SDU

2) Markov Decision Processes

Melih Kandemir

University of Southern Denmark
Department of Mathematics and Computer Science (IMADA)

What is a Markov Decision Process?

A Markov Decision Process (MDP) is a controlled Markov chain with an attached performance criterion. An MDP is defined as a tuple $M = \langle \mathcal{S}, \mathcal{A}, P, p_0, R \rangle$ where

- $\mathcal{S}$ is a state space as in DDPs.
- $\mathcal{A}$ is an action space as in DDPs.
- $r : \mathcal{S} \times \mathcal{A} \to [0, R_{\max}]$ is a reward function limited to the unit interval for convenience.
- $P : \mathcal{S} \times \mathcal{A} \to \mathcal{P}$ is a state transition distribution where $\mathcal{P}$ is the set of all probability distributions defined on $\mathcal{S}$.
- $p_0 \in \mathcal{P}$ is an initial state distribution.

The state transition distribution is defined as follows:

- $P(s'|s, a)$ where $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$.

The tuple $M = \langle \mathcal{S}, \mathcal{A}, P, p_0 \rangle$ is called a **controlled Markov chain**. The MDP above is defined based on a stationary controlled Markov chain for convenience (i.e. the elements do not depend on time). Its nonstationary counterpart exists.

Induced stochastic process

A control policy $\pi \in \Pi_{HS}$ and an MDP induce a probability distribution over any finite state-action sequence $h_T = (s_0, a_0, \ldots, s_{T-1}, a_{T-1}, s_T)$ given by

$$P(h_T) = p_0(s_0) \prod_{t=0}^{T-1} p_t(s_{t+1}|s_t, a_t)\pi_t(a_t|h_t),$$

where $h_t = (s_0, a_0, \ldots, s_{t-1}, a_{t-1}, s_t)$ because

$$\begin{aligned} P(h_{t+1}) &= P(h_t, a_t, s_{t+1}) \\ &= P(s_{t+1}|h_t, a_t)P(a_t|h_t)P(h_t) \\ &= \underbrace{p_t(s_{t+1}|s_t, a_t)}_{\text{Markov property}} \pi_t(a_t|h_t)P(h_t). \end{aligned}$$

The state-action sequence $h_\infty = (s_k, a_k)_{k\geq 0}$ can now be considered a stochastic process.
We denote the probability law of this stochastic process by $P^\pi(\cdot)$. The corresponding expectation operator is denoted by $\mathbb{E}^\pi[\cdot]$.

Induced stochastic process

Under a Markov control policy, the state sequence $(s_t)_{t\geq 0}$ becomes a Markov chain with transition probabilities:

$$P(s_{t+1} = s'|s_t = s) = \sum_{a\in\mathcal{A}_t} p_t(s'|s,a)\pi_t(a|s).$$

This follows since:

$$\begin{aligned} P(s_{t+1} = s'|s_t = s) &= \sum_{a\in\mathcal{A}_t} P(s_{t+1} = s', a|s_t = s) \\ &= \sum_{a\in\mathcal{A}_t} P(s_{t+1} = s'|s_t = s, a)P(a|s_t = s) \\ &= \sum_{a\in\mathcal{A}_t} p_t(s'|s,a)\pi_t(a|s) \end{aligned}$$

If the controlled Markov chain is stationary (time-invariant) and the control policy is stationary, then the induced Markov chain is stationary as well.

Expected return

We need to update the definition of the total reward to the stochastic dynamics case. Then our objective becomes to maximize

$$V_T^\pi(h_t) := \mathbb{E}^\pi \left[\sum_{t=0}^{T} r_t(s_t, a_t) \right]$$

which is called the **expected return**.
One can also define a risk-sensitive return function as below

$$V_{T,\lambda}^\pi(h_t) := \frac{1}{\lambda} \log \mathbb{E}^\pi \left[\exp \left(\lambda \sum_{t=0}^{T} r_t(s_t, a_t) \right) \right]$$

where $\lambda > 0$ gives higher weights to high rewards making the search **risk-seeking** and otherwise when $\lambda < 0$ the search is **risk-averse**.

Infinite horizon problems

Sometimes the system in question is expected to operate for a long time, or a large number of steps, possibly with no specific *closing* time. Infinite horizon problems are most often defined for stationary problems. In that case, optimal policies can be identified.

Discounted return:

$$V_\gamma^\pi(s) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s \right] \equiv \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]$$

where $0 < \gamma < 1$ is the discount factor. The discount factor ensures convergence of the sum.

Average return: Maximize the long-term average return:

$$V_{av}^\pi(s) = \liminf_{T \to \infty} \mathbb{E}^\pi \left[\frac{1}{T} \sum_{t=0}^{T-1} r(s_t, a_t) \right]$$

Sufficiency of Markov policies

In all the performance criteria defined above, the criterion is composed of sums of terms of the form $\mathbb{E}[r_t(s_t, a_t)]$. It follows that if two control policies induce the same marginal probability distributions $q_t(s_t, a_t)$ over the state-action pairs (s_t, a_t) for all $t \geq 0$, they will have the same performance for any linear return function. Using this observation, the next claim implies that it is enough to consider the set of (stochastic) Markov policies in the above planning problems.

Theorem

Let $\pi \in \Pi_{HS}$ be a general (history-dependent, stochastic) control policy. Let

$$p_t^{\pi,s_0}(s, a) = P^{\pi,s_0}(s_t = s, a_t = a), \quad (s, a) \in \mathcal{S}_t \times \mathcal{A}_t$$

Denote the marginal distributions induced by $q_t(s_t, a_t)$ on the state-action pairs (s_t, a_t), for all $t \geq 0$. Then there exists a stochastic Markov policy $\tilde{\pi} \in \Pi_{MS}$ that induces the same marginal probabilities (for all initial states s_0).

Dynamic programming for policy evaluation

Define the following **reward-to-go** function or **value function**:

$$V_k^\pi(s) = \mathbb{E}^\pi \left[\sum_{t=k}^{T} r(s_t, a_t) \middle| s_k = s \right]$$

Lemma (Value Iteration)

$V_k^\pi(s)$ may be computed by the backward recursion:

$$V_k^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} \left[r_k(s,a) + \sum_{s' \in \mathcal{S}_{k+1}} p_k(s'|s,a) V_{k+1}^\pi(s') \right], \quad \forall s \in \mathcal{S}_k$$

for $k = T-1, \ldots, 0$, starting with $V_T^\pi(s) = r_T(s)$.

Dynamic programming for policy optimization

Define the **optimal value function** at each time $k \geq 0$:

$$V_k^*(s) = \max_{\pi^k} \mathbb{E}^{\pi^k} \left[\sum_{t=k}^{T} r(s_k, a_k) \middle| s_k = s \right], \quad s \in \mathcal{S}_k,$$

where the maximum is taken over *tail* policies $\pi^k = (\pi_k, \ldots, \pi_{T-1})$ that start from time k. Note that π^k is allowed to be a general policy, i.e., history-dependent and stochastic.

Dynamic programming for policy optimization

Theorem (Finite-horizon Dynamic Programming)

1. **Backward recursion:** *Set* $V_T(s) = r_T(s)$ *for* $s \in \mathcal{S}_T$. *For* $k = T-1, \ldots, 0$, *compute* $V_k(s)$ *using the following recursion:*

$$V_k(s) = \max_{a \in \mathcal{A}_k} \left[r_k(s,a) + \sum_{s' \in \mathcal{S}_{k+1}} p_k(s'|s,a) V_{k+1}(s') \right], \quad s \in \mathcal{S}_k.$$

We have that $V_k(s) = V_k^*(s)$.

2. **Optimal policy:** *Any Markov policy* π_* *that satisfies, for* $t = 0, \ldots, T-1$,

$$\pi_t^*(s) \in \arg\max_{a \in \mathcal{A}_t} \left[r_t(s,a) + \sum_{s' \in \mathcal{S}_{t+1}} p_t(s'|s,a) V_{t+1}^*(s') \right], \quad \forall s \in \mathcal{S}_t,$$

is an optimal control policy. Furthermore, π_* *maximizes* $V^\pi(s_0)$ *simultaneously for every initial state* $s_0 \in \mathcal{S}_0$.

The Q function

The quantity

$$Q_k^*(s,a) \stackrel{\Delta}{=} r_k(s,a) + \sum_{s' \in \mathcal{S}_k} p_k(s'|s,a) V_{k+1}^*(s')$$

is known as the **optimal state-action value function**, or simply as the **Q-function**. $Q_k^*(s,a)$ is the expected return from stage k onward, if we choose $a_k = a$ and then proceed optimally. The result in the previous slide yields

$$V_k^*(s) = \max_{a \in \mathcal{A}_k} Q_k^*(s,a),$$

and

$$\pi_k^*(s) \in \arg\max_{a \in \mathcal{A}_k} Q_k^*(s,a).$$

The Q function provides the basis for the Q-learning algorithm.

Discounted MDPs

$$V_\gamma^\pi(s) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s \right] \equiv \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]$$

where $\gamma \in (0, 1)$ is a discount factor. Let $V_\gamma^*(s)$ denote the maximal expected value of the discounted return, over all (possibly history dependent and randomized) control policies, i.e.,

$$V_\gamma^*(s) = \sup_{\pi \in \Pi_{HS}} V_\gamma^\pi(s).$$

Our goal is to find an optimal control policy π_* that attains that maximum (for all initial states), and compute the numeric value of the optimal return $V_\gamma^*(s)$. As we shall see, for this problem there always exists an optimal policy which is a (deterministic) stationary policy.

Fixed-policy value function

For a stationary policy $\pi : S \to A$, we define the value function $V^\pi(s)$, $s \in \mathcal{S}$ as the corresponding discounted return:

$$V^\pi(s) \triangleq \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] = V_\gamma^\pi(s), \quad \forall s \in \mathcal{S}$$

Lemma

For $\pi \in \Pi_{SD}$, the value function V^π satisfies the following set of $|\mathcal{S}|$ linear equations:

$$V^\pi(s) = r(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, \pi(s)) V^\pi(s'), \quad \forall s \in \mathcal{S}.$$

Proof

$$V^\pi(s) \triangleq \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s \right] = \mathbb{E}^\pi \left[\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) \middle| s_1 = s \right],$$

since both $p(s'|s,a)$ and π are stationary. Now,

$$\begin{aligned}
V^\pi(s) &= r(s, \pi(s)) + \mathbb{E}^\pi \left[\sum_{t=1}^{\infty} \gamma^t r(s_t, \pi(s_t)) \middle| s_0 = s \right] \\
&= r(s, \pi(s)) + \mathbb{E}^\pi \left[\mathbb{E}^\pi \left[\sum_{t=1}^{\infty} \gamma^t r(s_t, \pi(s_t)) \middle| s_0 = s, s_1 = s' \right] \middle| s_0 = s \right] \\
&= r(s, \pi(s)) + \sum_{s' \in \mathcal{S}} p(s'|s, \pi(s)) \mathbb{E}^\pi \left[\sum_{t=1}^{\infty} \gamma^t r(s_t, \pi(s_t)) \middle| s_1 = s' \right] \\
&= r(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, \pi(s)) \mathbb{E}^\pi \left[\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, \pi(s_t)) \middle| s_1 = s' \right] \\
&= r(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, \pi(s)) V^\pi(s') \qquad \square
\end{aligned}$$

Vectoral view

Define the transition probabilities induced by a policy π as a matrix $[P^\pi]_{ij} := p(s' = j|s = i, \pi(i))$ and rewards as a column vector $[r^\pi] := r(s = i, \pi(i))$.

Lemma

The system of linear equations $V^\pi = r^\pi + \gamma P^\pi V^\pi$ has the unique solution below:

$$V^\pi = (I - \gamma P^\pi)^{-1} r^\pi.$$

Proof. We only need to show that the square matrix $I - \gamma P^\pi$ is non-singular. Let (λ_i) denote the eigenvalues of the matrix P^π. Since P^π is a stochastic matrix (row sums are 1), we have for an eigenvalue λ_i and its eigenvector x_i

$$\|P^\pi x_i\|_\infty = \|\lambda_i x_i\|_\infty = |\lambda_i| \|x_i\|_\infty \leq \|x_i\|_\infty$$

Then $|\lambda_i| \leq 1$. Furthermore, since $P^\pi \mathbf{1} = \mathbf{1}$, we have $\lambda = 1$ as the largest eigenvalue. Now, the eigenvalues of $I - \gamma P^\pi$ are $(1 - \gamma\lambda_i)$, and satisfy $|1 - \gamma\lambda_i| \geq 1 - \gamma > 0$ $\square$

Notation: Here $\mathbf{1}$ is a column vector of ones of an appropriate dimension.

Properties of geometric series

For some $a \neq 0$ and $|\gamma| < 1$ the following three identities hold

$$\sum_{k=0}^{n} a\gamma^k = \frac{a(1-\gamma^{n+1})}{1-\gamma},$$

$$\sum_{k=0}^{\infty} a\gamma^k = \frac{a}{1-\gamma},$$

$$\sum_{k=n}^{\infty} a\gamma^k = \frac{a\gamma^n}{1-\gamma}.$$

Fixed-policy value iteration

Algorithm Fixed-policy Value Iteration

1: $V_0 := (V_0(s))$ for some $s \in \mathcal{S}$.
2: **for** $n = 0, 1, 2, \ldots$ **do**
3: $\quad V_{n+1}(s) := r(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, \pi(s)) V_n(s'), \qquad \forall s \in \mathcal{S}$
4: **end for**

Theorem (Convergence of fixed-policy value iteration)

We have $V_n \to V^\pi$ component-wise, that is,

$$\lim_{n \to \infty} V_n(s) = V^\pi(s), \quad \forall s \in \mathcal{S}.$$

Proof

$$V_1(s) = r(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, \pi(s)) V_0(s')$$
$$= \mathbb{E}^{\pi}[r(s_0, a_0) + \gamma V_0(s_1) | s_0 = s].$$

Continuing similarly, we obtain that

$$V_n(s) = \mathbb{E}^{\pi} \left[\sum_{t=0}^{n-1} \gamma^t r(s_t, a_t) + \gamma^n V_0(s_n) \middle| s_0 = s \right].$$

$V_n(s)$ is the n-stage discounted return with $r_n(s_n) = V_0(s_n)$. Then

$$V^{\pi}(s) - V_n(s) = \mathbb{E}^{\pi} \left[\sum_{t=n}^{\infty} \gamma^t r(s_t, a_t) - \gamma^n V_0(s_n) \middle| s_0 = s \right].$$

Denoting $R_{\max} = \max_{s,a} |r(s, a)|$ and $\bar{V}_0 = \max_s |V_0(s)|$ we obtain

$$|V^{\pi}(s) - V_n(s)| \leq \gamma^n \left(\frac{R_{\max}}{1 - \gamma} + \bar{V}_0 \right) \qquad \text{which converges to 0 since } \gamma < 1 \quad \square$$

Optimal planning

$$V_\gamma^*(s) = \sup_{\pi \in \Pi_{HS}} V_\gamma^\pi(s)$$

is the optimal discounted return for state s. Let us denote

$$V^*(s) \stackrel{\Delta}{=} V_\gamma^*(s), \quad \forall s \in \mathcal{S},$$

Algorithm Value Iteration (VI)

1: $V_0 := (V_0(s))$ for some $s \in \mathcal{S}$.
2: **for** $n = 0, 1, 2, \ldots$ **do**
3: $\quad V_{n+1}(s) := \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) V_n(s') \right\}, \quad \forall s \in \mathcal{S}$
4: **end for**

Convergence

Theorem (Convergence of value iteration)

We have $\lim_{n\to\infty} V_n = V^*$ *(component-wise). The rate of convergence is exponential at rate* $O(\gamma^n)$.

Proof. The one in the literature does not make sense. Get back later.

Policy iteration

Algorithm Policy Iteration (PI)

1: Choose some stationary π_0.
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: $\quad V^{\pi_k} := (I - \gamma P^{\pi_k})^{-1} r^{\pi_k}$ (Policy evaluation)
4: $\quad \forall s \in \mathcal{S}$ do the following (Policy improvement)
5: $\quad \pi_{k+1}(s) \in \arg\max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) V^{\pi_k}(s') \right\}$.
6: $\quad$ Stop if $\pi_{k+1} = \pi_k$.
7: **end for**

Convergence

Theorem (Convergence of policy iteration)

The following statements hold:

1. *Each policy π_{k+1} is improving over the previous one π_k, in the sense that $V^{\pi_{k+1}} \geq V^{\pi_k}$ (component-wise).*
2. *$V^{\pi_{k+1}} = V^{\pi_k}$ if and only if π_k is an optimal policy.*
3. *Consequently, since the number of stationary policies is finite, π_k converges to the optimal policy after a finite number of steps.*

Contraction operators

A norm $\|\cdot\|$ over $\mathbb{R}^n$ is a real-valued function $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}^+$ such that, for any pair of vectors $x, y \in \mathbb{R}^d$ and scalar $a \in \mathbb{R}$,

1. $\|ax\| = |a| \cdot \|x\|$,
2. $\|x + y\| \leq \|x\| + \|y\|$,
3. $\|x\| = 0$ only if $x = 0$.

Examples:

- The p-norm $\|x\|_p = (\sum_{i=1}^{d} |x_i|^p)^{1/p}$ for $p \geq 1$ (Euclidean if $p = 2$)
- Max norm $\|x\|_\infty = \max_{1 \leq i \leq d} |x_i|$

Definition

The operator $T : \mathbb{R}^d \to \mathbb{R}^d$ is called a contraction operator if there exists $\beta \in (0, 1)$ (the contraction coefficient) such that

$$\|T(v_1) - T(v_2)\| \leq \beta \|v_1 - v_2\|,$$

for all $v_1, v_2 \in \mathbb{R}^d$. Similarly, such operator T is called a β-contraction operator.

Banach's fixed point theorem

Theorem

Let $T : \mathbb{R}^d \to \mathbb{R}^d$ be a contraction operator. Then

1. *The equation $T(v) = v$ has a unique solution $V^* \in \mathbb{R}^d$.*
2. *For any $v_0 \in \mathbb{R}^d$, $\lim_{n\to\infty} T^n(v_0) = V^*$. In fact, $\|T^n(v_0) - V^*\| \leq O(\beta^n)$, where β is the contraction coefficient.*

Dynamic programming operators

Definition

For a fixed stationary policy $\pi : \mathcal{S} \to \mathcal{A}$, define the **Bellman Operator** $T^\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ as follows: For any $V = (V(s)) \in \mathbb{R}^{|\mathcal{S}|}$,

$$(T^\pi(V))(s) = r(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, \pi(s))V(s'), \quad \forall s \in \mathcal{S}.$$

In column-vector notation $T^\pi(V) = r^\pi + \gamma P^\pi V$.

Definition

Define the **Bellman Optimality Operator** $T^* : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ as follows: For any $V = (V(s)) \in \mathbb{R}^{|\mathcal{S}|}$,

$$(T^*(V))(s) = \max_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a)V(s') \right\}, \quad \forall s \in \mathcal{S}.$$

In column-vector notation $T * (V) = \max_\pi \{r^\pi + \gamma P^\pi V\}$.

Bellman's Optimality Equation Properties

Let $\|V\|_\infty = \max_{s \in \mathcal{S}} |V(s)|$ denote the max-norm of V.

Theorem (Contraction property)

The following statements hold:

1. *T^π is a γ-contraction operator with respect to the max-norm, namely $\|T^\pi(V_1) - T^\pi(V_2)\|_\infty \leq \gamma \|V_1 - V_2\|_\infty$ for all $V_1, V_2 \in \mathbb{R}^{|\mathcal{S}|}$.*
2. *Similarly, T^* is a γ-contraction operator with respect to the max-norm.*

Proof (Claim 1)

Fix V_1, V_2. For every state s

$$\begin{aligned}
\forall s, |(T^\pi(V_1))(s) - (T^\pi(V_2))(s)| &= \left|\gamma \sum_{s' \in \mathcal{S}} p(s'|s, \pi(s))[V_1(s') - V_2(s')]\right| \\
&\leq \gamma \sum_{s' \in \mathcal{S}} p(s'|s, \pi(s))|V_1(s') - V_2(s')| \\
&\leq \gamma \sum_{s' \in \mathcal{S}} p(s'|s, \pi(s))\|V_1 - V_2\|_\infty \\
&= \gamma \|V_1 - V_2\|_\infty
\end{aligned}$$

Proof (Claim 2)

$$|T^*(V_1)(s) - T^*(V_2)(s)| \leq \gamma \|V_1 - V_2\|_\infty.$$

Fix the state s and check separately the positive and negative parts of the absolute value:

(a) $T^*(V_1)(s) - T^*(V_2)(s) \leq \gamma \|V_1 - V_2\|_\infty$: Let $\bar{a}$ denote an action that attains the maximum in $T^*(V_1)(s)$, namely

$$\bar{a} \in \arg\max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) V_1(s') \right\}.$$

Then,

$$T^*(V_1)(s) = r(s,\bar{a}) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,\bar{a}) V_1(s'),$$

and

$$T^*(V_2)(s) \geq r(s,\bar{a}) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,\bar{a}) V_2(s').$$

Proof (Claim 2) cont'd

Since the same action $\bar{a}$ appears in both expressions, we can now continue to show the inequality (a) similarly to 1:

$$\begin{aligned}(T^*(V_1))(s) - (T^*(V_2))(s) &\leq \gamma \sum_{s' \in \mathcal{S}} p(s'|s, \bar{a})(V_1(s') - V_2(s')) \\ &\leq \gamma \sum_{s' \in \mathcal{S}} p(s'|s, \bar{a}) \|V_1 - V_2\|_\infty = \gamma \|V_1 - V_2\|_\infty.\end{aligned}$$

(b) $T^*(V_2)(s) - T^*(V_1)(s) \leq \gamma \|V_1 - V_2\|_\infty$: Similarly to (a) we have

$$T^*(V_2)(s) - T^*(V_1)(s) \leq \gamma \|V_2 - V_1\|_\infty = \gamma \|V_1 - V_2\|_\infty.$$

(a) and (b) imply $|T^*(V_1)(s) - T^*(V_2)(s)| \leq \gamma \|V_1 - V_2\|_\infty$. Since this holds for any state s, we get $\|T^*(V_1) - T^*(V_2)\|_\infty \leq \gamma \|V_1 - V_2\|_\infty$.

Monotonicity of state transitions

One can treat taking the expectation of the value function with respect to the transition probability matrix as yet another operator: $P(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$. This operator is monotonic.

Lemma

For any $V_1, V_2 \in \mathbb{R}^d$ and any stochastic matrix P (i.e. rows sum up to one), if $V_1 \geq V_2$ then $PV_1 \geq PV_2$

Proof. $PV1 - PV_2 = P(V_1 - V_2) \geq 0$ as both factors of the last term are greater than zero, first because P is a probability matrix and second by assumption

Monotonicity of the Bellman operators

Notation: Denote vectors $V_1, V_2 \in \mathcal{R}^{|\mathcal{S}|}$ by $V_1 \geq V_2$ if $V_1(s) \geq V_2(s), \forall s \in \mathcal{S}$.

Lemma

If $V_1 \geq V_2$ then $T^\pi(V_1) \geq T^\pi(V_2)$ for all π and $T^*(V_1) \geq T^*(V_2)$.

Proof.

$$\begin{aligned} T^\pi V_1 - T^\pi V_2 &= (r^\pi + \gamma P^\pi V_1) - (r^\pi + \gamma P^\pi V_2) \\ &= \gamma(P^\pi V_1 - P^\pi V_2) \geq 0. \end{aligned}$$

The result for $T^*(V_1) \geq T^*(V_2)$ follows in the same way.

Bellman's Optimality Equation Properties

Theorem (Bellman's Optimality Equation)

The following statements hold:

1. *V^* is the unique solution of the following set of (nonlinear) equations:*

$$V(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) V(s') \right\}, \quad \forall s \in \mathcal{S}.$$

2. *Any stationary policy π_* that satisfies*

$$\pi_*(s) \in \arg\max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) V(s') \right\}, \quad \forall s \in \mathcal{S},$$

is an optimal policy (for any initial state $s_0 \in \mathcal{S}$).

Proof

1. As T^* is a contraction operator, there exists a unique $\widehat{V}$ s.t. $\widehat{V} = T^*(\widehat{V})$ due to the Banach fixed point theorem. $V^* \geq V \Rightarrow P^\pi V^* \geq P^\pi V$ for any π and $V \in \mathcal{V}$ where $\mathcal{V} = \{V : V^\pi, \forall \pi\}$. Then we have

$$V^* = \max_\pi \max_{V \in \mathcal{V}} \{r^\pi + \gamma P^\pi V\} = \max_\pi \{r^\pi + \gamma P^\pi V^*\} = T^*(V^*)$$

where the first step follows from the monotonicity of P^π and the second by the definition of T^*. Since the fixed point is unique, we get $\widehat{V} = V^*$.

2. By definition of π_* we have

$$T^{\pi_*}(V^*) = T^*(V^*) = V^*,$$

where the last equality follows from part 1. Thus the optimal value function satisfies the equation $T^{\pi_*}(V^*) = V^*$. Since V^{π_*} is the unique solution of that equation, we have $V^{\pi_*} = V^*$. This implies that π_* achieves the optimal value for any initial state.

Error bounds

Describe value iteration as $V_{n+1} = T^*(V_n)$. Note that value iteration does $O(|\mathcal{S}|^2 \cdot |\mathcal{A}|)$ computations. Let us have an observation-based criterion on how far we are to the fixed point.

Lemma

For π_{n+1} which is greedy with respect to V_{n+1}, we have

$$\|V^{\pi_{n+1}} - V^*\| \leq \frac{2\gamma}{1-\gamma}\|V_{n+1} - V_n\|$$

Proof

We consider the following:

$$\|V^{\pi_{n+1}} - V^*\| \leq \|V^{\pi_{n+1}} - V_{n+1}\| + \|V_{n+1} - V^*\|.$$

We now bound each part of the sum separately:

$$\begin{aligned}\|V^{\pi_{n+1}} - V_{n+1}\| &= \|T^{\pi_{n+1}}(V^{\pi_{n+1}}) - V_{n+1}\| \\ &\leq \|T^{\pi_{n+1}}(V^{\pi_{n+1}}) - T^*(V_{n+1})\| + \|T^*(V_{n+1}) - V_{n+1}\|.\end{aligned}$$

Since π_{n+1} is maximal over the actions using V_{n+1}, it implies that $T^{\pi_{n+1}}(V_{n+1}) = T^*(V_{n+1})$ and we conclude that:

$$\begin{aligned}\|V^{\pi_{n+1}} - V_{n+1}\| \leq& \\ &\|T^{\pi_{n+1}}(V^{\pi_{n+1}}) - T^{\pi_{n+1}}(V_{n+1})\| + \|T^*(V_{n+1}) - T^*(V_n)\| \\ &\leq \gamma\|V^{\pi_{n+1}} - V_{n+1}\| + \gamma\|V_{n+1} - V_n\|\end{aligned}$$

Rearranging, this implies that,

$$\|V^{\pi_{n+1}} - V_{n+1}\| \leq \frac{\gamma}{1-\gamma}\|V_{n+1} - V_n\|$$

Proof (cont'd)

For the second part of the sum we derive similarly that:

$$\begin{aligned}\|V_{n+1} - V^*\| &\leq \|V_{n+1} - T^*(V_{n+1})\| + \|T^*(V_{n+1}) - V^*\| \\ &= \|T^*(V_n) - T^*(V_{n+1})\| + \|T^*(V_{n+1}) - T^*(V^*)\| \\ &\leq \gamma\|V_n - V_{n+1}\| + \gamma\|V_{n+1} - V^*\|,\end{aligned}$$

and therefore

$$\|V_{n+1} - V^*\| \leq \frac{\gamma}{1-\gamma}\|V_{n+1} - V_n\|.$$

Combining the two results, we get:

$$\|V^{\pi_{n+1}} - V^*\| \leq \frac{2\gamma}{1-\gamma}\|V_{n+1} - V_n\|.$$

For any ε greater than the r.h.s. of the inequality above, π_{n+1} is $\varepsilon-$optimal.

Policy iteration

- Introduced by Howard, Dynamic Programming and Markov Processes, MIT Press, 1960
- Converges in $O(|\mathcal{S}|)$ steps.
- π-improving policy is defined as

$$\bar{\pi}(s) \in \arg\max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^{\pi}(s') \right\}, \quad \forall s \in \mathcal{S}.$$

- Each iteration makes $O(|\mathcal{S}|^2|\mathcal{A}| + |\mathcal{S}|^3)$ operations.
- The number of iterations of Value Iteration increases as γ approaches 1, while the number of policies (which upper bound the number of iterations of Policy Iteration) is independent of γ.

Lemma (Policy Improvement)

Let π be a stationary policy and $\bar{\pi}$ be a π-improving policy. We have $V^{\bar{\pi}} \geq V^{\pi}$ (component-wise), and $V^{\bar{\pi}} = V^{\pi}$ if and only if π is an optimal policy.

Proof of the policy improvement lemma

Observe first that

$$V^{\pi} = T^{\pi}(V^{\pi}) \leq T^{*}(V^{\pi}) = T^{\bar{\pi}}(V^{\pi})$$

Since $V_1 \leq V_2 \Rightarrow T^{\pi}(V_1) \leq T^{\pi}(V_2)$, applying $T^{\bar{\pi}}$ repeatedly to both sides of the above inequality $V^{\pi} \leq T^{\bar{\pi}}(V^{\pi})$ gives

$$V^{\pi} \leq T^{\bar{\pi}}(V^{\pi}) \leq (T^{\bar{\pi}})^2(V^{\pi}) \leq \cdots \leq \lim_{n \to \infty} (T^{\bar{\pi}})^n(V^{\pi}) = V^{\bar{\pi}}.$$

We now show that π is optimal if and only if $V^{\bar{\pi}} = V^{\pi}$. We showed that $V^{\bar{\pi}} \geq V^{\pi}$. If $V^{\bar{\pi}} > V^{\pi}$ then clearly π is not optimal. Assume that $V^{\bar{\pi}} = V^{\pi}$. We have the following identities:

$$V^{\pi} = V^{\bar{\pi}} = T^{\bar{\pi}}(V^{\bar{\pi}}) = T^{\bar{\pi}}(V^{\pi}) = T^{*}(V^{\pi}).$$

We have established that: $V^{\pi} = T^{*}(V^{\pi})$, and hence V^{π} and π is a fixed point of T^{*} and therefore by the result on Bellman's optimality equation, policy π is optimal.

Value Iteration (VI) versus Policy Iteration (PI)

Theorem

Let $\{VI_n\}$ be the sequence of values created by the VI algorithm, i.e., $VI_{n+1} = T^*(VI_n)$, and let $\{PI_n\}$ be the sequence of values created by PI algorithm, i.e., $PI_n = V^{\pi_n}$. If $VI_0 = PI_0$, then for all n we have

$$VI_n \leq PI_n \leq V^*.$$

Proof

Induction Basis: By construction $VI_0 = PI_0$. Since $PI_0 = V^{\pi_0}$, it is clearly bounded by V^*.
Induction Step: Assume that $VI_n \leq PI_n$. For VI_{n+1} we have,

$$VI_{n+1} = T^*(VI_n) = T^{\pi'}(VI_n),$$

where π' is the greedy policy w.r.t. VI_n, i.e.,

$$\pi'(s) \in \arg\max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s,a) VI_n(s') \right\}, \quad \forall s \in \mathcal{S}.$$

Since $VI_n \leq PI_n$, and $T^{\pi'}$ is monotonic, it follows that:

$$T^{\pi'}(VI_n) \leq T^{\pi'}(PI_n).$$

Proof cont'd

Since T^* is upper bounding any T^π:

$$T^{\pi'}(PI_n) \leq T^*(PI_n).$$

The policy determined by the PI algorithm in iteration $n+1$ is π_{n+1}, and we have:

$$T^*(PI_n) = T^{\pi_{n+1}}(PI_n).$$

From the definition of π_{n+1}, we have

$$T^{\pi_{n+1}}(PI_n) \leq V^{\pi_{n+1}} = PI_{n+1}.$$

Therefore, $VI_{n+1} \leq PI_{n+1}$. Since $PI_{n+1} = V^{\pi_{n+1}}$, we have $PI_{n+1} \leq V^*$.

Episodic Markov Decision Processes

Let $\mathcal{S}_G \subset \mathcal{S}$ define the **goal states**. The **termination time** is a random variable defined as

$$\tau = \inf\{t \geq 0 : s_t \in \mathcal{S}_G\},$$

the first time in which a goal state is reached, or infinity otherwise.
We assume the state space is finite $|\mathcal{S}| < \infty$ and for any π we have $\tau < \infty$. To simplify the notation, in the following we will assume a single goal state $S_G = s_G$ and that $r_G(s_\tau) = 0$. We therefore write the value of a **stochastic shortest path** problem as

$$V^\pi_{ssp}(s) = \begin{cases} \mathbb{E}^\pi \left[\sum_{t=0}^{\tau} r(s_t, a_t)\right], & s \neq s_G \\ 0, & s = s_G \end{cases}$$

Finite horizon return and discounted infinite return can be shown as special cases of the stochastic shortest path setting.