Learning and Optimization in Multiagent Decision-Making Systems
Lecture Notes on Markov Chains and Markov Decision Processes (by courtesy from R. Srikant)
Instructor: Rasoul Etesami

Discrete-time Markov Chains

A *stochastic process* or a *random process* is an indexed collection of random variables, and if the index set is discrete, then it is called a *discrete-time* process.

Let $\{X_k\}$ denote a discrete-time random process that takes values in a finite set $\mathscr{S}$ called the *state space*. A random process $\{X_k\}$ is called a *discrete-time Markov chain (DTMC)*, or simply a *Markov chain*, if for any sequence,

$$\mathbb{P}(X_k = s_k \mid X_{k-1} = s_{k-1}, X_{k-2} = s_{k-2}, \ldots, X_0 = s_0) = \mathbb{P}(X_k = s_k \mid X_{k-1} = s_{k-1}),$$

for any choice of $s_i \in \mathscr{S}$ and for all k.

Let p_k denote the probability distribution of the random variable X_k, i.e., a row vector of probabilities with $p_k(s) = \mathbb{P}(X_k = s)$. Assuming $\mathscr{S} = \{s_1, s_2, \ldots, s_n\}$, the distribution p_k is given by

$$p_k = (p_k(s_1), p_k(s_2), \ldots, p_k(s_n)) = (\mathbb{P}(X_k = s_1), \mathbb{P}(X_k = s_2), \ldots, \mathbb{P}(X_k = s_n)).$$

By the Markov property, the evolution of the probability distribution p_k is governed by

$$p_k(s) = \sum_{s' \in \mathscr{S}} p_{k-1}(s')\, \mathbb{P}(X_k = s \mid X_{k-1} = s').$$

Let p_k denote the probability distribution of the random variable X_k, i.e., a row vector of probabilities $p_k(s) = P(X_k = s)$. Assuming $S = \{s_1, s_2, \ldots, s_n\}$,

$$p_k = (p_k(s_1), p_k(s_2), \ldots, p_k(s_n)) = (P(X_k = s_1), P(X_k = s_2), \ldots, P(X_k = s_n)).$$

By the Markov property, the evolution of p_k is given by

$$p_k(s) = \sum_{s' \in S} p_{k-1}(s') P(X_k = s \mid X_{k-1} = s').$$

A Markov chain is said to be *time homogeneous* if $\mathbb{P}(X_k = s \mid X_{k-1} = s')$ is independent of the time index k. In this discussion, we only consider time-homogeneous Markov chains.

Associated with each time-homogeneous Markov chain is a matrix called the *transition probability matrix*, denoted by P, whose (s', s)-th element is given by

$$P(s', s) = \mathbb{P}(X_1 = s \mid X_0 = s').$$

Notice that

$$P^{\ell}(s', s) = \mathbb{P}(X_{\ell} = s \mid X_0 = s').$$

Using the transition matrix P, the evolution of p_k can be written in vector form as:

$$p_k = p_{k-1} P = p_0 P^k. \tag{13}$$

Thus, p_0 and P capture all the relevant information about the dynamics of the Markov chain. The matrix P can be encoded into a weighted directed graph, called the *transition diagram*, where the vertex set is $\mathscr{S}$, and the weight of an edge (s', s) equals $P(s', s)$.

Example Consider movements of a knight on a 4×4 chessboard starting at the most left corner.

1. X_k = knight's location; randomly walks: It is a Markov chain.
2. Y_k = knight's location with no backtracking: It is not a Markov chain.
3. $U_k = 0$ if on white, 1 otherwise; no backtracking: It is a Markov chain.
4. $Z_k = (Y_k, Y_{k-1})$: It is a Markov chain due to inclusion of sufficient history.

This example highlights the fact that the same dynamics may or may not be a Markov process depending on the choice of the state space.

Definition 40. *A state s is called reachable from s' if $\exists l \geq 1$ s.t. $P^l(s', s) > 0$. A Markov chain is* ***irreducible*** *if every state is reachable from any other state.*

Definition 41. *The period of a state $s \in \mathscr{S}$ is defined by $d_s = \gcd\{l : P^l(s,s) > 0\}$. A state s is* ***aperiodic*** *if $d_s = 1$. A Markov chain is aperiodic if all states are aperiodic.*

Lemma 42. *In an irreducible chain, all states share the same period. Thus, if one is aperiodic, all are, and the Markov chain becomes aperiodic.*

Theorem 43. *A finite, irreducible Markov chain has a unique stationary distribution μ. If also aperiodic, then $\lim_{k\to\infty} p_k = \mu$ for all p_0. Specifically, $\lim_{k\to\infty} P^k = 1^T \mu$.*

Example Consider a Markov chain with state space $\{1, 2, 3\}$ and the transition matrix:

$$P = \begin{bmatrix} 1/2 & 1/4 & 1/4 \\ 0 & 1/2 & 1/2 \\ 1 & 0 & 0 \end{bmatrix}$$

It is easy to see that the chain is irreducible, and state 1 is aperiodic. So, the chain is ergodic. We can solve for the stationary distribution $\mu = \mu P$ to get

$$\mu = [\mu(1), \mu(2), \mu(3)] = [1/2, 1/4, 1/4]$$

Now let T_1 be the first return time to state 1. Then the expected value of T_1 equals

$$E[T_1] = 1 \cdot \tfrac{1}{2} + 2 \cdot \tfrac{1}{4} + \sum_{k=0}^{\infty} (k+3) \cdot \tfrac{1}{8} \cdot \left(\tfrac{1}{2}\right)^k = 2.$$

An interesting observation in the above example is that $\mu(1) = \frac{1}{\mathbb{E}[T_1]}$. The following lemma confirms that this holds for any irreducible Markov chain.

Lemma 44. *The stationary distribution of a finite state space, irreducible Markov chain satisfies*

$$\mu(i) = \frac{1}{\mathbb{E}[T_i]},$$

where T_i is the first return time to state i.

Intuitively speaking, $1/\mathbb{E}[T_i]$ is the frequency of visits to state i. The ergodic theorem confirms this interpretation and enables us to estimate μ using a single sample path.

Theorem 45 (Ergodic Theorem). *For a finite state space, irreducible Markov chain $\{X_k\}$, we have*

$$\mathbb{P}\left(\lim_{\ell\to\infty}\frac{1}{\ell}\sum_{k=0}^{\ell-1}\mathbf{1}(X_k=i)=\frac{1}{\mathbb{E}[T_i]}\right)=1.$$

More generally, for any function $c:\mathscr{S}\to\mathbb{R}$, we have

$$\mathbb{P}\left(\lim_{\ell\to\infty}\frac{1}{\ell}\sum_{k=0}^{\ell-1}c(X_k)=\sum_{i\in\mathscr{S}}\mu(i)c(i)\right)=1,$$

and in particular,

$$\lim_{\ell\to\infty}\mathbb{E}\left[\frac{1}{\ell}\sum_{k=0}^{\ell-1}c(X_k)\right]=\sum_{i\in\mathscr{S}}\mu(i)c(i).$$

The above theorem is one of the most fundamental results in the theory of Markov chains. In particular, assuming there is a cost $c(\cdot)$ associated with the state of the Markov chain, the ergodic theorem states that the average cost observed on the sample path converges almost surely and in mean to

$$\sum_{i\in\mathscr{S}}\mu(i)c(i).$$

Example Consider a a wireless link with Bernoulli(λ) arrivals, Bernoulli(ν) service, buffer size B. Define state as number of packets in the buffer. Therefore, the state space equals $S=\{0,1,\ldots,B\}$. In order to find the stationary distribution, one needs to solve the equation

$$\mu=\mu P.$$

However, this task is almost impossible when the size of the matrix is large. Instead, we find the stationary distribution using the *flow equations*: if we partition the transition diagram into two sections A and B, the total transition flow from A to B in the steady state should be equal to the total flow from B to A. That is,

$$\sum_{i\in A}\sum_{j\in B}\mu(i)P(i,j)=\sum_{j\in B}\sum_{i\in A}\mu(j)P(j,i).$$

$$\mu(0)\lambda(1-\nu)=\mu(1)\nu(1-\lambda)\quad\Rightarrow\quad\mu(1)=\frac{\lambda(1-\nu)}{\nu(1-\lambda)}\mu(0)$$

$$\mu(1)\lambda(1-\nu)=\mu(2)\nu(1-\lambda)\quad\Rightarrow\quad\mu(2)=\left(\frac{\lambda(1-\nu)}{\nu(1-\lambda)}\right)^2\mu(0)$$

$$\vdots$$

$$\mu(B-2)\lambda(1-\nu)=\mu(B-1)\nu(1-\lambda)\quad\Rightarrow\quad\mu(B-1)=\left(\frac{\lambda(1-\nu)}{\nu(1-\lambda)}\right)^{B-1}\mu(0)$$

$$\mu(B-1)\lambda(1-\nu)=\mu(B)\nu\quad\Rightarrow\quad\mu(B)=(1-\lambda)\left(\frac{\lambda(1-\nu)}{\nu(1-\lambda)}\right)^{B}\mu(0)$$

The above equations together with

$$\sum_i\mu(i)=1,$$

give the stationary distribution of the Markov chain.

Markov Decision Processes

A Markov Decision Process (MDP) is a controlled Markov chain, i.e., a Markov chain in which the transition probabilities depend on an exogenous control parameter called action. More specifically, the probability of transitioning to state s' from state s upon taking action a is denoted by $P_{s,s'}(a)$ and is given by

$$P_{s,s'}(a) := \mathbb{P}(S_1 = s' \mid S_0 = s, A_0 = a), \quad \forall s, s' \in \mathscr{S}, \forall a \in \mathscr{A}_s.$$

In addition to the state space $\mathscr{S}$, for each $s \in \mathscr{S}$, we are given a finite action space $\mathscr{A}_s$ which denotes the set of possible actions at state s. For simplicity, we drop the dependency on s and assume the action space is $\mathscr{A}$ for all s.

Another component is a one-step random reward function $r : \mathscr{S} \times \mathscr{A} \to \mathbb{R}$ that assigns rewards to each (s, a) pair. We denote its expectation as $\bar{r}(s, a) = \mathbb{E}[r(s, a)]$. Assuming r is bounded, we can shift it to be non-negative without loss of generality.

Definition 46. *An MDP is a tuple $(\mathscr{S}, \mathscr{A}, P, r)$ where:*

- *$\mathscr{S}$ is the finite state space;*
- *$\mathscr{A}$ is the finite action space;*
- *$P_{s,s'}(a) = \mathbb{P}(S_1 = s' \mid S_0 = s, A_0 = a)$ is the transition probability;*
- *$r : \mathscr{S} \times \mathscr{A} \to [0, r_{\max}]$ is the one-step random reward function.*

Definition 47. *A policy is a probabilistic rule for choosing actions at each time step. Formally, a policy is a family of distributions $\pi = \{\pi_k\}_{k \geq 0}$ where $\pi_k : \mathscr{S}^{k+1} \times \mathscr{A}^k \to \Delta_{\mathscr{A}}$:*

$$\pi_k(a \mid \{s_i, a_i\}_{i=0}^{k-1}, s_k) := \mathbb{P}(A_k = a \mid S_0 = s_0, A_0 = a_0, \ldots, S_k = s_k).$$

*If a policy only depends on the current state, i.e., $\pi_k(a \mid \{s_i, a_i\}_{i=0}^{k-1}, s_k) = \pi_k(a \mid s_k)\ \forall k$, then it is called a **Markov** policy. If further, the policy is independent of time-step k, i.e., $\pi_k = \pi$ for all k, it is called a **stationary** policy.*

Remark 11. *Under a stationary policy π, the MDP behaves like a Markov chain with transition matrix $P^\pi = [P^\pi_{s,s'}]$, where*

$$P^\pi_{s,s'} = \sum_{a \in \mathscr{A}} \pi(a \mid s) P_{s,s'}(a), \quad \forall s, s' \in \mathscr{S}.$$

Performance Measure: Actions in an MDP are chosen to optimize some performance measure. One such measure is the **expected discounted total reward**, given by:

$$\mathbb{E}\left[\sum_{k=0}^{L} \gamma^k r(S_k, A_k)\right] \quad \text{or} \quad \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r(S_k, A_k)\right],$$

depending on whether the horizon is finite or infinite. The discount factor γ discounts future rewards: If horizon is finite: $\gamma \in [0, 1]$. Otherwise, if horizon is infinite: $\gamma \in [0, 1)$. Note that the infinite sum is well-defined since $r : \mathscr{S} \times \mathscr{A} \to [0, r_{\max}]$ is bounded.

Given an MDP $(\mathscr{S}, \mathscr{A}, P, r)$ and discount factor γ, the main goal is to find an optimal policy $\{\pi^*_k\}_{k \geq 0}$ that maximizes the expected discounted total reward. Fortunately, it suffices to focus on Markov policies.

Theorem 48. *For an MDP $(\mathscr{S}, \mathscr{A}, P, r)$ with discount factor γ (finite horizon or infinite horizon), there exists a* ***Markov policy*** $\{\pi_k^*\}_{k\geq 0}$ *that maximizes the expected discounted total reward.*

Our goal is to develop a systematic approach to find the optimum policy or evaluate the performance of a fixed policy for a given MDP.

Finite Horizon MDPs

Let $(\mathscr{S}, \mathscr{A}, P, r, \gamma, L)$ denote a discounted finite-horizon MDP, with discount factor $\gamma \in [0,1]$, and horizon $L < \infty$. Our goal is to find a Markov policy $\{\pi_k\}_{k=0}^L$ that maximizes the performance measure:

$$\mathbb{E}\left[\sum_{k=0}^{L} \gamma^k r(S_k, A_k)\right].$$

For a given policy $\pi = \{\pi_k\}_{k=0}^L$, define value functions $\{V_{k\to L}^\pi\}_{k=0}^L$, where each $V_{k\to L}^\pi : \mathscr{S} \to \mathbb{R}_+$ is the expected total discounted reward from time-step k to L, conditioned on $S_k = s$:

$$V_{k\to L}^\pi(s) := \mathbb{E}^\pi\left[\sum_{l=k}^{L} \gamma^{l-k} r(S_l, A_l) \,\middle|\, S_k = s\right].$$

Define expected reward and transition probability under policy π_l:

$$\bar{r}(s, \pi_l(s)) := \sum_{a\in\mathscr{A}} \pi_l(a \mid s)\bar{r}(s,a), \quad P_{s,s'}^{\pi_l} := \sum_{a\in\mathscr{A}} \pi_l(a \mid s) P_{s,s'}(a).$$

Then we can recursively express:

$$\begin{aligned}
V_{k\to L}^\pi(s) &= \mathbb{E}_\pi\left[\sum_{l=k}^{L} \gamma^{l-k} r(S_l, \pi_l(S_l)) \,\middle|\, S_k = s\right] \\
&= \bar{r}(s, \pi_k(s)) + \gamma\, \mathbb{E}_\pi\left[\sum_{l=k+1}^{L} \gamma^{l-k-1} r(S_l, \pi_l(S_l)) \,\middle|\, S_k = s\right] \\
&= \bar{r}(s, \pi_k(s)) + \gamma \sum_{s'\in\mathscr{S}} P_{s,s'}^{\pi_k}\, \mathbb{E}_\pi\left[\sum_{l=k+1}^{L} \gamma^{l-k-1} r(S_l, \pi_l(S_l)) \,\middle|\, S_{k+1} = s'\right].
\end{aligned}$$

Notice that the expectation in the right-hand side of the above equation equals $V_{k+1\to L}^\pi(s')$. Therefore, we have

$$V_{k\to L}^\pi(s) = \bar{r}(s, \pi_k(s)) + \gamma \sum_{s'\in\mathscr{S}} P_{s,s'}^{\pi_k} V_{k+1\to L}^\pi(s').$$

The functions $\{V_{k\to L}^\pi\}_{k=0}^L$ are called (state-)value functions, and they can be calculated using the following recursive equations, called the *Bellman equations*:

$$\begin{aligned}
&V_{L\to L}^\pi(s) = \bar{r}(s, \pi_L(s)), \\
&V_{k\to L}^\pi(s) = \bar{r}(s, \pi_k(s)) + \gamma \sum_{s'\in\mathscr{S}} P_{s,s'}^{\pi_k} V_{k+1\to L}^\pi(s'), \quad \forall k = 0, \ldots, L-1.
\end{aligned}$$

Abusing notation, we can rewrite the Bellman equations in vector form. To that end, let $\bar{r}^{\pi_k}$ denote the vector with entries $\bar{r}^{\pi_k}(s) := \bar{r}(s, \pi_k(s))$. Then:

$$V_{L\to L}^\pi = \bar{r}^{\pi_L}, \quad V_{k\to L}^\pi = \bar{r}^{\pi_k} + \gamma P^{\pi_k} V_{k+1\to L}^\pi \;\; \forall k = 0, \ldots, L-1.$$

Example: Consider a finite-horizon discounted MDP with $\gamma = 0.1$, $L = 2$, $\mathscr{S} = \{1,2,3\}$, $\mathscr{A} = \{a,b\}$, and transition probabilities:

$$P(a) = \begin{bmatrix} 0.2 & 0.2 & 0.6 \\ 0.3 & 0.4 & 0.3 \\ 0 & 1 & 0 \end{bmatrix}, \quad P(b) = \begin{bmatrix} 0.4 & 0.2 & 0.4 \\ 0.2 & 0.7 & 0.1 \\ 0 & 0.8 & 0.2 \end{bmatrix},$$

Also, assume the expected reward vector $\bar{r} = r$ has the form

$$\begin{aligned} &\bar{r}(1,a) = 2, \quad \bar{r}(1,b) = 1, \\ &\bar{r}(2,a) = -0.5, \quad \bar{r}(2,b) = 0, \\ &\bar{r}(3,a) = 3, \quad \bar{r}(3,b) = 1. \end{aligned}$$

Consider the stationary policy π that takes an action uniformly at random at each time-step. Calculate the family of value functions associated with this policy.

Solution: For such a uniform stationary policy π, we have

$$\bar{r}^{\pi} = \begin{bmatrix} 1.5 \\ -0.25 \\ 2 \end{bmatrix}, \quad P^{\pi} = \frac{1}{2}(P(a) + P(b)) = \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.25 & 0.55 & 0.2 \\ 0 & 0.9 & 0.1 \end{bmatrix}.$$

Then, the Bellman equations can be calculated recursively as

$$V^{\pi}_{2\to 2} = \bar{r}^{\pi} = \begin{bmatrix} 1.5 \\ -0.25 \\ 2 \end{bmatrix}, \quad V^{\pi}_{1\to 2} = \bar{r}^{\pi} + \gamma P^{\pi} V^{\pi}_{2\to 2} = \begin{bmatrix} 1.64 \\ -0.1892 \\ 1.9975 \end{bmatrix},$$

$$V^{\pi}_{0\to 2} = \bar{r}^{\pi} + \gamma P^{\pi} V^{\pi}_{1\to 2} = \begin{bmatrix} 1.6454 \\ -0.1793 \\ 2.0032 \end{bmatrix}.$$

Optimal Policies in Finite-Horizon MDPs

Our goal is to find a Markov policy that maximizes the expected discounted total reward over the horizon L. Specifically, we solve:

$$\max_{\pi = \{\pi_l\}_{l=0}^{L}} \mathbb{E}_{\pi}\left[\sum_{l=0}^{L} \gamma^l r(S_l, A_l) \,\middle|\, S_0 = s\right].$$

Let $V^*_{k\to L}(s)$ be the maximum expected discounted reward from time k to L:

$$V^*_{k\to L}(s) = \max_{\{\pi_l\}_{l=k}^{L}} \mathbb{E}_{\pi}\left[\sum_{l=k}^{L} \gamma^{l-k} r(S_l, A_l) \,\middle|\, S_k = s\right].$$

By the same logic as in deriving Bellman equations, we have

$$\begin{aligned}
V^*_{0\to L}(s) &= \max_{\pi=\{\pi_l\}_{l=0}^L} \mathbb{E}_\pi\left[\sum_{l=0}^{L}\gamma^l r(S_l,A_l)\,\middle|\, S_0=s\right]\\
&= \max_{\pi_0}\left(\mathbb{E}_{\pi_0}[r(S_0,A_0)\,|\,S_0=s]+\gamma\max_{\pi=\{\pi_l\}_{l=1}^L}\mathbb{E}_\pi\left[\sum_{l=1}^{L}\gamma^{l-1} r(S_l,A_l)\,\middle|\, S_0=s\right]\right)\\
&= \max_{\pi_0}\left(\bar r(s,\pi_0(s))+\gamma\max_{\pi=\{\pi_l\}_{l=1}^L}\sum_{s'\in\mathscr{S}}P^{\pi_0}_{s,s'}\,\mathbb{E}_\pi\left[\sum_{l=1}^{L}\gamma^{l-1} r(S_l,A_l)\,\middle|\, S_1=s'\right]\right)\\
&= \max_{\pi_0}\left(\bar r(s,\pi_0(s))+\gamma\sum_{s'\in\mathscr{S}}P^{\pi_0}_{s,s'}V^*_{1\to L}(s')\right).
\end{aligned}$$

Notice that given $V^*_{1\to L}:\mathscr{S}\to\mathbb{R}_+$, the right-hand side of the above equation is maximized for a deterministic policy π_0. Hence, we have

$$V^*_{0\to L}(s)=\max_{a\in\mathscr{A}}\left\{\bar r(s,a)+\gamma\sum_{s'\in\mathscr{S}}P_{s,s'}(a)V^*_{1\to L}(s')\right\}.$$

The same argument generalizes to $\{V^*_{k\to L}\}_{k=0}^L$, which results in the following set of recursive equations known as *Bellman optimality equations* or *dynamic programming equations*:

$$\begin{aligned}
V^*_{L\to L}(s) &= \max_{a\in\mathscr{A}}\bar r(s,a),\\
V^*_{k\to L}(s) &= \max_{a\in\mathscr{A}}\left(\bar r(s,a)+\gamma\sum_{s'}P_{s,s'}(a)V^*_{k+1\to L}(s')\right),\quad \forall k=0,\dots,L-1.
\end{aligned}$$

Once we compute $\{V^*_{k\to L}\}$, we define the optimal deterministic policy $\pi^*=\{\pi^*_k\}_{k=0}^L$:

$$\begin{aligned}
\pi^*_L(a\mid s) &= \begin{cases}1 & \text{if } a=\arg\max_{a'\in\mathscr{A}}\bar r(s,a'),\\ 0 & \text{otherwise,}\end{cases}\\
\pi^*_k(a\mid s) &= \begin{cases}1 & \text{if } a=\arg\max_{a'\in\mathscr{A}}\left(\bar r(s,a')+\gamma\sum_{s'}P_{s,s'}(a')V^*_{k+1\to L}(s')\right),\\ 0 & \text{otherwise,}\end{cases}\qquad \forall k=0,\dots,L-1,
\end{aligned}$$

where ties are broken arbitrarily.

Example Find the optimal deterministic policy for the MDP in the previous example with $L=3$ and $\gamma=0.65$. Moreover, determine i) the best action A_0 if $S_0=2$, and ii) the best action A_1 if $S_1=2$.

Solution. Using the Bellman optimality equation, we have

$$V^*_{3\to 3}=\begin{bmatrix}2\\0\\3\end{bmatrix},\quad \pi^*_3(a\mid 1)=1,\quad \pi^*_3(b\mid 2)=1,\quad \pi^*_3(a\mid 3)=1,$$

$$V^*_{2\to 3}=\begin{bmatrix}3.43\\0.475\\3\end{bmatrix},\quad \pi^*_2(a\mid 1)=1,\quad \pi^*_2(a\mid 2)=1,\quad \pi^*_2(a\mid 3)=1,$$

$$V^*_{1\to 3} = \begin{bmatrix} 3.6776 \\ 0.8773 \\ 3.3087 \end{bmatrix}, \quad \pi^*_1(a \mid 1) = 1, \quad \pi^*_1(a \mid 2) = 1, \quad \pi^*_1(a \mid 3) = 1,$$

$$V^*_{0\to 3} = \begin{bmatrix} 3.8826 \\ 1.0924 \\ 3.5703 \end{bmatrix}, \quad \pi^*_0(a \mid 1) = 1, \quad \pi^*_0(b \mid 2) = 1, \quad \pi^*_0(a \mid 3) = 1.$$

Thus, if $S_0 = 2$, best action is $A_0 = b$, and if $S_1 = 2$, the best action is $A_1 = a$. In particular, notice that the optimal policy π^* is not stationary.

Infinite Horizon MDPs

Let $(\mathscr{S}, \mathscr{A}, P, r, \gamma)$ denote a discounted infinite-horizon MDP with discount factor $\gamma \in [0, 1)$. The objective is to find a Markov policy $\pi = \{\pi_k\}_{k\geq 0}$ that maximizes:

$$\mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r(S_k, A_k) \right].$$

Similar to finite-horizon MDPs, associated with each Markov policy $\pi = \{\pi_k\}_{k=0}^{\infty}$, there is a set of value functions $\{V^\pi_{k\to\infty}\}_{k\geq 0}$. For now, we focus on stationary policies. Later on, we will show that there exists a stationary policy that maximizes the expected discounted total reward.

For each stationary policy π, define the value function $V^\pi : \mathscr{S} \to \mathbb{R}_+$:

$$V^\pi(s) := \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r(S_k, A_k) \,\middle|\, S_0 = s \right].$$

Following a similar idea as in finite-horizon MDPs, let us rewrite $V^\pi(s)$ as follows:

$$\begin{aligned} V^\pi(s) &= \mathbb{E}_\pi \left[r(S_0, A_0) + \gamma \sum_{k=1}^{\infty} \gamma^{k-1} r(S_k, A_k) \,\middle|\, S_0 = s \right] \\ &= \bar{r}(s, \pi(s)) + \gamma \sum_{s' \in \mathscr{S}} P^\pi_{s,s'} \, \mathbb{E}^\pi \left[\sum_{k=1}^{\infty} \gamma^{k-1} r(S_k, A_k) \,\middle|\, S_1 = s' \right] \\ &= \bar{r}(s, \pi(s)) + \gamma \sum_{s' \in \mathscr{S}} P^\pi_{s,s'} \, V^\pi(s') \end{aligned}$$

where $\bar{r}(s, \pi(s)) := \sum_{a \in \mathscr{A}} \pi(a \mid s) \bar{r}(s, a)$ and $P^\pi_{s,s'} := \sum_{a \in \mathscr{A}} \pi(a \mid s) P_{s,s'}(a)$, and the last equality follows from the fact that the policy π is stationary. Rewriting it in the vector form, we have:

$$V^\pi = \bar{r}^\pi + \gamma P^\pi V^\pi$$

Hence, V^π satisfies a fixed point equation. In particular, if $I - \gamma P^\pi$ is invertible, then we have

$$V^\pi = (I - \gamma P^\pi)^{-1} \bar{r}^\pi.$$

It is easy to check that $I - \gamma P^\pi$ is invertible. In fact, its inverse is given by

$$(I - \gamma P^\pi)^{-1} = \sum_{k=0}^{\infty} (\gamma P^\pi)^k,$$

where $(P^\pi)^0 := I$. However, calculating the matrix inverse is computationally expensive.

Associated with each policy π, there is the operator $T^\pi : \mathbb{R}^{|\mathscr{S}|} \to \mathbb{R}^{|\mathscr{S}|}$ defined as

$$T^\pi(V) = \bar{r}^\pi + \gamma P^\pi V,$$

which is called the *Bellman operator*. By the above discussion, the value function V^π is the unique fixed point of the operator T^π.

Remark 12. *Notice the similarity between the Bellman equation $T^\pi(V^\pi) = V^\pi$ and the Bellman equations for finite-horizon MDPs. Specifically, since the horizon is infinite here and the policy is stationary, the value function does not depend on time.*

Proposition 49. *The operator T^π is a contraction in ℓ_∞ norm with contraction factor γ:*

$$\|T^\pi(V) - T^\pi(V')\|_\infty \le \gamma \|V - V'\|_\infty,$$

where $\|V\|_\infty = \max_{s\in\mathscr{S}} |V(s)|$.

Proof: For any fixed $s \in \mathscr{S}$, we have

$$\begin{aligned} |T^\pi(V)(s) - T^\pi(V')(s)| &= \gamma \left| \sum_{s'} P^\pi_{s,s'} (V(s') - V'(s')) \right| \\ &\le \gamma \sum_{s'} P^\pi_{s,s'} |V(s') - V'(s')| \\ &\le \gamma \|V - V'\|_\infty \sum_{s'} P^\pi_{s,s'} = \gamma \|V - V'\|_\infty. \end{aligned}$$

Since this holds for all s, the operator T^π is a γ-contraction. □

A known general result for contraction operators is the contraction mapping theorem. It provides an algorithmic way to find the unique fixed point of a contraction operator defined on a general Banach spaces.

Theorem 50. *Let $U : \mathbb{R}^n \to \mathbb{R}^n$ be a contraction with factor $\gamma \in [0, 1)$ with respect to some norm $\|\cdot\|$, i.e.,*

$$\|U(x) - U(y)\| \le \gamma \|x - y\|.$$

Then U has a unique fixed point $U(x^*) = x^*$ such that for any $x_0 \in \mathbb{R}^n$:

$$\|U^k(x_0) - x^*\| \le \gamma^k \|x_0 - x^*\|, \quad \text{and} \quad \lim_{k\to\infty} U^k(x_0) = x^*.$$

Hence, to compute V^π, we can apply value iteration:

$$V_{k+1} := T^\pi(V_k) = \bar{r}^\pi + \gamma P^\pi V_k.$$

By Theorem 50 and Proposition 49, for any initial value vector V_0, we have

$$\|(T^\pi)^k(V_0) - V^\pi\|_\infty \le \gamma^k \|V_0 - V^\pi\|_\infty.$$

Optimal Policy for Infinite-Horizon MDPs

Next, we want to find the optimum stationary policy π^* for which $V^{\pi^*}(s) \geq V^{\pi}(s)$ for all stationary policies π. Notice that the existence of such a policy is not clear at this point. Nevertheless, let us pretend such an optimal policy exists.

Let V^* denote the value function of the optimal stationary policy π^*. Following the same argument as in the case of finite-horizon MDPs, we may expect V^* to satisfy some Bellman optimality equations. Since the policy π^* is stationary, the value function should not depend on time, as we discussed above. Hence, intuitively speaking, we may expect the value function V^* to satisfy the following Bellman optimality equation:

$$V^*(s) = \max_{a \in \mathscr{A}} \left(\bar{r}(s,a) + \gamma \sum_{s' \in \mathscr{S}} P_{s,s'}(a) V^*(s') \right).$$

We will show that:

1. The above equation has a unique solution which corresponds to the value function of some deterministic stationary policy π^*;
2. $V^*(s) \geq V^{\pi}(s)$ for all stationary policies π;
3. $V^*(s) \geq V^{\pi}(s)$ for any (possibly non-stationary) Markov policy π.

Let $T : \mathbb{R}^{|\mathscr{S}|} \to \mathbb{R}^{|\mathscr{S}|}$ denote the *Bellman optimality operator*, which is defined as follows:

$$T(V)(s) = \max_{a \in \mathscr{A}} \left(\bar{r}(s,a) + \gamma \sum_{s' \in \mathscr{S}} P_{s,s'}(a) V(s') \right), \quad \forall s \in \mathscr{S}.$$

Notice that for any vector $V \in \mathbb{R}^{|\mathscr{S}|}$, $T(V)$ corresponds to a stationary deterministic policy for which

$$T(V)(s) = \bar{r}(s, \pi(s)) + \gamma \sum_{s' \in \mathscr{S}} P_{s,s'}(\pi(s)) V(s') = T^{\pi}(V)(s),$$

where $\pi(s)$ denotes the deterministic action of policy π in state s, i.e.,

$$\pi(s) \in \arg\max_{a \in \mathscr{A}} \left(\bar{r}(s,a) + \gamma \sum_{s' \in \mathscr{S}} P_{s,s'}(a) V(s') \right).$$

Given the above equality, we say that the policy π is a *greedy policy* with respect to the vector V.

We want to show that T has a unique fixed point V^*, which by the above argument is the value function of some policy π^*. By the *contraction mapping theorem*, it is enough to show that T is a contraction operator.

Proposition 51. *The optimality operator T is a contraction in ℓ_∞ norm with factor γ:*

$$\|T(V) - T(V')\|_\infty \leq \gamma \|V - V'\|_\infty.$$

Proof: Fix $V, V' \in \mathbb{R}^{|\mathscr{S}|}$. Let π denote the greedy policy with respect to vector V. Consider a fixed $s \in \mathscr{S}$. We have:

$$
\begin{aligned}
T(V)(s) - T(V')(s) &= \bar{r}(s,\pi(s)) + \gamma \sum_{s' \in \mathscr{S}} P_{s,s'}(\pi(s)) V(s') - \max_{a \in \mathscr{A}} \left(\bar{r}(s,a) + \gamma \sum_{s' \in \mathscr{S}} P_{s,s'}(a) V'(s') \right) \\
&\leq \bar{r}(s,\pi(s)) + \gamma \sum_{s' \in \mathscr{S}} P_{s,s'}(\pi(s)) V(s') - \left(\bar{r}(s,\pi(s)) + \gamma \sum_{s' \in \mathscr{S}} P_{s,s'}(\pi(s)) V'(s') \right) \\
&= \gamma \sum_{s' \in \mathscr{S}} P_{s,s'}(\pi(s)) \big(V(s') - V'(s') \big) \\
&\leq \gamma \| V - V' \|_\infty .
\end{aligned}
$$

Notice that the right-hand side of the above inequality does not depend on $s \in \mathscr{S}$ nor the greedy policy π. Repeating the same argument for $T(V')(s) - T(V)(s)$, and then applying this bound for all $s \in \mathscr{S}$, the result follows. □

In conclusion, by the contraction mapping theorem, T has a unique fixed point V^*, which is the value function of a deterministic (greedy) stationary policy π^*:

$$
V^*(s) = T(V^*)(s) = \bar{r}(s,\pi^*(s)) + \gamma \sum_{s'} P_{s,s'}(\pi^*(s)) V^*(s') = T^{\pi^*}(V^*)(s).
$$

Next, we will show that π^* is the optimal stationary policy, i.e., $V^*(s) \geq V^\pi(s)$ for all $s \in \mathscr{S}$ and any stationary policy π. Let us first present two important properties of the Bellman optimality operator T.

Proposition 52. *The Bellman optimality operator T satisfies:*

(i) Monotonicity: If $V \leq V'$ (elementwise), then $T(V) \leq T(V')$.

(ii) Translation Invariance: For any $c \in \mathbb{R}$, $T(V + c \cdot \mathbf{1}) = T(V) + \gamma c \cdot \mathbf{1}$.

Proof: (i) For all s, since $V \leq V'$, we have for any a:

$$
\bar{r}(s,a) + \gamma \sum_{s'} P_{s,s'}(a) V(s') \leq \bar{r}(s,a) + \gamma \sum_{s'} P_{s,s'}(a) V'(s').
$$

Taking the max over a on both sides gives $T(V)(s) \leq T(V')(s)$.

(ii) For any s,

$$
\begin{aligned}
T(V + c\mathbf{1})(s) &= \max_a \left(\bar{r}(s,a) + \gamma \sum_{s'} P_{s,s'}(a) (V(s') + c) \right) \\
&= \max_a \left(\bar{r}(s,a) + \gamma \sum_{s'} P_{s,s'}(a) V(s') + \gamma c \sum_{s'} P_{s,s'}(a) \right) \\
&= T(V)(s) + \gamma c.
\end{aligned}
$$

since $\sum_{s'} P_{s,s'}(a) = 1$.

□

Fix a stationary policy π. Notice that by the definition of T and T^π, we have

$$T(V^\pi) \geq T^\pi(V^\pi) = V^\pi.$$

By the monotonicity of T, we have

$$T^2(V^\pi) \geq T(V^\pi) \geq V^\pi.$$

Repeating the same argument, we get

$$T^k(V^\pi) \geq V^\pi \quad \text{for all } k \in \mathbb{N}.$$

Taking the limit as $k \to \infty$ and invoking the contraction mapping theorem, we obtain

$$V^* = \lim_{k\to\infty} T^k(V^\pi) \geq V^\pi.$$

Hence, the policy π^* is the optimal policy among all stationary policies. It remains to show that π^* is the optimal policy among all Markov policies.

Theorem 53. *Given an infinite-horizon MDP $(\mathscr{S}, \mathscr{A}, P, r, \gamma)$, there exists a stationary policy π^* that maximizes the expected discounted total reward. Moreover, $V^* = T(V^*) = T^{\pi^*}(V^*)$.*

Proof: Let $\pi = \{\pi_k\}_{k\geq 0}$ be an optimal Markov policy. Define the time-dependent value function:

$$V^\pi_{k\to\infty}(s) := \mathbb{E}_\pi\left[\sum_{l=k}^{\infty} \gamma^{l-k} r(S_l, A_l) \,\middle|\, S_k = s\right].$$

Using Bellman equations, we have

$$V^\pi_{k\to\infty}(s) = \bar{r}(s, \pi_k(s)) + \gamma \sum_{s'} P^{\pi_k}_{s,s'} V^\pi_{k+1\to\infty}(s') \quad \forall k,$$

where

$$r(s, \pi_\ell(s)) := \sum_{a\in\mathscr{A}} \pi_\ell(a \mid s)\, r(s,a) \quad \text{and} \quad \bar{r}(s, \pi_\ell(s)) := \sum_{a\in\mathscr{A}} \pi_\ell(a \mid s)\, \bar{r}(s,a).$$

Next, we will show that

$$V^\pi_{k\to\infty} = T\left(V^\pi_{k+1\to\infty}\right) \quad \text{for all } k \geq 0.$$

Suppose the contrary, i.e., for some $k \geq 0$ and $s \in \mathscr{S}$, we have

$$V^\pi_{k\to\infty}(s) < T\left(V^\pi_{k+1\to\infty}\right)(s).$$

Let $\tilde{\pi}_k$ denote the greedy policy with respect to $T\left(V^\pi_{k+1\to\infty}\right)$, i.e.,

$$T\left(V^\pi_{k+1\to\infty}\right)(s) = \bar{r}(s, \tilde{\pi}_k(s)) + \gamma \sum_{s'\in\mathscr{S}} P^{\tilde{\pi}_k}_{s,s'} V^\pi_{k+1\to\infty}(s') = T^{\tilde{\pi}_k}\left(V^\pi_{k+1\to\infty}\right)(s).$$

Define the policy $\tilde{\pi}$ to be exactly the same as π, except at index k, for which π_k is replaced with $\tilde{\pi}_k$. By the above argument and simple induction, we have $V^\pi_{0\to\infty} \leq V^{\tilde{\pi}}_{0\to\infty}$ element-wise,

with inequality being strict for some $s \in \mathscr{S}$. This contradicts the assumption that π is optimal. Hence, for all $k \geq 0$, we have $V^{\pi}_{k\to\infty} = T\left(V^{\pi}_{k+1\to\infty}\right)$. Repetitive use of this identity yields

$$V^{\pi}_{0\to\infty} = T\left(V^{\pi}_{1\to\infty}\right) = T^2\left(V^{\pi}_{2\to\infty}\right) = \cdots = T^k\left(V^{\pi}_{k\to\infty}\right).$$

Since the random reward function is bounded by $r_{\max}$, it readily follows that

$$V^{\pi}_{k\to\infty} \leq \frac{r_{\max}}{1-\gamma} \cdot \mathbf{1},$$

where $\mathbf{1} \in \mathbb{R}^{|\mathscr{S}|}$ is the all-ones vector.

By Theorem 51 and using the above inequalities, for any $k > 0$ we have:

$$\begin{aligned}
\|V^{\pi}_{0\to\infty} - V^*\|_\infty &= \left\|T^k\left(V^{\pi}_{k\to\infty}\right) - V^*\right\|_\infty \\
&\leq \gamma^k \left\|V^{\pi}_{k\to\infty} - V^*\right\|_\infty \\
&\leq \frac{2\gamma^k r_{\max}}{1-\gamma}.
\end{aligned}$$

Taking the limit as $k \to \infty$, it follows that $V^{\pi}_{0\to\infty} = V^*$. By the same argument, we also have $V^{\pi}_{k\to\infty} = V^*$ for all $k \geq 0$. It remains to notice that V^* is the value function of some stationary policy π^*. □

Action-Value Function (Q-function)

So far, we have focused on the state-value function $V^{\pi} : \mathscr{S} \to \mathbb{R}_+$ for a policy π, and the optimal value function V^*. Another key object is the **action-value function** or **Q-function**:

Definition 54. *For a policy π, define the Q-function corresponding with π as*

$$Q^{\pi}(s,a) := \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k r(S_k, A_k) \,\middle|\, S_0 = s, A_0 = a\right], \quad \forall s \in \mathscr{S}, a \in \mathscr{A}.$$

That is, $Q^{\pi}(s,a)$ is the expected total reward when action a is taken at state s, and policy π is followed thereafter.

We note that by the above definition, for any $s \in \mathscr{S}$, we have

$$V^{\pi}(s) = \sum_{a \in .\mathscr{A}} \pi(a \mid s) Q^{\pi}(s,a).$$

Bellman Equation for Q-function. We can derive the Bellman equation for Q^{π} as follows:

$$\begin{aligned}
Q^{\pi}(s,a) &= \mathbb{E}[r(s,a)] + \mathbb{E}_{\pi}\left[\sum_{k=1}^{\infty} \gamma^k r(S_k, A_k) \,\middle|\, S_0 = s, A_0 = a\right] \\
&= \bar{r}(s,a) + \gamma \sum_{s',a'} \mathbb{P}(S_1 = s' \mid S_0 = s, A_0 = a) \cdot \pi(a' \mid s') \cdot Q^{\pi}(s',a').
\end{aligned}$$

Let the transition kernel under π be defined as:

$$P^{\pi,q}_{(s,a),(s',a')} := \mathbb{P}(S_1 = s' \mid S_0 = s, A_0 = a) \cdot \pi(a' \mid s').$$

Then the Bellman equation becomes:

$$Q^\pi = \bar{r} + \gamma P^{\pi,q} Q^\pi,$$

where $Q^\pi \in \mathbb{R}^{|\mathscr{S}||\mathscr{A}|}$, $\bar{r}$ is the reward vector over (s,a) pairs, and $P^{\pi,q} \in \mathbb{R}_+^{|\mathscr{S}||\mathscr{A}|\times|\mathscr{S}||\mathscr{A}|}$ is the transition matrix under π.

Bellman Operator for Q. Define the operator $T^{\pi,q} : \mathbb{R}^{|\mathscr{S}||\mathscr{A}|} \to \mathbb{R}^{|\mathscr{S}||\mathscr{A}|}$:

$$T^{\pi,q}(Q) = \bar{r} + \gamma P^{\pi,q} Q.$$

Again, one can show that Q^π is the unique fixed point of $T^{\pi,q}$, and this operator is also a γ-contraction.

Bellman Optimality Operator for Q. Define the Bellman optimality operator $T^q : \mathbb{R}^{|\mathscr{S}||\mathscr{A}|} \to \mathbb{R}^{|\mathscr{S}||\mathscr{A}|}$:

$$T^q(Q)(s,a) = \bar{r}(s,a) + \gamma \sum_{s'} P_{s,s'}(a) \max_{a'} Q(s',a'), \quad \forall s \in \mathscr{S}, a \in \mathscr{A}.$$

Then,

- T^q is a γ-contraction in ℓ_∞ norm.
- The unique fixed point Q^* of T^q satisfies:

$$Q^*(s,a) = \bar{r}(s,a) + \gamma \sum_{s'} P_{s,s'}(a) \max_{a'} Q^*(s',a').$$

- The optimal deterministic stationary policy π^* satisfies:

$$\pi^*(a \mid s) = \begin{cases} 1, & \text{if } a = \arg\max_{a'} Q^*(s,a'), \\ 0, & \text{otherwise.} \end{cases}$$

- The optimal value function is:

$$V^*(s) = \sum_{a \in \mathscr{A}} \pi^*(a|s) Q^*(s,a) = \max_{a \in \mathscr{A}} Q^*(s,a).$$