Reinforcement Learning

Introduction to Artificial Intelligence with Mathematics
Lecture Notes

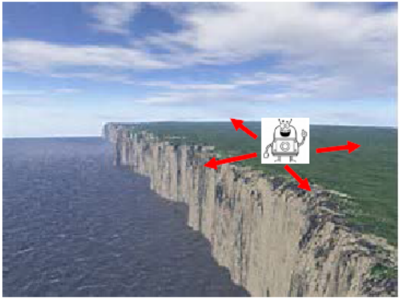
Ganguk Hwang

Department of Mathematical Sciences
KAIST

Introduction

Reinforcement Learning (RL) is concerned with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward.

(positive, negative)



Robot → agent

understanding environment

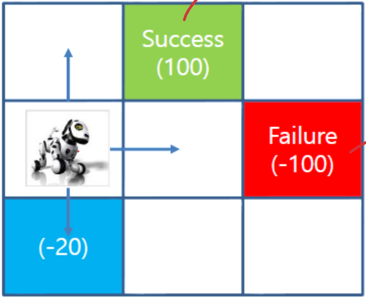
(some trials)

Figure: Agent and Environment

- Markov Decision Process

Example

(1,2) destination state



(2,3)

Figure: example

Robot moves to right ⇝ state (2,2)

- The environment or state $\{(1,1),(1,2),\cdots,(3,3)\}$ is (stochastically) changed according to the action that the agent takes
- The set of actions {up, down, left, right}
- The set of rewards $\{-100,-20,0,100\}$

- Agent - Environment Interaction

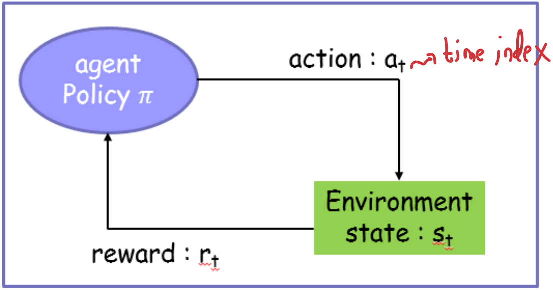


Figure: Markov Decision Process

- The agent tries to maximize the cumulative reward
- Reinforcement learning (RL) is learning from interaction
    - approach to sequential decision making

Overview of Markov Decision Process

- Markov Decision Process (MDP) consists of
  - a set of decision epochs or stages t time index of process
    - finite/infinite horizon, discrete/continuous
  - state space S
    - discrete: finite/countable, continuous
    - s_t: the state at stage t
  - action set A → all possible actions
  - (stationary) transition probability matrix $p(j|s, a)$ → of next stage j
    - $p : S \times A \rightarrow S$ → invariant
  - (stationary and bounded) reward function
    - $R : S \times A \rightarrow \mathbb{R}$
    - $r(s, a)$: the reward when the state is s and the action a is taken
  - policy π → set of actions
    - $\pi : S \rightarrow A$
    - $\pi(s)$: the action taken at state s
  - MDP is sometimes called a controlled Markov chain.

- at stage t
  - observe the state s_t
  - select an action based on the state s_t
  - obtain the resulting reward $r_t(s, a)$
- Influence diagram of MDP

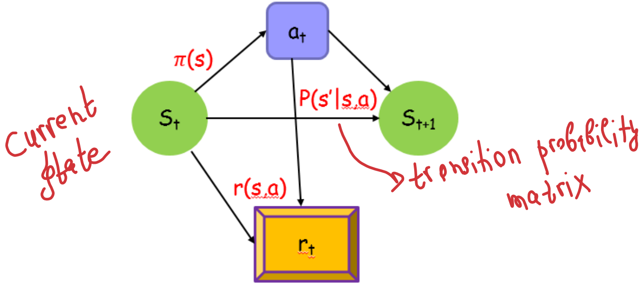


Figure: Influence diagram of MDP

Basic Assumptions invariants in time

1. stationary rewards $r(s, a)$ and transition probabilities $p(j|s, a)$
2. bounded rewards, i.e., $|r(s, a)| \leq M < \infty$ for all $a \in A$ and $s \in S$
3. discounting with λ $(0 \leq \lambda < 1)$ discounting vector
4. discrete state spaces S which is finite or countable

Remarks on policies

A decision rule d_t: $S \to A$ at stage t.

A policy π specifies the decision rules to be used at all decision epochs.

$$\pi = (d_1, d_2, \cdots).$$

There are different types of decision rules.

- Markovian and randomized (MR) $d_t : S \to \mathcal{P}(A)$
- Markovian and deterministic (MD) $d_t : S \to A$

We call a policy *stationary* if it uses the same decision rule for all stages, i.e.,

$$\pi = (d, d, d, \cdots).$$

Basic Setting

Let

- X_t: the state at stage t
- Y_t: the action taken at stage t

For a Markovian and deterministic decision rule d_t at stage t

$$Y_t = d_t(X_t) \text{ for } d_t \in D^{MD}$$

For a Markovian and randomized decision rule d_t,

$$P\{Y_t = a\} = q_{d_t(X_t)}(a) \text{ for } d_t \in D^{MR}$$

where $q_{d_t(s)}(\cdot)$ is a probability distribution for the action taken at state s.

The values of a policy in infinite horizon models

- The expected total reward of a policy π

$$v^{\pi}(s) = \lim_{N \to \infty} E_s^{\pi} \left[\sum_{t=1}^{N} r^{\pi}(X_t, Y_t) \right], \; s_1 = s \in S$$

- The expected total discounted reward of a policy π

$$v_{\lambda}^{\pi}(s) = \lim_{N \to \infty} E_s^{\pi} \left[\sum_{t=1}^{N} \lambda^{t-1} r^{\pi}(X_t, Y_t) \right], \; s_1 = s \in S$$

- The average reward or gain of a policy π

$$v^{\pi}(s) = \lim_{N \to \infty} \frac{1}{N} E_s^{\pi} \left[\sum_{t=1}^{N} r^{\pi}(X_t, Y_t) \right], \; s_1 = s \in S$$

The objective of the expected total discounted reward

- The objective of the expected total discounted reward

Find an optimal policy π^* that maximizes the expected total discounted reward

$$v_\lambda^\pi(s) = E^\pi\left[\sum_{t=1}^{\infty} \lambda^{t-1} r(X_t, Y_t)\right], \; s_1 = s \in S$$

(called the value function of a policy π). Here, λ is a discounted factor in $[0, 1)$.

Remarks:

- If the rewards are stochastic, the expectation of each reward is considered.
- Why a discount factor?
  - future rewards are not worth as much as current reward
  - there is a possibility that the process is terminated with probability $1-\lambda$
  - mathematically tractable

Recall that

$$v_\lambda^\pi(s) = E^\pi \left[\sum_{t=1}^{\infty} \lambda^{t-1} r(X_t, Y_t) \right], \; s_1 = s \in S$$

Which is given by

$$v_\lambda^\pi(s) \quad = \quad \sum_{t=1}^{\infty} \sum_{j \in S} \sum_{a \in A} \lambda^{t-1} r(j, a) P^\pi \{X_t = j, Y_t = a | X_1 = s\},$$

Policy Evaluation

For $d \in D^{MD}$, define

$$r_d(s) = r(s, d(s)), p_d(j|s) = p(j|s, d(s)).$$

For $d \in D^{MR}$, define

$$r_d(s) = \sum_{a \in A} q_{d(s)}(a) r(s, a), p_d(j|s) = \sum_{a \in A} q_{d(s)}(a) p(j|s, a).$$

r_d and P_d are the corresponding vector and matrix to $r_d(s)$ and $p_d(j|s)$, respectively.

Let $\pi = (d_1, d_2, d_3, \cdots) \in \Pi^{MR}$, $P_\pi^0 = I$, and $P_\pi^{(t)}$ be a matrix whose (s, j)-component is given by

$$P_\pi^{(t)}(j|s) = P^\pi\{X_{t+1} = j | X_1 = s\}.$$

Then $P_\pi^{(t)} = P_{d_1} P_{d_2} \cdots P_{d_t}$.

For instance,

$$\begin{aligned}
(P_\pi^{(1)})_{sj} &= P^\pi\{X_2 = j | X_1 = s\} \\
&= \sum_{a \in A} P^\pi\{X_2 = j, d_1(s) = a | X_1 = s\} \\
&= \sum_{a \in A} P^\pi\{X_2 = j | d_1(s) = a, X_1 = s\} P^\pi\{d_1(s) = a | X_1 = s\} \\
&= \sum_{a \in A} q_{d_1(s)}(a) p(j|s, a) \\
&= (P_{d_1})_{sj} \text{ (the } (s, j)\text{-th element of } P_{d_1})
\end{aligned}$$

$$
\begin{aligned}
(P_\pi^{(2)})_{sj} &= P^\pi\{X_3 = j | X_1 = s\} \\
&= \sum_{a\in A} P^\pi\{X_3 = j, d_1(s) = a | X_1 = s\} \\
&= \sum_{a\in A} P^\pi\{X_3 = j | d_1(s) = a, X_1 = s\} P^\pi\{d_1(s) = a | X_1 = s\} \\
&= \sum_{a\in A} q_{d_1(s)}(a) \sum_{i\in S}\sum_{b\in A} P^\pi\{X_3 = j, X_2 = i, d_2(i) = b | d_1(s) = a, X_1 = s \\
&= \sum_{a\in A} q_{d_1(s)}(a) \sum_{i\in S}\sum_{b\in A} P^\pi\{X_3 = j | X_2 = i, d_2(i) = b, X_1 = s, d_1(s) = a \\
&\qquad\qquad\qquad\qquad P^\pi\{X_2 = i, d_2(i) = b | d_1(s) = a, X_1 = s\} \\
&= \sum_{a\in A} q_{d_1(s)}(a) \sum_{i\in S}\sum_{b\in A} P^\pi\{X_3 = j | X_2 = i, d_2(i) = b\} \\
&\quad P^\pi\{d_2(i) = b | X_2 = i, d_1(s) = a, X_1 = s\} P^\pi\{X_2 = i | d_1(s) = a, X_1 = \\
&= \sum_{i\in S}\sum_{a\in A} q_{d_1(s)}(a) p(i|s,a) \sum_{b\in A} q_{d_2(i)}(b) p(j|i,b) \\
&= \sum_{i\in S} (P_{d_1})_{si}(P_{d_2})_{ij} = (P_{d_1}P_{d_2})_{sj}
\end{aligned}
$$

Note that the value function v_λ^π of a policy π is given as follows:

$$\begin{aligned}
v_\lambda^\pi(s) &= \sum_{t=1}^{\infty}\sum_{j\in S}\sum_{a\in A}\lambda^{t-1}r(j,a)P^\pi\{X_t=j,Y_t=a|X_1=s\}\\
&= \sum_{t=1}^{\infty}\sum_{j\in S}\sum_{a\in A}\lambda^{t-1}r(j,a)P^\pi\{Y_t=a|X_t=j,X_1=s\}\\
&\qquad\qquad P^\pi\{X_t=j|X_1=s\}\\
&= \sum_{t=1}^{\infty}\sum_{j\in S}\sum_{a\in A}\lambda^{t-1}r(j,a)q_{d_t(j)}(a)P^\pi\{X_t=j|X_1=s\}\\
&= \sum_{t=1}^{\infty}\sum_{j\in S}\lambda^{t-1}\sum_{a\in A}r(j,a)q_{d_t(j)}(a)P^\pi\{X_t=j|X_1=s\}\\
&= \sum_{t=1}^{\infty}\sum_{j\in S}\lambda^{t-1}r_{d_t}(j)P^\pi\{X_t=j|X_1=s\}\\
&= \sum_{t=1}^{\infty}\lambda^{t-1}(P_\pi^{(t-1)}r_{d_t})_s.
\end{aligned}$$

It then follows that

$$
\begin{aligned}
v_\lambda^\pi &= \sum_{t=1}^{\infty} \lambda^{t-1} P_\pi^{(t-1)} r_{d_t} \\
&= r_{d_1} + \lambda P_{d_1} r_{d_2} + \lambda^2 P_{d_1} P_{d_2} r_{d_3} + \cdots \\
&= r_{d_1} + \lambda P_{d_1} (r_{d_2} + \lambda P_{d_2} r_{d_3} + \cdots) \\
&= r_{d_1} + \lambda P_{d_1} v_\lambda^{\pi'}
\end{aligned}
$$

where $\pi' = (d_2, d_3, \cdots)$.

The value function for a stationary policy

When $\pi = d^\infty = (d, d, d, \cdots)$, i.e., π is stationary,

$$
v_\lambda^{d^\infty} = r_d + \lambda P_d v_\lambda^{d^\infty}.
$$

Bellman Equations (Optimality Equations)

Let V denote the set of bounded real valued functions on S and we use the supnorm on V. The corresponding matrix norm is

$$\|M\| = \sup_{s \in S} \sum_{j \in S} |M(j|s)|$$

Note that, for all $v \in V$ and $d \in D^{MR}$,

$$r_d + \lambda P_d v \in V.$$

For $v \in V$, define a linear transformation L_d by

$$L_d v = r_d + \lambda P_d v.$$

Then we know that $L_d : V \to V$ and $v_\lambda^{d^\infty}$ is a fixed point of L_d in V.

Theorem 1

For any stationary policy $\pi = d^\infty$ with $d \in D^{MR}$, $v_\lambda^{d^\infty}$ is the unique solution in V of

$$v = L_d v.$$

Moreover, $v_\lambda^{d^\infty}$ is written as

$$v_\lambda^{d^\infty} = (I - \lambda P_d)^{-1} r_d$$

where $(I - M)^{-1} = \sum_{n=0}^{\infty} M^n$.

Proof: We use the following corollary.

Corollary 2

Let Q be a bounded linear transformation on a Banach Space V, and suppose that the spectral radius satisfies $\rho(Q) < 1$. Then $(I - Q)^{-1}$ exists and satisfies

$$(I - Q)^{-1} = \sum_{n=0}^{\infty} Q^n$$

Since $\|P_d\| = 1$ and $\lambda = \|\lambda P_d\| \geq \rho(\lambda P_d)$, $(I - \lambda P_d)^{-1}$ exists. From $v = L_d v = r_d + \lambda P_d v$, we obtain

$$v = (I - \lambda P_d)^{-1} r_d = \sum_{t=1}^{\infty} \lambda^{t-1} P_d^{t-1} r_d = v_\lambda^{d^\infty}.$$

We now consider the following system of equations, called the optimality equations or Bellman equations.

$$v(s) = \sup_{a \in A} \left\{ r(s,a) + \lambda \sum_{j \in S} p(j|s,a) v(j) \right\}$$

For $v \in V$, define an operator $\mathcal{L}$ on V by

$$\mathcal{L}v = \sup_{d \in D^{MR}} \{ r_d + \lambda P_d v \}.$$

When the supremum is attained for all $v \in V$, we define L by

$$Lv = \max_{d \in D^{MR}} \{ r_d + \lambda P_d v \}.$$

The Bellman equations are given by

$$\mathcal{L}v = v \text{ or } Lv = v.$$

Proposition 3

For all $v \in V$*,*

$$\sup_{d \in D^{MR}} \{r_d + \lambda P_d v\} = \sup_{d \in D^{MD}} \{r_d + \lambda P_d v\}$$

Proof: Since $D^{MD} \subset D^{MR}$, $\geq$ is trivial. To prove the $\leq$ part, for $v \in V$, $d \in D^{MR}$ and observe that, for all $s \in S$

$$\begin{aligned} &\sup_{a \in A} \left\{ r(s,a) + \lambda \sum_{j \in S} p(j|s,a) v(j) \right\} \\ &\geq \sum_{a \in A} q_{d(s)}(a) \left\{ r(s,a) + \lambda \sum_{j \in S} p(j|s,a) v(j) \right\}. \end{aligned}$$

Let

$$v_\lambda^* = \sup_{\pi \in \Pi^{MR}} v_\lambda^\pi.$$

Theorem 4

Suppose that there exists a $v \in V$ *for which*

- $v \geq \mathcal{L}v$, *then* $v \geq v_\lambda^*$.
- $v \leq \mathcal{L}v$, *then* $v \leq v_\lambda^*$.
- $v = \mathcal{L}v$, *then* $v = v_\lambda^*$.

Proof:

1. Choose $\pi = (d_1, d_2, \cdots) \in \Pi^{MR}$. Then,

$$v \geq \sup_{d \in D^{MR}} \{r_d + \lambda P_d v\} = \sup_{d \in D^{MD}} \{r_d + \lambda P_d v\},$$

so that

$$\begin{aligned} v &\geq r_{d_1} + \lambda P_{d_1} v \geq r_{d_1} + \lambda P_{d_1} r_{d_2} + \lambda^2 P_{d_1} P_{d_2} v \\ &\geq \cdots \\ &\geq r_{d_1} + \lambda P_{d_1} r_{d_2} + \cdots + \lambda^{n-1} P_{d_1} \cdots P_{d_{n-1}} r_{d_n} + \lambda^n P_\pi^{(n)} v. \end{aligned}$$

Hence, from $v_\lambda^\pi = r_{d_1} + \lambda P_{d_1} r_{d_2} + \lambda^2 P_{d_1} P_{d_2} r_{d_3} + \cdots$ we obtain

$$v - v_\lambda^\pi \geq \lambda^n P_\pi^{(n)} v - \sum_{k=n}^{\infty} \lambda^k P_\pi^{(k)} r_{d_{k+1}}.$$

Choose $\epsilon > 0$. Since $\|\lambda^n P_\pi^{(n)} v\| \leq \lambda^n \|v\|$ and $0 \leq \lambda < 1$, for sufficiently large n

$$-\frac{\epsilon}{2} e \leq \lambda^n P_\pi^{(n)} v \leq \frac{\epsilon}{2} e$$

where e is the column vector whose elements are all equal to 1. Moreover, for sufficiently large n

$$\sum_{k=n}^{\infty} \lambda^k P_\pi^{(k)} r_{d_{k+1}} \leq \sum_{k=n}^{\infty} \lambda^k M e = \frac{\lambda^n M e}{1-\lambda} < \frac{\epsilon}{2} e.$$

Therefore, for all $s \in S$

$$v(s) \geq v_\lambda^\pi(s) - \epsilon.$$

2. Since $v \leq \mathcal{L}v = \sup_{d \in D^{MD}} \{r_d + \lambda P_d v\}$, for each $s \in S$ there exists $a_s \in A$ such that

$$v(s) \leq r(s, a_s) + \lambda \sum_{j \in S} p(j|s, a_s) v(j) + \epsilon.$$

Choose $d(s) = a_s$. Then,

$$v \leq r_d + \lambda P_d v + \epsilon e, \text{ i.e.,}$$

$$(I - \lambda P_d) v \leq r_d + \epsilon e.$$

It then follows that

$$v \leq (I - \lambda P_d)^{-1} (r_d + \epsilon e) = v_\lambda^{d^\infty} + (1 - \lambda)^{-1} \epsilon e.$$

Banach Fixed Point Theorem

To go further, first observe that V is a Banach space.

We say that an operator $T : V \to V$ is a contraction mapping if there exists $\lambda (0 \leq \lambda < 1)$ such that $\|Tv - Tu\| \leq \lambda \|v - u\|$ for all $u, v \in V$.

Theorem 5 (Banach Fixed Point Theorem)

Suppose that V is a Banach space and $T : V \to V$ is a contraction mapping. Then

1. *there exists a unique v^* in V such that $Tv^* = v^*$; and*
2. *for arbitrary $v^{(0)}$ in V, the sequence $\{v^{(n)}\}$ defined by*

$$v^{(n+1)} = Tv^{(n)} = T^{n+1} v^{(0)},$$

converges to v^*.

Proof: For any $m \geq 1$

$$\begin{aligned}
\|v^{(n+m)} - v^{(n)}\| &\leq \sum_{k=0}^{m-1} \|v^{(n+k+1)} - v^{(n+k)}\| \\
&= \sum_{k=0}^{m-1} \|T^{n+k} v^{(1)} - T^{n+k} v^{(0)}\| \\
&\leq \sum_{k=0}^{m-1} \lambda^{n+k} \|v^{(1)} - v^{(0)}\| \\
&= \frac{\lambda^n (1 - \lambda^m)}{1 - \lambda} \|v^{(1)} - v^{(0)}\|.
\end{aligned}$$

For sufficiently large n, the RHS can be made arbitrarily small, which means that $\{v^{(n)}\}$ is Cauchy. Since V is complete, the sequence $\{v^{(n)}\}$ has a limit $v^* \in V$.

By the properties of norms and contraction mapping,

$$\begin{aligned} 0 &\leq \|Tv^* - v^*\| \leq \|Tv^* - v^{(n)}\| + \|v^{(n)} - v^*\| \\ &\leq \|Tv^* - Tv^{(n-1)}\| + \|v^{(n)} - v^*\| \leq \lambda\|v^* - v^{(n-1)}\| + \|v^{(n)} - v^*\|. \end{aligned}$$

By letting n go to ∞, the RHS goes to 0, which implies that v^* is a fixed point of T.

Now, let v^* and u^* be two fixed points of T. Then, we have

$$\|v^* - u^*\| = \|Tv^* - Tu^*\| \leq \lambda\|v^* - u^*\|$$

and it implies that $v^* = u^*$.

The existence of an optimal policy

W now show the existence and uniqueness of a solution v^*_λ of the Bellman equations. To this end, we first show that the operator

$$\mathcal{L}(v) := \sup_{a \in A} \left\{ r(\cdot) + \lambda \sum p(\cdot|\cdot) v(\cdot) \right\}$$

is a contraction mapping, i.e.,

$$\|\mathcal{L}(v) - \mathcal{L}(u)\| \leq \lambda \|v - u\|.$$

Proof of Contraction:
Consider u and v. Fix $s \in S$ and assume that $\mathcal{L}(v)(s) \geq \mathcal{L}(u)(s)$.
There exists $a_s \in A$ such that

$$r(s, a_s) + \lambda \sum_{j \in S} p(j|s, a_s)v(j) + \epsilon > \sup_{a \in A} \left\{ r(s, a) + \lambda \sum_{j \in S} p(j|s, a)v(j) \right\}$$

Then,

$$\begin{aligned} 0 &\leq \mathcal{L}(v)(s) - \mathcal{L}(u)(s) \\ &\leq \lambda \sum_{j \in S} p(j|s, a_s)v(j) + \epsilon - \lambda \sum_{j \in S} p(j|s, a_s)u(j) \\ &\leq \lambda \sum_{j \in S} p(j|s, a_s)\|v - u\| + \epsilon \\ &\leq \lambda\|v - u\| + \epsilon. \end{aligned}$$

Repeating the argument for $\mathcal{L}(v)(s) \leq \mathcal{L}(u)(s)$ shows that

$$\|\mathcal{L}(v) - \mathcal{L}(u)\| \leq \lambda\|v - u\|.$$

Since $\mathcal{L}$ is a contraction mapping on V, by applying Banach Fixed Point Theorem there exists the unique solution $v^{(f)}$ such that $\mathcal{L}(v^{(f)}) = v^{(f)}$. Moreover, by Theorem 4 we have

$$v^{(f)} = v_\lambda^*.$$

We also have the following theorem.

Theorem 6

A policy $\pi^* \in \Pi^{MR}$ is optimal iff $v_\lambda^{\pi^*}$ is a solution of the optimal equations.

Existence of Optimal Policies

Given $v \in V$, call a decision rule $d_v \in D^{MD}$, *v-improving* if

$$d_v \in \operatorname{argmax}_{d \in D^{MD}} \{r_d + \lambda P_d v\}.$$

Equivalently,

$$r_{d_v} + \lambda P_{d_v} v = \max_{d \in D^{MD}} \{r_d + \lambda P_d v\}.$$

A decision rule $d^* \in D^{MD}$ is *conserving* (equivalently, v_λ^*-improving) if

$$L_{d^*} v_\lambda^* = L v_\lambda^* = v_\lambda^*.$$

Equivalently,

$$d^* \in \operatorname{argmax}_{d \in D^{MD}} \{r_d + \lambda P_d v_\lambda^*\}.$$

Suppose that d^* exists and is conserving. Then, the deterministic stationary policy $\pi = (d^*)^\infty = (d^*, d^*, \cdots)$ is the optimal policy.

Theorem 7

For all $\epsilon > 0$, there exists an ϵ-optimal deterministic stationary policy.

proof: For $\epsilon > 0$, choose $d_\epsilon \in D^{MD}$ such that

$$\begin{aligned} r_{d_\epsilon} + \lambda P_{d_\epsilon} v_\lambda^* &\geq \sup_{d \in D^{MD}} \{r_d + \lambda P_d v_\lambda^*\} - (1-\lambda)\epsilon e \\ &= v_\lambda^* - (1-\lambda)\epsilon e. \end{aligned}$$

So we have

$$r_{d_\epsilon} \geq (I - \lambda P_{d_\epsilon}) v_\lambda^* - (1-\lambda)\epsilon e.$$

Using $v_\lambda^{(d_\epsilon)^\infty} = (I - \lambda P_{d_\epsilon})^{-1} r_{d_\epsilon}$, we have

$$v_\lambda^{(d_\epsilon)^\infty} \geq v_\lambda^* - \epsilon e.$$

Hence, $(d_\epsilon)^\infty$ is ϵ-optimal.

Finding an optimal policy

There are three popular ways of finding an optimal policy in MDP.

- Value Iteration
- Policy Iteration
- Linear Programming

Their variants are

- Splitting Method
- Modified Policy Iteration
- etc.

Value Iteration

1. Start with an initial value function $v_0, \epsilon > 0$, and set $n = 0$.
2. For each $s \in S$, compute $v^{(n+1)}$ by

$$v^{(n+1)}(s) = \max_{a \in A} \left\{ r(s,a) + \lambda \sum_{j \in S} p(j|s,a) v^{(n)}(j) \right\}$$

3. If $\|v^{(n+1)} - v^{(n)}\| < \epsilon \frac{1-\lambda}{2\lambda}$, go to step 4. Otherwise, increment n by 1 and return to step 2.
4. For each $s \in S$, choose

$$d_\epsilon(s) = \text{argmax}_{a \in A} \left\{ r(s,a) + \lambda \sum_{j \in S} p(j|s,a) v^{(n+1)}(j) \right\}$$

and stop.

Convergence in Value Iteration

The value iteration algorithm finds a stationary policy that is ϵ-optimal within a finite number of iteration, i.e., $\|v_\lambda^{(d_\epsilon)^\infty} - v_\lambda^*\| < \epsilon$. Why?

First, observe that

$$\|v_\lambda^{(d_\epsilon)^\infty} - v_\lambda^*\| \leq \|v_\lambda^{(d_\epsilon)^\infty} - v^{(n+1)}\| + \|v^{(n+1)} - v_\lambda^*\|$$

Since $v_\lambda^{(d_\epsilon)^\infty}$ is a fixed point of L_{d_ϵ} and $Lv^{(n+1)} = L_{d_\epsilon} v^{(n+1)}$,

$$\begin{aligned}
\|v_\lambda^{(d_\epsilon)^\infty} - v^{(n+1)}\| &= \|L_{d_\epsilon} v_\lambda^{(d_\epsilon)^\infty} - v^{(n+1)}\| \\
&\leq \|L_{d_\epsilon} v_\lambda^{(d_\epsilon)^\infty} - Lv^{(n+1)}\| + \|Lv^{(n+1)} - v^{(n+1)}\| \\
&\leq \|L_{d_\epsilon} v_\lambda^{(d_\epsilon)^\infty} - L_{d_\epsilon} v^{(n+1)}\| + \|Lv^{(n+1)} - Lv^{(n)}\| \\
&\leq \lambda \|v_\lambda^{(d_\epsilon)^\infty} - v^{(n+1)}\| + \lambda \|v^{(n+1)} - v^{(n)}\|.
\end{aligned}$$

So we have, if $\|v^{(n+1)} - v^{(n)}\| < \epsilon \frac{1-\lambda}{2\lambda}$

$$\|v_\lambda^{(d_\epsilon)^\infty} - v^{(n+1)}\| \leq \frac{\lambda}{1-\lambda} \|v^{(n+1)} - v^{(n)}\| < \frac{\epsilon}{2}.$$

Similarly,

$$\begin{aligned}\|v^{(n+1)} - v_\lambda^*\| &\leq \|v^{(n+1)} - Lv^{(n+1)}\| + \|Lv^{(n+1)} - Lv_\lambda^*\| \\ &\leq \lambda\|v^{(n+1)} - v^{(n)}\| + \lambda\|v^{(n+1)} - v_\lambda^*\|\end{aligned}$$

So,

$$\|v^{(n+1)} - v_\lambda^*\| \leq \frac{\lambda}{1-\lambda}\|v^{(n+1)} - v^{(n)}\| < \frac{\epsilon}{2}.$$

Therefore,

$$\|v_\lambda^{(d_\epsilon)^\infty} - v_\lambda^*\| \leq \|v_\lambda^{(d_\epsilon)^\infty} - v^{(n+1)}\| + \|v^{(n+1)} - v_\lambda^*\| < \epsilon.$$

Policy Iteration

1. Set $n = 0$ and select an arbitrary policy $d_0 \in D^{MD}$.
2. (policy evaluation)
   - Obtain $v^{(n)}$ by solving
   $$(I - \lambda P_{d_n})v^{(n)} = r_{d_n}.$$
3. (policy improvement)
   - Choose d_{n+1} to satisfy
   $$d_{n+1} \in \mathrm{argmax}_{d \in D^{MD}} \left\{ r_d + \lambda P_d v^{(n)} \right\}$$
   setting $d_{n+1} = d_n$ if possible.
4. If $d_{n+1} = d_n$, stop and set $d^* = d_n$. Otherwise, increase n by 1 and return to step 2.

Convergence in Policy Iteration

- For two successive value functions $v^{(n)}$ and $v^{(n+1)}$, we have

$$v^{(n+1)} \geq v^{(n)}$$

which can be shown as follows.

$$r_{d_{n+1}} + \lambda P_{d_{n+1}} v^{(n)} \geq r_{d_n} + \lambda P_{d_n} v^{(n)} = v^{(n)},$$

i.e.,

$$r_{d_{n+1}} \geq (I - \lambda P_{d_{n+1}}) v^{(n)}.$$

Hence

$$v^{(n+1)} = (I - \lambda P_{d_{n+1}})^{-1} r_{d_{n+1}} \geq v^{(n)}.$$

- The algorithm is terminated in a finite time when there are finite policies and states.
- The convergence is also proved for more general state and action spaces under some suitable assumption.

More on RL: Full Observability vs. Partial Observability

- Full Observability
  - An agent directly observes the environment state.
  - A Markov Decision Process is used in this case.
- Partial Observability
  - An agent indirectly observe the environment state.
    For instance, a poker playing agent only observes public cards.
  - In this case, the agent state is not identical to the environment state.
  - A Partially Observable Markov Decision Process is used in this case.
  - The agent constructs its own state representation.

Model based RL vs. Model free RL

- Model based RL
  - The environment is modeled by a mathematical model such as MDP.
  - Algorithms which use a model are called model-based methods.
  - Learing from experience: $P(j|s,a) \propto \#(s, a \to j)$
- Model free RL
  - A ground-truth model of the environment is usually not available to the agent.
  - It uses the experience to directly learn a value function (e.g., Q-learning).
  - The model-free methods are more popular and have been more extensively developed and tested than the model-based methods.

References

- M.L. Puterman, Markov Decision Processes Discrete Stochastic Dynamic Programming, Willy, 2005.
- R.S. Sutton and A.G. Barto, Reinforcement Learing, The MIT Press, 1998.