stochastic matrix A has a nonnegative row eigenvector corresponding to the eigenvalue $\lambda = 1$

If π is a nonnegative row eigenvector, it can be scaled so that it is a probability distribution (components of π add up to one). If so π is said to be a stationary distribution of A, because if $\pi A = \pi$, and X_t has the distribution π, so does X_{t+1}.

However, nothing is said about the uniqueness of π.

Irreducible Markov Processes

Definition

A row-stochastic matrix A is said to be irreducible if it is not possible to partition the permute the rows and columns symmetrically (via a permutation matrix Π) such that

$$\Pi^{-1} A \Pi = \begin{bmatrix} B_{11} & 0 \\ B_{21} & B_{22} \end{bmatrix}.$$

Equivalent Characterization of Irreducibility

Lemma

A row-stochastic matrix A is irreducible if and only if, for any pair of states $y_s, y_f \in \mathcal{X}$, there exists a sequence of states $y_1, \cdots y_l \in \mathcal{X}$ such that, with $y_0 = y_s$ and $y_{l+1} = y_f$, we have that

$$a_{y_k y_{k+1}} > 0, k = 0, \ldots, l.$$

Thus the matrix A is irreducible if and only if, for every pair of states y_s and y_f, there is a path from y_s to y_f such that every step in the path has a positive probability.

Every state is reachable from every other state (including itself) with positive probability.

Equivalent Characterization of Irreducibility – Cont'd

Theorem

A row-stochastic matrix A is irreducible if and only if

$$\sum_{l=0}^{n-1} A^l > 0,$$

where $A^0 = I$ and the inequality is componentwise.

So we can start with $M_0 = I$ and define recursively $M_{l+1} = I + AM_l$. If $M_l > 0$ for any l, then A is irreducible. If we get up to M_{n-1} and if this matrix is not strictly positive, then A is not irreducible.

Useful Properties of Irreducible Markov Processes

Theorem

Suppose A is an irreducible row-stochastic matrix. Then

1. $\lambda = 1$ *is a simple eigenvalue of* A.
2. *The corresponding row eigenvector of* A *has all positive elements.*
3. *Thus* A *has a unique stationary distribution, whose elements are all positive.*
4. *There is an integer* p, *called the* period *of* A, *such that the spectrum of* A *is invariant under rotation by* $\exp(\mathbf{i}2\pi/p)$.
5. *In particular, all* p*-th roots of unity namely* $\exp(\mathbf{i}2k\pi/p)$, $k = 0, \cdots, p-1$ *are all eigenvalues of* A.

Primitive Matrices

Definition

A row-stochastic matrix A is said to be primitive if there exists an integer l such that $A^l > 0$.

Definition

An irreducible row-stochastic matrix A is said to be aperiodic if $\lambda = 1$ is the only eigenvalue of A with magnitude one.

Theorem

A row-stochastic matrix A is primitive if and only if it is irreducible and aperiodic.

Some Examples

Suppose

$$A_1 = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0.5 & 0 \end{bmatrix}, A_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Then A_1 is primitive because $A_1^2 > 0$

However A_2 is irreducible but not primitive; its 3 eigenvalues are the cube roots of unity and it has a period $p = 3$.

Limiting Behavior of Markov Processes

Theorem

Suppose A is an irreducible row-stochastic matrix, and let π denote the corresponding stationary distribution. Then

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} A^t = \mathbf{1}_n \pi.$$

Suppose A is a primitive row-stochastic matrix, and let π denote the corresponding stationary distribution. Then

$$A^t \to \mathbf{1}_n \pi \text{ as } t \to \infty.$$

Examples

Suppose

$$A_1 = \begin{bmatrix} 0.3 & 0.7 \\ 0.7 & 0.3 \end{bmatrix}, A_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Then A_1 is strictly positive and hence primitive, while A_2 is irreducible with period 2. Both matrices have the same stationary distribution $\pi = [0.5 \;\; 0.5]$. So

$$A_1^l \to \mathbf{1}_2 \pi = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix},$$

while A_2^l equals I if l is even and A if l is odd. So A_2^l has no limit as $l \to \infty$. However, the average

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} A_2^t = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}.$$

Markov Chain Monte Carlo Method

Application: For any probability distribution ϕ,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \phi A^t = \phi \mathbf{1}_n \pi = \pi, \ \forall \phi.$$

Application: Suppose $f : \mathcal{X} \to \mathbb{R}$, and we wish to compute $E[f(\cdot), \pi]$. If $\{x_t\}_{t\geq 0}$ is any sample path of the Markov process, then define

$$\hat{f}_T = \frac{1}{T} \sum_{t=t_0+1}^{t_0+T} f(X_t).$$

Then $\hat{f}_T \to E[f(\cdot), \pi]$ as $T \to \infty$.

Outline

1. Markov Processes
   - Markov Processes: Basics
   - Markov Processes with Absorbing States

2. Markov Reward Processes

Absorbing State: Definition

Definition

A state $x_i \in \mathcal{X}$ is said to be an **absorbing state** if $X_t = x_i$ implies that $X_{t+1} = x_i$, or equivalently, that $X_\tau = x_i$ for all $\tau \geq t$. Another equivalent defintion is that row i of the state transtion matrix A consists of a 1 in column i and zeros elsewhere.

Let A be the state transition matrix. Then x_i is an absorbing state if and only if $a_{ii} = 1$ (which automatically implies that $a_{ij} = 0$ for $j \neq i$.)

A sample path of a Markov process that terminates in an absorbing state is called an episode.

Hitting Times and Hitting Probabilities of Absorbing States

Suppose a Markov process has nonabsorbing states $x_1, \cdots, x_n$ and absorbing states $a_1, \cdots, a_s$.
Assume it is possible to go from any nonabsorbing state to at least one absorbing state in a finite number of steps. (Note the change in notation.)

The state transition matrix looks like

$$M = \begin{bmatrix} A & B \\ 0 & I_s \end{bmatrix}.$$

We can ask two questions:

1. What is the average time needed to hit an absorbing state?
2. What is the probability of hitting an absorbing state?

Hitting Time: Solution

Theorem

Suppose

$$M = \begin{bmatrix} A & B \\ 0 & I_s \end{bmatrix}.$$

Then the vector of average times needed to hit an absorbing state from each nonabsorbing state is given by

$$\boldsymbol{\theta} = (I - A)^{-1}\mathbf{1}_n.$$

(It can be shown that $\rho(A) < 1$ so that $I - A$ is nonsingular.)

Hitting Probability: Solution

Theorem

Suppose

$$M = \begin{bmatrix} A & B \\ 0 & I_s \end{bmatrix}.$$

For each absorbing state a_j, the vector of probabilities of a sample path reaching the state a_j from each nonabsorbing state is given by

$$\mathbf{p}_j = (I - A)^{-1} B_j, \tag{1}$$

where B_j denotes the j-th column of the matrix B.

Outline

1. Markov Processes
   - Markov Processes: Basics
   - Markov Processes with Absorbing States

2. Markov Reward Processes

Markov Reward Process: Definition

Suppose $\{X_t\}_{t\geq 0}$ is a Markov process on $\mathcal{X}$ with state transition matrix A. Suppose that, in addition, there is a reward function $R : \mathcal{X} \to \mathbb{R}$, as well as a "discount" factor $\gamma \in (0, 1)$. Define the expected discounted future reward $V(x_i)$ as

$$V(x_i) = E\left[\sum_{t=0}^{\infty} \gamma^t R(X_t) | X_0 = x_i\right].$$

The sum is convergent because $\gamma < 1$ and $\mathcal{X}$ is finite. Note: Even if R is random but bounded, the sum would still converge.

Question: How can we compute $V(x_i)$ for each state x_i?

Recursive Relationship for Expected Discounted Reward

Define the vectors

$$\mathbf{v} = [\ V(x_1) \quad \cdots \quad V(x_n)\ ]^\top,$$

$$\mathbf{r} = [\ R(x_1) \quad \cdots \quad R(x_n)\ ]^\top.$$

Theorem

The vector **v** *satisfies the recursive relationship*

$$\mathbf{v} = \mathbf{r} + \gamma A \mathbf{v}.$$

Some Generalizations

If the reward function is random, then above relationship still holds, with $\mathbf{r}$ defined as

$$\mathbf{r} = [E[R(x_1)] \cdots E[R(x_n)]].$$

If the reward is paid at the next time instant, then $\mathbf{r}$ is defined as

$$\mathbf{r} = [r_1 \cdots r_n],$$

where

$$r_i = E[R(X_1)|X_0 = x_i].$$

Computing V

Note that $\rho(A) = 1$, so that $\rho(\gamma A) = \gamma < 1$. So we could write

$$\mathbf{v} = (I - \gamma A)^{-1}\mathbf{r}.$$

But the complexity would be $O(n^3)$. Is there another way?

Contraction Mapping Theorem

Theorem

Suppose $f : \mathbb{R}^n \to \mathbb{R}^n$ *and that there exists a constant* $\rho < 1$ *such that*

$$\|f(x) - f(y)\| \leq \rho \|x - y\|, \ \forall x, y \in \mathbb{R}^n,$$

where $\|\cdot\|$ *on* $\mathbb{R}^n$*. Then there is a unique* $x^* \in \mathbb{R}^n$ *such that*

$$f(x^*) = x^*.$$

To find x^**, choose an arbitrary* $x_0 \in \mathbb{R}^n$ *and define* $x_{l+1} = f(x_l)$*. Then* $\{x_l\} \to x^*$ *as* $l \to \infty$*. Moreover, we have the explicit estimate*

$$\|x^* - x_l\| \leq \frac{\rho^l}{1 - \rho} \|x_1 - x_0\|.$$

Computing V by Value Iteration

Theorem

The map $\mathbf{y} \mapsto T\mathbf{y} := \mathbf{r} + \gamma A\mathbf{y}$ is monotone and is a contraction with constant γ.

Therefore, if we choose $\mathbf{y}_0$ as we wish, and define $\{\mathbf{y}_i\}$ by

$$\mathbf{y}_{i+1} = T\mathbf{y}_i = \mathbf{r} + \gamma A\mathbf{y}_i,$$

then

$$\|\mathbf{y}_{i+1} - \mathbf{y}_i\|_\infty \le \gamma\|\mathbf{y}_i - \mathbf{y}_{i-1}\|_\infty.$$

So $\mathbf{y}_i \to \mathbf{x}^*$, and for each l, we have

$$\|\mathbf{v} - \mathbf{y}_l\| \le \frac{\gamma^l}{1-\gamma}\|\mathbf{y}_1 - \mathbf{y}_0\|.$$

How Many Iterations?

Define the initial error as

$$c := \|\mathbf{y}^1 - \mathbf{y}^0\|_\infty = \|\mathbf{r} + \gamma A\mathbf{y}^0 - \mathbf{y}^0\|_\infty .$$

Then, to ensure that $\|\mathbf{y}^L - \mathbf{v}\|_\infty \le \epsilon$, it is enough to perform

$$L = \left\lceil \frac{1}{1-\gamma} \log \frac{c}{\epsilon(1-\gamma)} \right\rceil$$

iterations. Complexity of $O(Ln^2)$ versus $O(n^3)$.

Note that L does not depend on n.

The Case of Nonnegative Rewards

The map T is monotone. So if $\mathbf{y}^1 \le \mathbf{y}^2$, then $T\mathbf{y}^1 \le T\mathbf{y}^2$ where the inequality is componentwise.

Hence, if we can choose $\mathbf{y}_0$ such that $\mathbf{y}_1 = T\mathbf{y}_0 \ge \mathbf{y}_0$, then $T\mathbf{y}_1 = T^2\mathbf{y}_0 \ge T\mathbf{y}_0 \ge \mathbf{y}_0$. Therefore $\mathbf{y}_i \uparrow \mathbf{v}^*$.

Sufficient Condition: If $\mathbf{r} \ge \mathbf{0}$, and we choose $\mathbf{y}_0 = \mathbf{r}$, then $\mathbf{y}_i \uparrow \mathbf{v}^*$.