Reinforcement Learning

Discounted MDP

Lecturer: *Nicolò Cesa-Bianchi* version January 31, 2024

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 $\{p(\cdot \mid s, a) : s \in \mathcal{S}, a \in \mathcal{A}\}$. However, for simplicity we assume a time-independent reward function $r : \mathcal{S} \times \mathcal{A} \to [-1, 1]$.

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 $V^{\boldsymbol{\pi}} : \mathcal{S} \to \mathbb{R}$ for a policy $\boldsymbol{\pi}$ gives the γ-discounted return from any initial state s,

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

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

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

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

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

where $s_0 \sim \mu$.

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

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

PROOF.

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

□

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 $\boldsymbol{v}^\pi, \boldsymbol{r}^\pi \in [-1,1]^{|\mathcal{S}|}$ with components $v^\pi(s) = V^\pi(s)$ and $r^\pi(s) = r\big(s,\pi(s)\big)$. Let also $\boldsymbol{P}^\pi$ be a $|\mathcal{S}|\times|\mathcal{S}|$ matrix with components $P^\pi(s,s') = p\big(s' \mid s,\pi(s)\big)$. Then

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

Note that the above is equivalent to $(I-\gamma\boldsymbol{P}^\pi)\boldsymbol{v}^\pi = \boldsymbol{r}^\pi$. Note also that $\boldsymbol{P}^\pi$ is a row-stochastic matrix, and therefore its eigenvalues λ_i satisfy $|\lambda_i| \le 1$. Since the eigenvalues of $I-\gamma\boldsymbol{P}^\pi$ are of the form $1-\gamma\lambda_i$ with $0<\gamma<1$, they are all positive and so $I-\gamma\boldsymbol{P}^\pi$ is non-singular. We thus find that

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

Since inverting the $|\mathcal{S}|\times|\mathcal{S}|$ matrix $I-\gamma\boldsymbol{P}^\pi$ requires order of $|\mathcal{S}|^3$ operations, an alternative way to compute V^π is via (fixed-policy) value iteration: $\boldsymbol{v}^\pi_{n+1} = \boldsymbol{r}^\pi + \gamma\boldsymbol{P}^\pi\boldsymbol{v}^\pi_n$ where $\boldsymbol{v}^\pi_0$ is an arbitrary initial vector.

We now show that fixed-policy value iteration converges exponentially fast. First, note that

$$\boldsymbol{v}^\pi_1 = \boldsymbol{r}^\pi + \gamma\boldsymbol{P}^\pi\boldsymbol{v}^\pi_0 = \boldsymbol{r}^\pi + \gamma\mathbb{E}_1\big[\boldsymbol{v}^\pi_0\big]$$

where $\mathbb{E}_1\big[\boldsymbol{v}^\pi_0\big]$ is a vector whose s-th component is

$$\mathbb{E}_1\big[\boldsymbol{v}^\pi_0\big](s) = \sum_{s'\in\mathcal{S}} \boldsymbol{v}^\pi_s(s')p\big(s' \mid s,\pi(s)\big)$$

Iterating n times, we get

$$\boldsymbol{v}^\pi_n = \sum_{t=0}^{n-1}\gamma^t\mathbb{E}_t\big[\boldsymbol{r}^\pi\big] + \gamma^n\mathbb{E}_n\big[\boldsymbol{v}^\pi_0\big]$$

where the s-th components of $\mathbb{E}_t\big[\boldsymbol{r}^\pi\big]$ and $\mathbb{E}_n\big[\boldsymbol{v}_0^\pi\big]$ are defined by

$$\mathbb{E}_t\big[\boldsymbol{r}^\pi\big](s) = \sum_{s_1\in\mathcal{S}}\cdots\sum_{s_t\in\mathcal{S}} \boldsymbol{r}^\pi\big(s_t,\pi(s_t)\big)\, p\big(s_t \mid s_{t-1},\pi(s_{t-1})\big)\cdots p\big(s_1 \mid s,\pi(s)\big)$$
$$\mathbb{E}_n\big[\boldsymbol{v}_0^\pi\big](s) = \sum_{s_1\in\mathcal{S}}\cdots\sum_{s_n\in\mathcal{S}} \boldsymbol{v}_0^\pi(s_n)\, p\big(s_n \mid s_{n-1},\pi(s_{n-1})\big)\cdots p\big(s_1 \mid s,\pi(s)\big)$$

Now,

$$\lim_{n\to\infty}\boldsymbol{v}_n^\pi = \sum_{t=0}^{\infty}\gamma^t\mathbb{E}_t\big[\boldsymbol{r}^\pi\big] = \boldsymbol{v}_n^\pi$$

Hence

$$\boldsymbol{v}^\pi - \boldsymbol{v}_n^\pi = \sum_{t=n}^{\infty}\gamma^t\,\mathbb{E}_t\big[\boldsymbol{r}^\pi\big] - \gamma^n\mathbb{E}_n\big[\boldsymbol{v}_0^\pi\big] = \gamma^n\left(\sum_{t=0}^{\infty}\gamma^t\mathbb{E}_{n+t}\big[\boldsymbol{r}^\pi\big] - \mathbb{E}_n\big[\boldsymbol{v}_0^\pi\big]\right)$$

Since the components of $\mathbb{E}_{n+t}\big[\boldsymbol{r}^\pi\big]$ are all in $[-1,1]$, we have that

$$\left|\sum_{t=0}^{\infty}\gamma^t\mathbb{E}_{n+t}\big[\boldsymbol{r}^\pi\big]\right| \le \frac{\mathbf{1}}{1-\gamma}$$

where $\mathbf{1} = (1,\dots,1)$. Moreover, since the expectation is always smaller that the maximum, $\mathbb{E}_n\big[\boldsymbol{v}_0^\pi\big] \le \mathbf{1}\,\|\boldsymbol{v}_0^\pi\|_\infty$. This implies

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

showing 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 Optimality Equations) *The following statements hold:*

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

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

2. *Any stationary policy π^* satisfying*

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

   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):

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

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

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

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 $\boldsymbol{v}^\pi = (I - \gamma \boldsymbol{P}^\pi)^{-1} \boldsymbol{r}^\pi$
2. Policy improvement:

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

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 $\mathcal{O}\big(|\mathcal{A}|\,|\mathcal{S}|^2 + |\mathcal{S}|^3\big)$ operations per iteration, while VI only requires $\mathcal{O}\big(|\mathcal{A}|\,|\mathcal{S}|^2\big)$ operations per iteration. However, in many cases PI has a smaller number of iterations than VI. Indeed, one can show that $V_n^{\mathrm{VI}} \leq V_n^{\mathrm{PI}} \leq V^*$ 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 $V_0^{\mathrm{VI}} = V_0^{\mathrm{PI}}$.

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

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

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,

$$
\begin{aligned}
V^{\boldsymbol{\pi}}(\mu) &= \sum_{s' \in \mathcal{S}} \mu(s') V(s') \\
&= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \, r\big(s_t, \pi(s_t)\big)\right] \qquad \text{(where } s_0 \sim \mu \text{ and } s_t \sim p(\cdot | s_{t-1}, \pi(s_{t-1}))\text{)} \\
&= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sum_{t=0}^{\infty} \gamma^t \, r(s,a) \mathbb{P}^{\boldsymbol{\pi}}\big(s_t = s, \pi(s_t) = a\big) \\
&= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s,a) \sum_{t=0}^{\infty} \gamma^t \, \mathbb{P}^{\boldsymbol{\pi}}\big(s_t = s, \pi(s_t) = a\big) \\
&= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s,a) q^{\boldsymbol{\pi}}(s,a)
\end{aligned}
$$

This shows that $V^{\boldsymbol{\pi}}(\mu)$ is linear in $q^{\boldsymbol{\pi}}(s,a)$ for any π. In particular,

$$
V^*(\mu) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s,a) q^*(s,a)
$$

where q^* is the discounted occupancy measure of π^*. This also shows that we can find π^* by solving the following linear program (LP)

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

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

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

and it is easy to verify that the discounted occupancy measure of π^* is indeed q^*. The dual program

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

reveals that the Bellman Optimality equations arise as constraints of the dual program, and that the state-value function and the discounted occupancy measure are dual decision variables.