STATS 701 – Theory of Reinforcement Learning
Markov Reward Processes, Part 2

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

Outline

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 $\mathbf{v}$ *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

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

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

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

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 $\|\mathbf{y}^L - \mathbf{v}\|_\infty \leq \epsilon$, 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 $\mathbf{y}_i \uparrow \mathbf{v}^*$.

Outline

Average 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 $R : \mathcal{X} \to \mathbb{R}$ (no discount factor now)

Define the average reward w.r.t. an initial state distribution ϕ as

$$c^\star := \lim_{T \to \infty} \frac{1}{T} E \left[\sum_{t=0}^{T} R(X_t) | X_0 \sim \phi \right]$$

Question: Does $c^\star$ depend on ϕ? How can we compute it?

Average cost in terms of stationary distribution

Note that $X_t \sim \phi A^t$ and therefore $E[R(X_t)|X_0 \sim \phi] = \phi A^t \mathbf{r}$

Suppose A is irreducible with (unique) stationary distribution $\boldsymbol{\mu}$

$$
\begin{aligned}
c^\star &:= \lim_{T\to\infty} \frac{1}{T} E\left[\sum_{t=0}^{T} R(X_t)|X_0 \sim \phi\right] \\
&= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \phi A^t \mathbf{r} \\
&= \phi \left(\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} A^t\right) \mathbf{r} \\
&= \phi \mathbf{1}_n \boldsymbol{\mu} \mathbf{r} = \boldsymbol{\mu} \mathbf{r}
\end{aligned}
$$

$c^\star$ is independent of ϕ under irreducibility

Bias or transient reward

Recall the recursive relationship $\mathbf{v} = \mathbf{r} + \gamma A \mathbf{v}$ for discounted MRPs

In order to derive an analogue for average reward MPs, assume the process is primitive (which is the same as irreducible and aperiodic)

Define the bias or transient reward

$$J_i^\star := \sum_{t=0}^{\infty} \left(E[R(X_t) | X_0 = x_i] - c^\star \right)$$

Note no discounting — not clear if this is even well defined!

Bias or transient reward in vector form

We defined the bias or transient reward

$$J_i^\star := \sum_{t=0}^{\infty} (E[R(X_t)|X_0 = x_i] - c^\star)$$

Note that if $X_0 \sim \mathbf{e}_i^\top$ then $X_t \sim \mathbf{e}_i^\top A^t$. Therefore

$$J_i^\star = \sum_{t=0}^{\infty} (\mathbf{e}_i^\top A^t \mathbf{r} - c^\star)$$

which in vector notation becomes (using $A^t \mathbf{1}_n = \mathbf{1}_n$)

$$\mathbf{J}^\star = \sum_{t=0}^{\infty} (A^t \mathbf{r} - c^\star \mathbf{1}_n) = \sum_{t=0}^{\infty} A^t (\mathbf{r} - c^\star \mathbf{1}_n)$$

Why is bias well defined?

By aperiodicity, $\lambda = 1$ is the only eigenvalue of magnitude 1

Recall that $\boldsymbol{\mu}, \mathbf{1}_n$ are left, right eigenvectors for $\lambda = 1$

So $A_2 = A - \mathbf{1}_n \boldsymbol{\mu} =: A - M$ has the same spectrum as A except that the eigenvalue at 1 is replaced by 0

Since $\rho(A_2) < 1$, we have

$$\sum_{t=0}^{\infty} A_2^t = (1 - A_2)^{-1} = (I - A + M)^{-1}$$

Why is bias well defined? Contd.

Let $\mathbf{u} = \mathbf{r} - c^\star \mathbf{1}_n$

Note that $\boldsymbol{\mu}\mathbf{u} = \boldsymbol{\mu}\mathbf{r} - c^\star \boldsymbol{\mu}\mathbf{1}_n = c^\star - c^\star = 0$

Therefore, $A_2\mathbf{u} = (A - \mathbf{1}_n\boldsymbol{\mu})\mathbf{u} = A\mathbf{u} - \mathbf{1}_n 0 = A\mathbf{u}$

Note that $\boldsymbol{\mu}A\mathbf{u} = \boldsymbol{\mu}\mathbf{u}$ is also 0

Thus, $A^2\mathbf{u} = A(A\mathbf{u}) = A_2(A\mathbf{u}) = A_2 A_2 \mathbf{u} = A_2^2\mathbf{u}$

$\vdots$

$\forall t \geq 0$, $A^t\mathbf{u} = A_2^t\mathbf{u}$ and $\boldsymbol{\mu}A^t\mathbf{u} = 0$

Why is bias well defined? Contd.

$$\mathbf{J}^\star = \sum_{t=0}^{\infty} A^t(\mathbf{r} - c^\star \mathbf{1}_n)$$
$$= \sum_{t=0}^{\infty} A_2^t(\mathbf{r} - c^\star \mathbf{1}_n)$$
$$= (I - A + M)^{-1}(\mathbf{r} - c^\star \mathbf{1}_n)$$

Observe that

$$\boldsymbol{\mu} \mathbf{J}^\star = \sum_{t=0}^{\infty} \boldsymbol{\mu} A^t(\mathbf{r} - c^\star \mathbf{1}_n) = 0$$

A recursive relation

$$
\begin{aligned}
J_i^\star &:= \sum_{t=0}^{\infty} \left(E[R(X_t)|X_0 = x_i] - c^\star \right) \\
&= R(x_i) - c^\star + \sum_{t=1}^{\infty} \left(E[R(X_t)|X_0 = x_i] - c^\star \right) \\
&= r_i - c^\star + \sum_{j=1}^{n} a_{ij} \sum_{t=1}^{\infty} \left(E[R(X_t)|X_i = x_j] - c^\star \right) \\
&= r_i - c^\star + \sum_{j=1}^{n} a_{ij} J_j^\star
\end{aligned}
$$

or, in vector notation,

$$\mathbf{J}^\star = \mathbf{r} - c^\star \mathbf{1}_n + A\mathbf{J}^\star$$

Uniqueness

The "Poisson equation"

$$\mathbf{J} = \mathbf{r} - c^\star \mathbf{1}_n + A\mathbf{J}$$

does not have a unique solution: if $\mathbf{J}$ is a solution then so is $\mathbf{J} + \alpha \mathbf{1}_n$

Turns out the only solution of the Poisson equation that also satisfies $\boldsymbol{\mu}\mathbf{J} = 0$ is $\mathbf{J}^\star$