Markov Decision Processes and Performance Metrics

Shangtong Zhang

University of Virginia

Discrete-time MDP

Markov chains:

$$S_0, S_1, S_2, \ldots$$

MDP:

$$S_0, A_0, R_1, S_1, A_1, R_1, \ldots$$

Discrete-time MDP

- State space $\mathcal{S}$
- Action space $\mathcal{A}$
- A Markov Policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$

$$\pi(s), \pi(a|s), \pi(\cdot|s)$$

- Reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- Transition function $p : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$

$$p(s|s, a), p(\cdot|s, a)$$

- Initial distribution $p_0 \in \mathcal{P}(\mathcal{S})$

Infinite-horizon discrete-time MDP

$S_0 \sim p_0(\cdot)$
For $t = 0, 1, 2, \ldots$

- $A_t \sim \pi(\cdot | S_t)$
- $R_{t+1} \doteq r(S_t, A_t)$
- $S_{t+1} \sim p(\cdot | S_t, A_t)$

A policy in an MDP induces a Markov chain

Performance metrics

- Total rewards
- Average reward

Total rewards

For a discount factor $\gamma \in [0, 1]$,

$$
\begin{aligned}
J_{\pi,\gamma} &\doteq \mathbb{E}\left[R_1 + \gamma R_2 + \gamma^2 R_3 + \cdots \mid p_0, \pi, p, r\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} R_t\right]
\end{aligned}
$$

- $\gamma < 1$: discounted total rewards
- $\gamma = 1$: undiscounted total rewards

Return

$$
\begin{aligned}
G_t &\doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \\
&= \sum_{i=1}^{\infty} \gamma^{i-1} R_{t+i} \\
J_{\pi,\gamma} &= \mathbb{E}\left[G_0 \mid p_0, \pi, p, r\right]
\end{aligned}
$$

Value functions

State-value function $v_\pi : \mathcal{S} \to \mathbb{R}$

$$v_\pi(s) \doteq \mathbb{E}\left[G_t \mid S_t = s, \pi, p, r\right]$$

Action-value function $q_\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

$$q_\pi(s, a) \doteq \mathbb{E}\left[G_t \mid S_t = s, \underline{A_t = a}, \pi, p, r\right]$$

Law of total expectation:

$$\begin{aligned} v_\pi(s) &= \mathbb{E}\left[G_t \mid S_t = s\right] \\ &= \mathbb{E}_{A_t}\left[\mathbb{E}\left[G_t \mid S_t = s, A_t\right]\right] \\ &= \sum_a \pi(a|s) \mathbb{E}\left[G_t \mid S_t = s, A_t = a\right] \\ &= \sum_a \pi(a|s) q_\pi(s, a) \end{aligned}$$

Bellman equations

$$
\begin{aligned}
& v_\pi(s) \\
=& \mathbb{E}\left[G_t | S_t = s\right] \\
=& \mathbb{E}\left[R_{t+1} + \gamma G_{t+1} | S_t = s\right] \\
=& \sum_a \pi(a|s) \mathbb{E}\left[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a\right] \\
=& \sum_a \pi(a|s) r(s,a) + \gamma \sum_a \pi(a|s) \mathbb{E}\left[G_{t+1} | S_t = s, A_t = a\right] \\
=& r_\pi(s) + \gamma \sum_{a,s'} \pi(a|s) p(s'|s,a) \mathbb{E}\left[G_{t+1} | S_t = s, A_t = a, S_{t+1} = s'\right] \\
\\
=& r_\pi(s) + \gamma \sum_{a,s'} \pi(a|s) p(s'|s,a) v_\pi(s')
\end{aligned}
$$

Bellman equations

$$q_\pi(s,a) = r(s,a) + \sum_{s',a'} p(s'|s,a)\pi(a'|s')q_\pi(s',a')$$

Vector forms of Bellman equations

$v_\pi \in \mathbb{R}^{|\mathcal{S}|}$, $r_\pi \in \mathbb{R}^{|\mathcal{S}|}$
$P_\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$:

$$P_\pi(s, s') = \sum_a \pi(a|s)p(s'|s, a)$$

$$v_\pi = r_\pi + \gamma P_\pi v_\pi$$

v_π is the unique v satisfying

$$v = r_\pi + \gamma P_\pi v$$

Neumann series

If $\rho(X) < 1$, then

$$\sum_{t=0}^{\infty} X^t = (I - X)^{-1}$$

Vector forms of Bellman equations

$$
\begin{aligned}
v_\pi =& (I - \gamma P_\pi)^{-1} r_\pi \\
v_\pi =& (I + \gamma P_\pi + \gamma^2 P_\pi^2 + \dots) r_\pi
\end{aligned}
$$

Bellman operator

$$\mathcal{T}_\pi v \doteq r_\pi + \gamma P_\pi v$$

Contraction mapping

A map $\mathcal{T} : \mathcal{X} \to \mathcal{X}$ is a contraction mapping on $\mathcal{X}$ if there exists some $\gamma \in (0, 1)$ such that $\forall x, x'$

$$\|\mathcal{T}(x) - \mathcal{T}(x')\| \leq \gamma \|x - x'\|$$

Banach fixed-point theorem

Let $\mathcal{X}$ be a non-empty complete space with a norm $\|\cdot\|$. Let $\mathcal{T}$ be a contraction mapping on $\mathcal{X}$. Then there exists a unique $x_* \in \mathcal{X}$ such that

$$\mathcal{T}(x_*) = x_*.$$

Furthermore, for any x,

$$\lim_{n\to\infty} \mathcal{T}^{(n)}(x) = x_*$$

Contraction of the Bellman operator

$\mathcal{T}_\pi$ is a γ-contraction w.r.t. $\|\cdot\|_\infty$

$$
\begin{aligned}
&\left|(\mathcal{T}_\pi v)(s) - (\mathcal{T}_\pi v')(s)\right| \\
=&\gamma \sum_{a,s'} \pi(a|s)p(s'|s,a)\left|v(s) - v(s')\right| \\
\leq&\gamma \sum_{a,s'} \pi(a|s)p(s'|s,a) \max_z \left|v(z) - v'(z)\right| \\
=&\gamma \max_s \left|v(s) - v'(s)\right| \\
=&\gamma \left\|v - v'\right\|_\infty
\end{aligned}
$$

Vector forms of Bellman equations

$q_\pi \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$, $r \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$
$P \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}\times\mathcal{A}|}$:

$$P((s,a),(s',a')) = p(s'|s,a)\pi(a'|s')$$
$$q_\pi = r + \gamma P q_\pi$$

Optimal policy

A policy π_* is called an optimal policy if $\forall s, \pi$

$$v_{\pi_*}(s) \geq v_\pi(s)$$

Does an optimal policy always exist?

Existence of the optimal value function

Bellman operator

$$(\mathcal{T}_\pi v)(s) = \sum_a \pi(a|s) \left(r(s,a) + \gamma \sum_{s'} p(s'|s,a) v(s') \right)$$

Bellman optimality operator:

$$(\mathcal{T}_* v)(s) = \max_a \left(r(s,a) + \gamma \sum_{s'} p(s'|s,a) v(s') \right)$$

Contraction of the Bellman optimality operator

$$\begin{aligned}\max_x f(x) - \max_x g(x) =& f(x_0) - \max_x g(x)\\ \leq& f(x_0) - g(x_0) \leq \max_x |f(x) - g(x)|\end{aligned}$$

Let v_* denote the unique fixed point of $\mathcal{T}_*$

Existence of the optimal value function

$$
\begin{aligned}
v_\pi &= \mathcal{T}_\pi v_\pi \preceq \mathcal{T}_* v_\pi \\
\mathcal{T}_\pi v_\pi &\preceq \mathcal{T}_\pi \mathcal{T}_* v_\pi \\
v_\pi &\preceq \mathcal{T}_*^{(2)} v_\pi \\
&\cdots \\
v_\pi &\preceq v_*
\end{aligned}
$$

Existence of an optimal policy

$$\pi_{v_*}(a|s) = \begin{cases} 1, & a = \arg\max_b \left(r(s,b) + \gamma \sum_{s'} p(s'|s,b) v_*(s')\right) \\ 0, & \quad \text{otherwise} \end{cases}$$

$$\begin{aligned} v_{\pi_{v_*}} &= \mathcal{T}_{\pi_{v_*}} v_{\pi_{v_*}} \\ v_* &= \mathcal{T}_* v_* = \mathcal{T}_{\pi_{v_*}} v_* \\ \implies v_{\pi_*} &= v_* \end{aligned}$$

Optimal action-value function

$$
\begin{aligned}
q_*(s,a) \geq& q_\pi(s,a)\ \forall \pi, s, a \\
q_*(s,a) =& r(s,a) + \gamma \sum_{s'} p(s'|s,a) \max_{a'} q_*(s',a') \\
(\mathcal{T}_* q)(s,a) =& r(s,a) + \gamma \sum_{s'} p(s'|s,a) \max_{a'} q(s',a') \\
v_*(s) =& \max_a q_*(s,a) \\
\pi_*(a|s) \doteq& \begin{cases} 1, & a = \arg\max_b q_*(s,b) \\ 0, & \text{otherwise} \end{cases}
\end{aligned}
$$

Two fundamental tasks

- Prediction (Policy Evaluation)
  Given π, estimate $J_{\pi,\gamma}$, v_π, q_π
- Control
  Find π_*, v_*, q_*

Average reward (gain)

$$\bar{J}_\pi \doteq \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} R_t \mid p_0, \pi, p, r\right]$$

If the Markov chain induced by π is ergodic, then $\bar{J}_\pi$ is independent of p_0

Alternative form of average reward

$$
\begin{aligned}
\bar{J}_\pi =& \mathbb{E}\left[\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} r_\pi(S_t)\right] = \mathbb{E}_{s\sim d_\pi(s)}\left[r_\pi(s)\right] \\
=& d_\pi^\top r_\pi
\end{aligned}
$$

Differential value function (bias)

Expected total differences between the immediate reward and the average reward

$$\bar{v}_\pi(s) \doteq \mathbb{E}\left[\sum_{i=1}^{\infty}\left(R_{t+i} - \bar{J}_\pi\right) \mid S_t = s\right]$$

$$\bar{q}_\pi(s, a) \doteq \mathbb{E}\left[\sum_{i=1}^{\infty}\left(R_{t+i} - \bar{J}_\pi\right) \mid S_t = s, A_t = a\right]$$

The fundamental matrix

$$\bar{v}_\pi(s) \doteq \sum_{i=0}^{\infty} \left(\sum_{s'} P_\pi^i(s, s') r_\pi(s') - \bar{J}_\pi \right)$$

$$\bar{v}_\pi = \sum_{i=0}^{\infty} \left(P_\pi^i r_\pi - P_* r_\pi \right) = \left(\sum_{i=0}^{\infty} \left(P_\pi^i - P_* \right) \right) r_\pi$$

The fundamental matrix

$$
\begin{aligned}
(P_\pi^i - P_*)(I - P_\pi + P_*) &= P_\pi^i - P_\pi^{i+1} \\
\sum_{i=0}^{\infty}(P_\pi^i - P_*)(I - P_\pi + P_*) &= I - P_* \\
H_\pi \doteq \sum_{i=0}^{\infty}(P_\pi^i - P_*) &= (I - P_\pi + P_*)^{-1}(I - P_*) \\
\bar{v}_\pi &= H_\pi r_\pi
\end{aligned}
$$

Decomposition of transition matrix

Let P_π be a finite ergodic chain, then there exists a nonsingular W such that

$$P_\pi = W^{-1} \begin{bmatrix} Q & 0 \\ 0 & 1 \end{bmatrix} W,$$
$$P_* = W^{-1} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} W,$$

where $\sigma(Q) < 1$ and $\sigma(I - P_\pi) = \sigma(I - Q)$

Properties of the fundamental matrix

$$
\begin{aligned}
H_\pi =& W^{-1}\begin{bmatrix}(I-Q)^{-1} & 0\\ 0 & 0\end{bmatrix} W\\
H_\pi P_* =& P_* H_\pi = 0\\
P_* v_\pi =& P_* H_\pi r_\pi = 0
\end{aligned}
$$

Differential Bellman equations

$$
\begin{aligned}
\bar{v}_\pi(s) =& \sum_a \pi(a|s) \left(r(s,a) - \bar{J}_\pi + \sum_{s'} p(s'|s,a) \bar{v}_\pi(s') \right) \\
\bar{v}_\pi =& r_\pi - \bar{J}_\pi 1 + P_\pi \bar{v}_\pi \\
\bar{q}_\pi(s,a) =& r(s,a) - \bar{J}_\pi + \sum_{s',a'} p(s'|s,a) \bar{q}_\pi(s',a') \\
\bar{q}_\pi =& r - \bar{J}_\pi 1 + P_\pi \bar{q}_\pi
\end{aligned}
$$

Solving differential Bellman equations

$$v = r_\pi - J1 + P_\pi v$$

$$\left\{(v, J) \mid v = \bar{v}_\pi + c1, c \in \mathbb{R}, J = \bar{J}_\pi\right\}$$

Solving differential Bellman equations

$$
\begin{aligned}
J\mathbf{1} &= r_\pi + (P_\pi - I)v \\
J &= d_\pi^\top r_\pi \\
(P_\pi - I)(v_1 - v_2) &= 0
\end{aligned}
$$

Discounted and differential value functions

$$v_{\pi,\gamma} = \frac{\bar{J}_\pi}{1-\gamma}1 + \bar{v}_\pi + f(\gamma),$$

where

$$\lim_{\gamma \to 1} f(\gamma) = 0$$

Optimal average reward

$$\bar{J}_* \doteq \sup_{\pi} \bar{J}_{\pi}$$

Existence of an optimal policy

d_π is (Lipschitz) continuous

$$
\begin{aligned}
\bar{J}_\pi =& r_\pi^\top d_\pi \\
\begin{bmatrix} (P_\pi^\top - I) \\ 1^\top \end{bmatrix} d_\pi =& \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
d_\pi =& \left(\begin{bmatrix} (P_\pi^\top - I) \\ 1^\top \end{bmatrix}^\top \begin{bmatrix} (P_\pi^\top - I) \\ 1^\top \end{bmatrix} \right)^{-1} \begin{bmatrix} (P_\pi^\top - I) \\ 1^\top \end{bmatrix}^\top \begin{bmatrix} 0 \\ 1 \end{bmatrix}
\end{aligned}
$$

Optimality equation

$$
\begin{aligned}
v(s) &= \max_a \pi(a|s) \left(r(s,a) - J + \sum_{s'} p(s'|s,a) v(s) \right) \\
0 &= \max_\pi \{ r_\pi - J1 + P_\pi v - v \}
\end{aligned}
$$

Optimality equation identifies the optimal average reward

If (v, J) is a solution, then $J = \bar{J}_*$

$$
\begin{aligned}
J1 \geq & r_\pi + P_\pi v - v \\
J1 \geq & P_\pi r_\pi + P_\pi^2 v - P_\pi v \\
& \dots \\
J1 \geq & P_\pi^{t-1} r_\pi + P_\pi^t v - P_\pi^{t-1} v \\
J1 \geq & \frac{1}{t} \sum_{i=0}^{t-1} P_\pi^i r_\pi + \frac{P_\pi^t v - v}{t} \\
J \geq & J_\pi
\end{aligned}
$$

There is a π such that the equality holds

Existence of solutions

Choose a sequence $\{\gamma_n\}$ such that $\lim_{n\to\infty} \gamma_n = 1$ such that they share the same optimal policy, say μ

$$
\begin{aligned}
0 =& r_\mu + (\gamma_n P_\mu - I) v_{*,\gamma_n} \\
=& \max_{\mu'} \left\{ r_{\mu'} + (\gamma_n P_{\mu'} - I) v_{\mu,\gamma_n} \right\} \\
\geq& r_\pi + (\gamma_n P_\pi - I) v_{\mu,\gamma_n}
\end{aligned}
$$

Existence of solutions

$$
\begin{aligned}
v_{\mu,\gamma_n} =& \frac{\bar{J}_\mu}{1-\gamma_n} 1 + \bar{v}_\mu + f(\gamma_n) \\
(\gamma_n P_\pi - I) v_{\mu,\gamma_n} =& -\bar{J}_\mu 1 + (P_\pi - I)\bar{v}_\mu + f_\pi(\gamma_n) \\
0 \geq& r_\pi - \bar{J}_\mu 1 + (P_\pi - I)\bar{v}_\mu + f_\pi(\gamma_n) \\
0 \geq& \max_\pi \left\{ r_\pi - \bar{J}_\mu 1 + (P_\pi - I)\bar{v}_\mu + f_\pi(\gamma_n) \right\} \\
0 \geq& \max_\pi \left\{ r_\pi - \bar{J}_\mu 1 + (P_\pi - I)\bar{v}_\mu \right\} \\
0 =& \max_\pi \left\{ r_\pi - \bar{J}_\mu 1 + (P_\pi - I)\bar{v}_\mu \right\}
\end{aligned}
$$

Full characterization of solutions

Does the following set contains <u>all</u> solutions?

$$\left\{(J, v) \mid J = \bar{J}_*, v = \bar{v}_\mu + c1\right\}$$

Identifying an optimal policy

Suppose $\bar{J}_*, \bar{v}_*$ satisfy

$$0 = \max_{\pi} \left\{ r_\pi - \bar{J}_* 1 + (P_\pi - I)\bar{v}_* \right\},$$

then $\pi_{\bar{v}_*}$, a policy that is greedy w.r.t. to $\bar{v}_*$, is an optimal policy.

$$\begin{aligned} 0 &= r_{\pi_{\bar{v}_*}} - \bar{J}_* 1 + (P_{\pi_{\bar{v}_*}} - I)\bar{v}_* \\ 0 &= r_{\pi_{\bar{v}_*}} - \bar{J}_{\pi_{\bar{v}_*}} 1 + (P_{\pi_{\bar{v}}} - I)\bar{v}_{\pi_{\bar{v}_*}} \end{aligned}$$

References

- Markov Decision Processes: Discrete Stochastic Dynamic Programming by Martin Puterman
- Neuro-Dynamic Programming by Dimitri Bertsekas