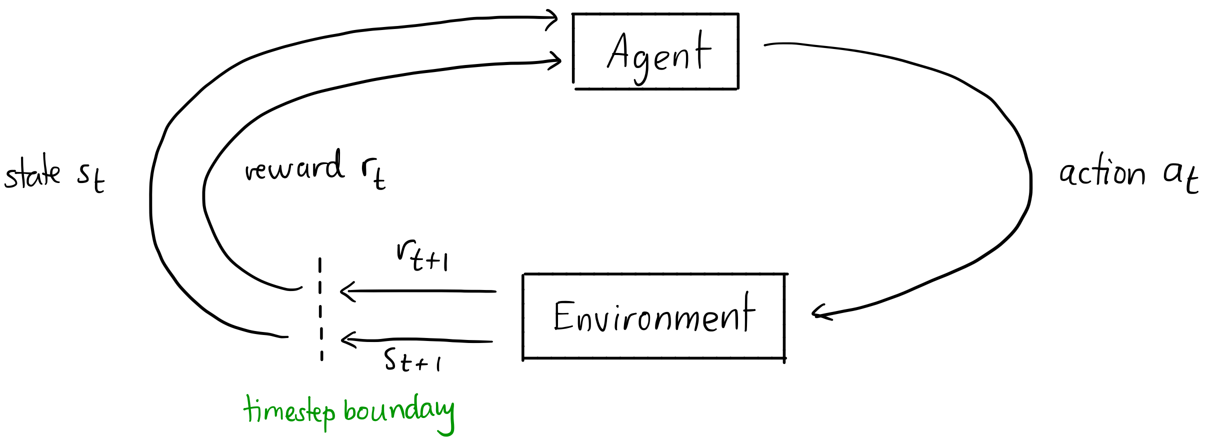
Introduction to Reinforcement Learning

This lecture introduces the basic reinforcement learning setup of a finite Markov decision process (MDP), the concepts of policies and value functions the "duality" between them, and the proof that "optimal" policies and value functions exist provided future rewards are discounted. There is substantial overlap between this lecture and Lecture 14 of my class MAST30026 (see http://therisingsea.org/post/mast30026/) which you can consult for some details omitted here. The standard references include:

- S. Russell, P. Norvig "Artificial intelligence: a modern approach" 3rd ed. §17.2.3
- R.S. Sutton, A.G. Barto "Reinforcement learning"

Def^n A finite MDP is a finite set S of states, a finite set $\mathcal{A}$ of actions, for each $s \in S$ a subset $\mathcal{A}(s) \subseteq \mathcal{A}$ of allowed actions in state s, a reward function $R : S \longrightarrow \mathbb{R}$ and for each pair $s \in S$, $a \in \mathcal{A}(s)$ a probability distribution $P(s' \mid s, a)$ over states $s' \in S$.

The interpretation is that an agent interacts with an environment, which has state space S, via actions which cause the environment to undergo transitions according to the distribution P, and in each (discrete) timestep the agent receives rewards. The goal of the agent is to act in such a way to obtain the maximal reward, in a sense to be specified more carefully in a moment.



For simplicity we assume there is a special <u>initial state</u> $s_{init} \in S$ and a subset of <u>terminal states</u> $S_{term} \subseteq S$. Assume $R(s_{init}) = 0$.

<u>Def^n</u> An <u>episode</u> e is a finite sequence

$$r_0, s_0, a_0, r_1, s_1, a_1, r_2, s_2, a_2, \dots, r_n, s_n \quad \text{i.e.} \quad e = \{(r_i, s_i, a_i)\}_{i=1}^{n}$$

satisfying the following conditions:

(i) $s_0 = s_{init}$, $s_n \in S_{term}$
(ii) $s_i \in S$ and $a_i \in \mathcal{A}(s_i)$ for all $0 \leq i \leq n$
(iii) $r_i = R(s_i)$ for $0 \leq i \leq n$.

The set of all episodes is denoted $\mathcal{E}$. (it may be infinite!)

The <u>discounted reward</u> (with fixed discount factor $0 < \gamma < 1$) of a sequence $\underline{s} = (s_i)_{i=0}^{n}$ of states is

$$R(\underline{s}, \gamma) := \sum_{t \geq 0} \gamma^t R(s_t).$$

The problem of optimal control is to determine how an agent should behave (that is, what actions it should choose) so as to maximise the expected value of the discounted reward over all episodes. More precisely, with $\Delta X \subseteq \mathbb{R}^X$ denoting the space of probability distributions on a finite set X, with the subspace topology:

<u>Def^n</u> A <u>policy</u> is a function $\pi : S \longrightarrow \Delta \mathcal{A}$ such that for all $s \in S$, the distribution $\pi(s)$ (which we write as $\pi(a|s) := \pi(s)(a)$) satisfies $\pi(a|s) = 0$ whenever $a \notin \mathcal{A}(s)$. (sometimes called a <u>stochastic</u> policy)

Giving S the discrete topology, let $\mathcal{P} \subseteq \mathrm{Cts}(S, \Delta\mathcal{A})$ be the set of all policies with the subspace topology (giving $\mathrm{Cts}(S,\Delta\mathcal{A}) = (\Delta\mathcal{A})^S$ the compact-open topology, or equivalently the product topology). This topology is determined by the metric d_∞ on $\mathcal{P}$, where (we can use any Lipschitz equiv. metric on $\mathbb{R}^{|S|} \supseteq \Delta\mathcal{A}$)

$$d_\infty(\pi, \pi') = \sup_{s\in S} \sup_{a\in\mathcal{A}(s)} |\pi(a|s) - \pi'(a|s)|.$$

For the reader's convenience I will refer to proofs of various facts below given in my MAST30026 class, but this is standard material which can be found in many places.

Def^n Given an episode $e = \{(r_i, s_i, a_i)\}_{i=1}^n$ and policy $\pi \in \mathcal{P}$ the probability of e occurring if the agent acts according to π is

$$P_\pi(e) := \prod_{i=0}^{n-1} \pi(a_i|s_i) \cdot \prod_{i=0}^{n} P(s_{i+1}|s_i, a_i).$$

The expected discounted reward of π is

$$\mathbb{E}(R_\pi) := \sum_{e=(\underline{r},\underline{s},\underline{a})} P_\pi(e) \cdot R(\underline{s}, \gamma)$$

Lemma There exists a policy π^* which is optimal, in the sense that for all $\rho \in \mathcal{P}$, $\mathbb{E}(R_{\pi^*}) \geqslant \mathbb{E}(R_\rho)$.

Proof $\mathbb{E}(R_{(-)}) : \mathcal{P} \longrightarrow \mathbb{R}$ is continuous, and we claim $\mathcal{P}$ is compact, so that the claim follows from the extreme value theorem (Corollary L9-4). Note that $(\Delta\mathcal{A})^S = \prod_{s\in S} \Delta\mathcal{A}$ is a finite product of compact spaces ($\Delta\mathcal{A} \subseteq \mathbb{R}^{|\mathcal{A}|}$ is closed and bounded) hence compact.

For each $s \in S$, $\mathcal{A}(s) \subseteq \mathcal{A}$ means $\Delta\mathcal{A}(s)$ is a closed subset of $\Delta\mathcal{A}$, hence

$$\mathcal{P} = \prod_{s \in S} \Delta\mathcal{A}(s) \subseteq \prod_{s \in S} \Delta\mathcal{A}$$

is a closed subset, hence compact. $\square$

Of course there may be more than one optimal policy. However, the above argument is not constructive, so it is not clear how to find such a policy. However there is a general trick: if you want to optimise a function f, rephrase the optimisation problem as a fixed point problem for a different function g (see Lecture 14 for examples of this). This leads us to value functions.

Def^n Set $r_{max} = \sup_{s \in S} |R(s)|$ and $H = \dfrac{r_{max}}{1-\gamma}$.

Def^n The space of value functions is $\mathcal{V} = \mathrm{Cts}(S, [-H,H]) = [-H,H]^S$ with the compact-open topology (i.e. the sup metric). This is a complete metric space, which is also compact.

Lemma Let $\pi \in \mathcal{P}$ be a policy. Then $\Phi_\pi : \mathcal{V} \longrightarrow \mathcal{V}$ defined by

$$\Phi_\pi(v)(s) = R(s) + \gamma \sum_{s' \in S} \sum_{a \in \mathcal{A}(s)} \pi(a|s) P(s'|s,a) v(s')$$

is a contraction mapping, with contraction factor γ.

Proof Set $\Phi = \Phi_\pi$. First we should check this is well-defined, i.e. if $|v(s)| \leq H$ for all $s \in S$ then $|\Phi(v)(s)| \leq H$ for all $s \in S$. But

$$|\Phi(v)(s)| \leq r_{max} + \gamma H = H$$

To prove Φ is a contraction observe that

$$\begin{aligned} |\Phi(v)(s) - \Phi(w)(s)| &= \gamma \Big| \sum_{s' \in S} \sum_{a \in A(s)} \pi(a|s) P(s'|s,a) \big(v(s') - w(s')\big) \Big| \\ &\leq \gamma \sum_{s' \in S} \sum_{a \in A(s)} \pi(a|s) P(s'|s,a) \,|v(s') - w(s')| \\ &\leq \gamma\, d_\infty(v, w). \quad \square \end{aligned}$$

Recall that by the Banach fixed point theorem any contraction mapping $\Phi : \mathcal{V} \longrightarrow \mathcal{V}$ on a complete metric space $\mathcal{V}$ has a unique fixed point $fix(\Phi)$ which may be computed from any initial $v_0 \in \mathcal{V}$ by iterating Φ. In the situation of the lemma, beginning with $v_0 \equiv 0$ we obtain $v_0, v_1 = \Phi_\pi(v_0), v_2 = \Phi_\pi^2(v_0), \ldots$

$$v_1(s) = \Phi_\pi(v_0)(s) = R(s)$$

$$v_2(s) = \Phi_\pi(v_1)(s) = R(s) + \gamma \sum_{s' \in S} \sum_{a \in A(s)} \pi(a|s) P(s'|s,a) R(s')$$

$$\begin{aligned} v_3(s) = \Phi_\pi(v_2)(s) = R(s) &+ \gamma \sum_{s' \in S} \sum_{a \in A(s)} \pi(a|s) P(s'|s,a) R(s') \\ &+ \gamma^2 \sum_{s', s'' \in S} \sum_{\substack{a \in A(s) \\ a' \in A(s')}} \pi(a|s) P(s'|s,a) \cdot \pi(a'|s') P(s''|s',a') R(s'') \end{aligned}$$

<u>Def</u>

Def^n Let $v_\pi \in \mathcal{V}$ denote the unique fixed point of Φ_π. We call v_π the <u>evaluation</u> of the policy π.

The above shows that $v_\pi(s)$ contains contributions from all paths in state space beginning at s.

If $\Phi_n \to \Phi$ is a sequence of contraction mappings with the same contraction factor γ converging uniformly to Φ, then $fix(\Phi_n) \to fix(\Phi)$ since (writing $u_n = fix(\Phi_n)$ and $u = fix(\Phi)$)

$$\begin{aligned} d_{\mathcal{V}}(u_n, u) &= d_{\mathcal{V}}(\Phi_n(u_n), \Phi(u)) \\ &\leq d_{\mathcal{V}}(\Phi_n(u_n), \Phi_n(u)) + d_{\mathcal{V}}(\Phi_n(u), \Phi(u)) \\ &\leq \gamma\, d_{\mathcal{V}}(u_n, u) + d_\infty(\Phi_n, \Phi) \end{aligned}$$

and hence $d_{\mathcal{V}}(u_n, u) \leq (1-\gamma)^{-1} d_\infty(\Phi_n, \Phi)$.

<u>Remark</u> Since $\mathcal{V}$ is compact, the compact-open topology on $Cts(\mathcal{V}, \mathcal{V})$ agrees with the topology associated to the sup-metric d_∞.

Let $Ctr_\gamma(\mathcal{V}, \mathcal{V}) \subseteq Cts(\mathcal{V}, \mathcal{V})$ denote the set of γ-contraction mappings with the subspace topology.

<u>Def^n</u> <u>Policy evaluation</u> is the continuous function

$$\begin{aligned} \mathcal{P} &\xrightarrow{\Phi_{(-)}} Ctr_\gamma(\mathcal{V}, \mathcal{V}) \xrightarrow{fix} \mathcal{V} \\ \pi &\longmapsto \Phi_\pi \longmapsto fix(\Phi_\pi) \end{aligned}$$

which sends a policy π to the unique solution in $\mathcal{V}$ of the equation

$$v(s) = R(s) + \gamma \sum_{s' \in S} \sum_{a \in A(s)} \pi(a|s) P(s'|s,a) v(s')$$

It is traditional to denote this value function by v_π.

<u>Remark</u> To see $\Phi_{(-)}$ is continuous, note that

$$
\begin{aligned}
d_\infty(\Phi_\pi, \Phi_\rho) &= \sup_{v \in \mathcal{V}} d_{\mathcal{V}}(\Phi_\pi(v), \Phi_\rho(v)) \\
&= \sup_{v \in \mathcal{V}} \sup_{s \in S} \left| \Phi_\pi(v)(s) - \Phi_\rho(v)(s) \right| \\
&\leq \sup_{v \in \mathcal{V}} \sup_{s \in S} \sum_{s' \in S} \sum_{a \in \mathcal{A}(s)} \left| \pi(a|s) - \rho(a|s) \right| P(s'|s,a) |v(s')| \\
&\leq \sup_{v \in \mathcal{V}} \sup_{s \in S} \sum_{s' \in S} \sum_{a \in \mathcal{A}(s)} \left| \pi(a|s) - \rho(a|s) \right| P(s'|s,a)\, H \\
&\leq \sup_{v \in \mathcal{V}} \sup_{s \in S} \sum_{a \in \mathcal{A}(s)} d_\infty(\pi, \rho) \cdot H \\
&\leq |\mathcal{A}| \cdot H \cdot d_\infty(\pi, \rho)
\end{aligned}
$$

To briefly summarise: associated to any finite MDP we have a compact space $\mathcal{P}$ of <u>policies</u>, a compact space $\mathcal{V}$ of <u>value functions</u>, a continuous function $\mathbb{E}(R_{(-)}) : \mathcal{P} \longrightarrow \mathbb{R}$ assigning to each policy the expected discounted reward, and a continuous <u>policy evaluation</u> $\mathcal{P} \longrightarrow \mathcal{V}$ sending π to v_π.

<u>Theorem</u> There is a unique solution $v^* \in \mathcal{V}$ of the <u>Bellman equation</u>

$$v(s) = R(s) + \gamma \sup_{a \in \mathcal{A}(s)} \sum_{s' \in S} P(s'|s,a)\, v(s'). \qquad (*)$$

This v^* is the evaluation of an optimal policy π^*, and we call v^* the <u>optimal value function</u>.

<u>Proof</u> The Bellman equation gives a contraction map $\Phi : \mathcal{V} \to \mathcal{V}$ which has a unique fixed point, i.e. a unique solution to $(*)$ exists. If we define π^* to be the deterministic policy

$$\pi^*(s) = \operatorname{argmax}_{a \in \mathcal{A}(s)} \sum_{s' \in S} P(s'|s,a)\, v^*(s')$$

then it is easy to check $v_{\pi^*} = v^*$, so it only remains to show π^* is optimal. Let v be a value function and suppose $v(s) \le v^*(s)$ for all $s \in S$. Then for any policy π

$$\begin{aligned}
\Phi_\pi(v)(s) &= R(s) + \gamma \sum_{a \in \mathcal{A}(s)} \pi(a|s) \sum_{s' \in S} P(s'|s,a)\, v(s') \\
&\le R(s) + \gamma \sum_{a \in \mathcal{A}(s)} \pi(a|s) \sup_{a \in \mathcal{A}(s)} \Big\{ \sum_{s' \in S} P(s'|s,a)\, v(s') \Big\} \\
&= R(s) + \gamma \sup_{a \in \mathcal{A}(s)} \sum_{s' \in S} P(s'|s,a)\, v(s') \\
&\le R(s) + \gamma \sup_{a \in \mathcal{A}(s)} \sum_{s' \in S} P(s'|s,a)\, v^*(s') = v^*(s)
\end{aligned}$$

Hence $v \le v^*$ implies $\Phi_\pi(v) \le v^*$. But taking the limit we obtain that $v_\pi \le v^*$ (since we may start with $v \equiv -M$), and in particular $\mathbb{E}(R_\pi) = v_\pi(s_{init}) \le v^*(s_{init}) = \mathbb{E}(R_{\pi^*})$ so that π^* is optimal. $\square$

<u>Remark</u> A policy is "implicit" in the sense that it dictates the immediate behaviour in a given state, whereas a value function is "explicit" in the sense that it contains global information about the long-run consequences of a behaviour. It seems reasonable to compare the policy π to an <u>algorithm</u> and V_π to the function that this algorithm computes, with the fixed point iteration being analogous to the process of computation itself.