**Markov Decision Processes** **April 22, 2024**

# Module 2.C: Infinite-Horizon Discounted Markov Decision Processes

*Lecturer: Avik Kar*

**Note**: *LaTeX template courtesy of UC Berkeley EECS dept.*

**Disclaimer**: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

## 1 The problem

**Definition 1.1** (Infinite-horizon Discounted Markov decision process)**.** We refer to the tuple $(X, U, \gamma, p, \bar{r}) =: M$ as a Markov decision process where

1. $\mathcal{X}$ is the state space of the system.
2. $\mathcal{U}$ is the action space. At system state $x \in \mathcal{X}$, the agent can take action from $\mathcal{U}_x \subseteq \mathcal{U}$.
3. $\gamma \in (0,1)$ is the discount factor.
4. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space and $\mathcal{B}_{\mathcal{X}}$ be a collection measurable sets in $\mathcal{X}$. The transition kernel, $p : \mathcal{X} \times \mathcal{U} \times \mathcal{B}_{\mathcal{X}} \to [0,1]$, governs the state transitions as
$$p(x,u,B) := \mathbb{P}\left(X_{t+1} \in B \mid X_t = x, U_t = u\right) \ \forall (x,u,B) \in \mathcal{X} \times \mathcal{U} \times \mathcal{B}_{\mathcal{X}}, t \in \mathbb{N}. \tag{1}$$
5. $\bar{r} : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ be the collection of reward maps such that the agent receives reward $\bar{r}(x,u)$ whenever it takes action $u$ at state $x$.

**Assumption 1.2.** We assume the following:

1. Both $\mathcal{X}$ and $\mathcal{U}$ are finite.
2. $\bar{r} \in [0,1]$.

Given an initial state $x_0$, we want to find a policy $\pi = \{\phi_0, \phi_1, \ldots\}$ where $\phi_t : \mathcal{X} \to \Delta_{\mathcal{U}}$, $t = 0,1,\ldots$, that minimizes the value function of the policy $\pi$, defined as

$$V_\pi(x_0) := \lim_{T \to \infty} \mathbb{E}_\pi \left[ \sum_{t=0}^{T-1} \gamma^t \bar{r}(X_t, U_t) \mid X_0 = x_0 \right]. \tag{2}$$

Note that under the assumption $\bar{r} \in [0,1]$, $\sum_{t=0}^{T-1} \gamma^t \bar{r}(X_t, U_t)$ is positive, bounded above by $\frac{1}{1-\gamma}$ and is monotonically increasing with $T$. So, by the monotone convergence theorem, we can bring the limit inside expectation in the definition of $V_\pi$ and $V_\pi \leq \frac{1}{1-\gamma}$. Let $\Pi$ be the set of all admissible policies. The optimal value function $V^*$ is defined as

$$V^*(x_0) := \sup_{\pi \in \Pi} V_\pi(x_0), \ x_0 \in \mathcal{X}. \tag{3}$$

An optimal policy, for a given initial state $x$, is one that attains the optimal value $V^*(x)$. The policy may depend on $x$, but we will generally find that for most problems, an optimal policy, when it exists, may be chosen to be independent of the initial state.

## 2 Bellman Equation and Bellman Operator

Notice that under policy $\pi \in \Pi$, any initial state $x_0 \in \mathcal{X}$

$$\begin{aligned} V_\pi(x_0) &= \lim_{T\to\infty} \mathbb{E}_\pi \left[ \sum_{t=0}^{T-1} \gamma^t \bar{r}(X_t, U_t) \mid X_0 = x_0 \right] \\ &= \mathbb{E}_\pi \left[ \bar{r}(X_0, U_0) + \gamma \lim_{T\to\infty} \sum_{t=1}^{T} \gamma^{t-1} \bar{r}(X_t, U_t) \mid X_0 = x_0 \right] \\ &= \mathbb{E}_\pi \left[ \bar{r}(X_0, U_0) + \gamma V_\pi(X_1) \mid X_0 = x_0 \right]. \end{aligned} \tag{4}$$

If there exists an optimal policy $\pi^*$, then

$$\begin{aligned} V^*(x_0) &= V_{\pi^*}(x_0) \\ &= \lim_{T\to\infty} \mathbb{E}_{\pi^*} \left[ \sum_{t=0}^{T-1} \gamma^t \bar{r}(x_t, u_t) \mid X_0 = x_0 \right] \\ &= \mathbb{E}_{\pi^*} \left[ \bar{r}(X_0, U_0) + \gamma \lim_{T\to\infty} \sum_{t=1}^{T} \gamma^{t-1} \bar{r}(X_t, U_t) \mid X_0 = x_0 \right] \\ &= \mathbb{E}_{\pi^*} \left[ \bar{r}(X_0, U_0) + \gamma V_{\pi^*}(X_1) \mid X_0 = x_0 \right] \\ &= \mathbb{E}_{\pi^*} \left[ \bar{r}(X_0, U_0) + \gamma V^*(X_1) \mid X_0 = x_0 \right] \\ &= \sup_{\pi\in\Pi} \mathbb{E}_\pi \left[ \bar{r}(X_0, U_0) + \gamma V^*(X_1) \mid X_0 = x_0 \right] \\ &= \max_{u_0 \in \mathcal{U}_{x_0}} \left\{ \bar{r}(x_0, u_0) + \gamma \sum_{x_1 \in \mathcal{X}} p(x_0, u_0, x_1) V^*(x_1) \right\}. \end{aligned} \tag{5}$$

The above equation is called the Bellman equation. Next, we define the Bellman operator, $\mathcal{T}$. For any function $V : \mathcal{X} \to \mathbb{R}$, the Bellman operator maps $V$ to $\mathcal{T}V$, which is defined as

$$(\mathcal{T}V)(x) := \max_{u\in\mathcal{U}_x} \left\{ \bar{r}(x,u) + \gamma \sum_{y\in\mathcal{X}} p(x,u,y) V(y) \right\}, \ \forall x \in \mathcal{X}. \tag{6}$$

Note that $V^*$ is a fixed point of the Bellman operator. Later, we will show that $V^*$ is the unique fixed point of $\mathcal{T}$. We also define the Bellman operator corresponding to a stationary deterministic policy $\phi : \mathcal{X} \to \mathbb{R}$, $\mathcal{T}_\phi$ as follows.

$$(\mathcal{T}_\phi V)(x) := \bar{r}(x, \phi(x)) + \gamma \sum_{y\in\mathcal{X}} p(x, \phi(x), y) V(y), \ \forall x \in \mathcal{X}. \tag{7}$$

We denote by $\mathcal{T}^k$ the composition of the mapping $\mathcal{T}$ with itself $k$ times, i.e., for all $k$ we write

$$(\mathcal{T}^k V)(x) = (\mathcal{T}(\mathcal{T}^{k-1}))(x), \ x \in \mathcal{X}. \tag{8}$$

Similarly, $\mathcal{T}_\phi^k$ is defined by

$$(\mathcal{T}_\phi^k V)(x) = (\mathcal{T}_\phi(\mathcal{T}_\phi^{k-1}))(x), \ x \in \mathcal{X}. \tag{9}$$

For convenience, we also write $\mathcal{T}^0 V = V$ and $\mathcal{T}_\phi^0 V = V$. Now, we will show two elementary properties of the Bellman operator in the next two lemmas.

**Lemma 2.1** (Monotonicity Lemma)**.** *For any function $V : \mathcal{X} \to \mathbb{R}$ and $V' : \mathcal{X} \to \mathbb{R}$, such that*

$$V(x) \le V'(x), \ \forall x \in \mathcal{X},$$

*and for any stationary policy $\phi : \mathcal{X} \to \mathbb{R}$, we have*

$$\begin{aligned} (\mathcal{T}^k V)(x) &\le (\mathcal{T}^k V')(x), \ \forall x \in \mathcal{X}, k = 1, 2, \ldots, \\ (\mathcal{T}_\phi^k V)(x) &\le (\mathcal{T}_\phi^k V')(x), \ \forall x \in \mathcal{X}, k = 1, 2, \ldots. \end{aligned}$$

*Proof.* Use mathematical induction. □

Let us denote by $e : \mathcal{X} \to \mathbb{R}$ the unit function that takes the value 1 identically on $\mathcal{X}$. Let us now state and prove the next lemma.

**Lemma 2.2.** *For every $k$, function $V : \mathcal{X} \to \mathbb{R}$, stationary policy $\phi$ and scalar $\alpha$*

$$(\mathcal{T}^k(V + \alpha e))(x) \leq (\mathcal{T}^k V)(x) + \gamma^k \alpha, \ \forall x \in \mathcal{X},$$
$$(\mathcal{T}_\phi^k(V + \alpha e))(x) \leq (\mathcal{T}_\phi^k V)(x) + \gamma^k \alpha, \ \forall x \in \mathcal{X}.$$

*Proof.* Use mathematical induction. □

# 3 Value Iteration

In this section, we will first show that the Bellman operator takes any bounded function $V$ to $V^*$ asymptotically when applied repetitively, and the convergence speed of the iterates is geometric with rate $\gamma$.

**Proposition 3.1** (Convergence of value iteration)**.** *For any bounded function $V : \mathcal{X} \to \mathbb{R}$, the optimal value function satisfies*

$$V^*(x) = \lim_{k \to \infty} (\mathcal{T}^k V)(x), \ \forall x \in \mathcal{X}.$$

*Proof.* For every positive integer $K$, initial state $x_0 \in \mathcal{X}$, and policy $\pi = \{\phi_0, \phi_1, \ldots\}$, we break down the cost $V_\pi(x_0)$ into the portions incurred over the first $K$ stages and over the remaining stages, i.e.,

$$\begin{aligned} V_\pi(x_0) &= \lim_{T \to \infty} \mathbb{E}_\pi \left[ \sum_{t=0}^{T-1} \gamma^t \bar{r}(X_t, U_t) \mid X_0 = x_0 \right] \\ &= \mathbb{E}_\pi \left[ \sum_{t=0}^{K-1} \gamma^t \bar{r}(X_t, U_t) \mid X_0 = x_0 \right] + \lim_{T \to \infty} \mathbb{E}_\pi \left[ \sum_{t=K}^{T-1} \gamma^t \bar{r}(X_t, U_t) \mid X_0 = x_0 \right]. \end{aligned}$$

Since $\bar{r}(x, u) \in [0, 1]$, we also obtain

$$\lim_{T \to \infty} \mathbb{E}_\pi \left[ \sum_{t=K}^{T-1} \gamma^t \bar{r}(X_t, U_t) \mid X_0 = x_0 \right] \leq \sum_{t=K}^{\infty} \gamma^t = \frac{\gamma^K}{1 - \gamma}.$$

Using these relations, it follows that

$$V_\pi(x_0) - \frac{\gamma^K}{1-\gamma} - \gamma^K \max_{x \in \mathcal{X}} |V(x)| \leq \mathbb{E} \left[ \gamma^K V(x_K) + \sum_{t=0}^{K-1} \gamma^t \bar{r}(X_t, U_t) \mid x_0 \right] \leq V_\pi(x_0) + \frac{\gamma^K}{1-\gamma} + \gamma^K \max_{x \in \mathcal{X}} |V(x)|.$$

By taking the minimum over $\pi$, we obtain for all $x_0$, $K$,

$$V^*(x_0) - \frac{\gamma^K}{1-\gamma} - \gamma^K \max_{x \in \mathcal{X}} |V(x)| \leq (\mathcal{T}^K V)(x_0) \leq V^*(x_0) + \frac{\gamma^K}{1-\gamma} + \gamma^K \max_{x \in \mathcal{X}} |V(x)|,$$

and by taking limit $K \to \infty$, the result follows. □

**Corollary 3.2** (Bellman equation)**.** *The optimal value function satisfies the Bellman equation, i.e.,*

$$V^*(x) = \max_{u \in \mathcal{U}_x} \mathbb{E} \left[ \bar{r}(x, u) + \gamma V(X_1) \mid X_0 = x \right], \ \forall x \in \mathcal{X},$$

*or, equivalently,*

$$V^* = \mathcal{T} V^*.$$

*Furthermore, $V^*$ is the unique solution of this equation of bounded functions.*

*Proof.* Let $V_0(x) = 0$ for all $x \in \mathcal{X}$. So, we have that

$$V^*(x) - \frac{\gamma^K}{1-\gamma} \leq (\mathcal{T}^K V_0)(x) \leq V^*(x) + \frac{\gamma^K}{1-\gamma}.$$

Applying the operator $\mathcal{T}$ to this relation and using Lemma 2.1 and Lemma 2.2, we obtain for all $x \in \mathcal{X}$ and $K$

$$(\mathcal{T}V^*)(x) - \frac{\gamma^K}{1-\gamma} \leq (\mathcal{T}^{K+1} V_0)(x) \leq (\mathcal{T}V^*)(x) + \frac{\gamma^K}{1-\gamma}.$$

By taking the limit as $K \to \infty$ in the preceding relation and using the fact

$$\lim_{K\to\infty} (\mathcal{T}^{K+1} V_0)(x) = V^*(x),\ \forall x \in \mathcal{X},$$

we obtain $V^* = \mathcal{T}V^*$.

To show uniqueness, observe that if $V$ is bounded and satisfies $V = \mathcal{T}V$, then $V = \lim_{K\to\infty} \mathcal{T}^K V$, so by proposition 3.1, we have $V = V^*$. □

Solving an MDP essentially means finding an optimal policy for the MDP. The next proposition will help us to find the optimal policy.

**Proposition 3.3** (Necessary and sufficient condition for optimality)**.** *A stationary policy $\phi$ is optimal if and only if $\phi(x)$ attains the minimum in Bellman's equation* (5) *for each $x \in \mathcal{X}$, i.e.,*

$$\mathcal{T}V^* = \mathcal{T}_\phi V^*.$$

*Proof.* If $\mathcal{T}V^* = \mathcal{T}_\phi V^*$, then that $V^* = \mathcal{T}_\phi V^*$. Also, $V_\phi$ is the unique fixed point of $\mathcal{T}_\phi$. Hence, $V_\phi = V^*$ and $\phi$ is optimal.

On the other hand, if $\phi$ is optimal, we have $V^* = V_\phi$, which means $J^* = \mathcal{T}_\phi J^*$. Combining this with Bellman's equation, we obtain $\mathcal{T}J^* = \mathcal{T}_\phi J^*$. □

**Proposition 3.4.** *For any two bounded function $V : \mathcal{X} \to \mathbb{R}$, $V' : \mathcal{X} \to \mathbb{R}$, and for all $k = 0, 1, \ldots$, there holds*

$$\max_{x\in\mathcal{X}} \left| (\mathcal{T}^k V)(x) - (\mathcal{T}^k V')(x) \right| \leq \gamma^k \max_{x\in\mathcal{X}} |V(x) - V'(x)|. \tag{10}$$

*Proof.* Denote $c := \max_{x\in X} |V(x) - V'(x)|$. Then we have

$$V(x) - c \leq V'(x) \leq V(x) + c,\ \forall x \in X.$$

Applying $\mathcal{T}^k$ in the above relation and using Lemma 2.1 and Lemma 2.2, we get that

$$(\mathcal{T}^k V)(x) - \gamma^k c \leq (\mathcal{T}^k V')(x) \leq (\mathcal{T}^k V)(x) + \gamma^k c,\ \forall x \in X.$$

It follows that

$$\left| (\mathcal{T}^k V)(x) - (\mathcal{T}^k V')(x) \right| \leq \gamma^k c,\ \forall x \in X, \tag{11}$$

which proves the result. □

**Corollary 3.5.**

$$\lim_{k\to\infty} \mathcal{T}^k V = V^*. \tag{12}$$

*Also, $V^*$ is the unique fixed point of the map $\mathcal{T}$.*

**Proposition 3.6.**

Here is the value iteration algorithm for discounted MDP:

**Algorithm 1** Value Iteration

Input: Discounted MDP, $M$; error tolerance, $\epsilon$.
Initilize $V_0 = 0$, $k = 0$
**while** True **do**
  $k \leftarrow k + 1$
  **for** $x \in X$ **do**
    $V_k(x) = \max_{u \in U} r(x,u) + \gamma \sum_{y \in X} p(x,u,y) V_{k-1}(y)$
  **end for**
  $\phi_k(x) \in \arg\max_{u \in U} r(x,u) + \sum_{y \in X} p(x,u,y) V_{k-1}(y)$
  **if** $\max_{x \in \mathcal{X}} |V_k(x) - V_{k-1}(x)| < \epsilon$ **then**
    Break
  **end if**
**end while**
Return $V_k$, $\pi_k$

# 4 Examples