

This material is partially based on the book draft “Reinforcement Learning: Foundations” by Shie Mannor, Yishay Mansour, and Aviv Tamar.

Similarly to before, we consider an MDP with finite state space \mathcal{S} , finite action space \mathcal{A} such that $\mathcal{A}(s) = \mathcal{A}$ for all $s \in \mathcal{S}$, and transition kernel $\{p(\cdot | s, a) : s \in \mathcal{S}, a \in \mathcal{A}\}$. However, for simplicity we assume a time-independent reward function $r : \mathcal{S} \times \mathcal{A} \rightarrow [-1, 1]$.

We now want to derive the Bellman optimality equations for the discounted horizon case. We can not use backward induction because the horizon is stochastic. For any fixed $0 < \gamma < 1$, the state-value function $V^\pi : \mathcal{S} \rightarrow \mathbb{R}$ for a policy π gives the γ -discounted return from any initial state s_0 ,

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

where $a_t \sim \pi_t(\cdot | s_t)$. Note that, since rewards are bounded in $[-1, 1]$,

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

We can also define the state-value function with respect to an initial state distribution μ ,

$$V^\pi(\mu) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]$$

where $s_0 \sim \mu$.

Our goal is to find the policy $\pi^* = (\pi_0^*, \pi_1^*, \dots)$ that maximizes $V^\pi(s_0)$ for each initial state s_0 with respect to all policies π . Let V^* the state-value function for the optimal policy π^* . Since $\mathbb{P}(T = t) = \gamma^{t-1}(1 - \gamma)$, we know that $\pi^* = (\pi_0, \pi_1, \dots)$ is Markov and deterministic.

Next, we prove an important property of the state-value function.

Lemma 1 *For any stationary and deterministic Markov policy π , V^π satisfies the following $|\mathcal{S}|$ linear equations*

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

PROOF.

$$\begin{aligned}
V^\pi(s) &= \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t)) \middle| s_0 = s \right] \\
&= r(s, \pi(s)) + \sum_{s' \in \mathcal{S}} p(s' | s, \pi(s)) \mathbb{E} \left[\sum_{t=1}^{\infty} \gamma^t r(s_t, \pi(s_t)) \middle| s_1 = s' \right] \\
&= r(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} p(s' | s, \pi(s)) \mathbb{E} \left[\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, \pi(s_t)) \middle| s_1 = s' \right] \\
&= r(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} p(s' | s, \pi(s)) \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t)) \middle| s_0 = s' \right] \quad (\text{because } \pi \text{ is stationary}) \\
&= r(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} p(s' | s, \pi(s)) V^\pi(s')
\end{aligned}$$

□

This property suggests a way of computing V^π for a fixed policy π , which could be useful if we want to find π maximizing V^π .

In view of that, it is convenient to rephrase the above property using vector notation. Let $\mathbf{v}^\pi, \mathbf{r}^\pi \in [-1, 1]^{|\mathcal{S}|}$ with components $v^\pi(s) = V^\pi(s)$ and $r^\pi(s) = r(s, \pi(s))$. Let also \mathbf{P} be a $|\mathcal{S}| \times |\mathcal{S}|$ matrix with components $P(s', s) = p(s' | s, \pi(s))$. Then

$$\mathbf{v}^\pi = \mathbf{r}^\pi + \gamma \mathbf{P}^\pi \mathbf{v}^\pi$$

Note that the above is equivalent to $(I - \gamma \mathbf{P}^\pi) \mathbf{v}^\pi = \mathbf{r}^\pi$. Since $I - \gamma \mathbf{P}^\pi$ is non-singular (because \mathbf{P}^π is a row-stochastic matrix and therefore the eigenvalues λ_i satisfy $|\lambda_i| \leq 1$), we find that

$$\mathbf{v}^\pi = (I - \gamma \mathbf{P}^\pi)^{-1} \mathbf{r}^\pi$$

Since inverting the $|\mathcal{S}| \times |\mathcal{S}|$ matrix $I - \gamma \mathbf{P}^\pi$ requires order of $|\mathcal{S}|^3$ operations, an alternative way to compute V^π is via (fixed-policy) value iteration: $\mathbf{v}_{n+1}^\pi = \mathbf{r}^\pi + \gamma \mathbf{P}^\pi \mathbf{v}_n^\pi$ where \mathbf{v}_0^π is an arbitrary initial vector. It can be proven that

$$\|\mathbf{v}^\pi - \mathbf{v}_n^\pi\|_\infty \leq \left(\frac{1}{1 - \gamma} - \|\mathbf{v}_0^\pi\|_\infty \right) \gamma^n$$

Implying that fixed-policy value iteration converges exponentially fast.

The following result provides an explicit characterization of the optimal state-value function V^* and shows that the optimal policy is stationary and can easily be computed if the state-value function is known.

Theorem 2 (Bellman's Optimality Equations) *The following statements hold:*

1. V^* is the unique solution of the following system of nonlinear equations

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

2. Any stationary policy π^* satisfying

$$\pi^*(s) \in \operatorname{argmax}_{a \in \mathcal{A}} \left(r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s' | s, a) V^*(s') \right) \quad s \in \mathcal{S}$$

is such that $V^{\pi^*} = V^*$.

Unlike V^π , the equation defining V^* is not linear. Yet, similarly to V^π , we can compute V^* using **value iteration** (VI):

$$V_{n+1}(s) = \max_{a \in \mathcal{A}} \left(r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s' | s, a) V_n(s') \right) \quad s \in \mathcal{S}$$

where V_0 is arbitrary. Similarly to fixed-policy value iteration, one can show that

$$\max_{s \in \mathcal{S}} |V^*(s) - V_n(s)| \leq \left(\frac{1}{1 - \gamma} - \max_{s \in \mathcal{S}} |V_0(s)| \right) \gamma^n$$

A different method, called **policy iteration** (PI), constructs a sequence of policies converging to the optimal policy:

For $n = 0, 1, \dots$

1. Policy evaluation: Compute V^{π_n} using $\mathbf{v}^\pi = (I - \gamma \mathbf{P}^\pi)^{-1} \mathbf{r}^\pi$
2. Policy improvement:

$$\pi_{n+1}(s) \in \operatorname{argmax}_{a \in \mathcal{A}} \left(r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s' | s, a) V^{\pi_n}(s') \right) \quad s \in \mathcal{S}$$

PI performs successive rounds of policy improvement, where each policy π_{n+1} improves on the previous one π_n . Since the number of stationary policies is bounded, so is the number of strict improvements, and PI must terminate with an optimal policy after a finite number policy updates.

In terms of running time, PI requires $\mathcal{O}(|\mathcal{A}| |\mathcal{S}|^2 + |\mathcal{S}|^3)$ operations per iteration, while VI only requires $\mathcal{O}(|\mathcal{A}| |\mathcal{S}|^2)$ operations per iteration. However, in many cases PI has a smaller number of iterations than VI. Indeed, one can show that $V_n^{\text{VI}} \leq V_n^{\text{PI}} \leq V^*$ for all $n \geq 0$, where V_n^{VI} and V_n^{PI} are the sequences of state-value functions produced, respectively, by VI and PI, and we assume $V_0^{\text{VI}} = V_0^{\text{PI}}$.

Linear programming duality. In order to obtain more insights on the Bellman's equations it is useful to introduce the notion of **discounted occupancy measure** q^π , which adapts to the discounted infinite horizon criterion the quantity q_t^π introduced earlier,

$$q^\pi(s, a) = \sum_{t=0}^{\infty} q_t^\pi(s, a) = \sum_{t=0}^{\infty} \mathbb{P}^\pi(s_t = s, \pi(s) = a, T \geq t) = \sum_{t=0}^{\infty} \gamma^t \mathbb{P}^\pi(s_t = s, \pi(s) = a)$$

where

$$\sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} q^\pi(s, a) = \frac{1}{1 - \gamma}$$

We can now express the return in terms of the discounted occupancy measure,

$$\begin{aligned}
V^\pi(\mu) &= \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t)) \right] \\
&= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sum_{t=0}^{\infty} \gamma^t r(s, a) \mathbb{P}^\pi(s_t = s, \pi(s_t) = a) \\
&= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s, a) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}^\pi(s_t = s, \pi(s_t) = a) \\
&= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s, a) q^\pi(s, a)
\end{aligned}$$

This shows that $V^\pi(\mu)$ is linear in $q^\pi(s, a)$ for any π . We can exploit this fact and write $V^*(\mu)$ as the solution of the following linear program (LP)

$$\begin{aligned}
&\max_{q: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s, a) q(s, a) \\
&\text{subject to: } q(s, a) \geq 0 && (s, a) \in \mathcal{S} \times \mathcal{A} \\
&\sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} q(s, a) = \frac{1}{1 - \gamma} \\
&\sum_{a \in \mathcal{A}} q(s', a) = \mu(s') + \gamma \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} p(s' | s, a) q(s, a) \quad s' \in \mathcal{S}
\end{aligned}$$

where the constraints define the set of all feasible discounted occupancy measures. The optimal stationary policy π^* can be directly obtained from the solution q^* of the LP as

$$\pi^*(a | s) = \frac{q^*(s, a)}{\sum_{a' \in \mathcal{A}} q^*(s, a')}$$

The dual linear program is then written as

$$\begin{aligned}
&\min_{V: \mathcal{S} \rightarrow \mathbb{R}} \sum_{s \in \mathcal{S}} \mu(s) V(s) \\
&\text{subject to:} \\
&V(s) \geq r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s' | s, a) V(s') \quad (s, a) \in \mathcal{S} \times \mathcal{A}
\end{aligned}$$

where the constraints enforce the first Bellman's Optimality equation. By strong duality, the optimal solution of the dual is also $V^*(\mu)$. Moreover, the state-value function and the discounted occupancy measure are dual decision variables.