9.1 Introduction

A Markov decision process is a discrete time stochastic control process. It provides a mathematical framework for modeling decision making in situations where outcomes are partly random and partly under the control of a decision maker. In control theory, "controlled" means the action can be controlled. However, a Markov chain is controlled if the state transition probabilities ($P_{ij}(u)$: $u$ is the control action or decision) can be controlled. We will assume that $u \in U$ and $U$ is a finite set. When the action $u$ is taken, a cost $C(x,u)$ (can be deterministic or random) is incurred. For the cost $C(x,u)$, we assume that cost $C(x,u) \geq 0$ (without loss of generality), which means if there is a sequential cost $1, 0, -1$, we add 1 to all the cost, then all the cost become greater or equal to 0, but the generality remains.

To solve Markov Decision Problem, our goal is for a given initial state: $x_0 = i$, and discount factor: $\alpha \in [0, 1)$, choosing a sequence of action $\{u_k\}$ to minimize cost:

$$J_\mu(i) = \lim_{N \to \infty} \mathbb{E}\left[\sum_{k=0}^{N} \alpha^k C(x_k, u_k) | x_0 = i\right] \quad \alpha \in [0, 1)$$

Since in above problem time is from 0 to $\infty$ and also have a discount factor $\alpha$. The above problem is called "infinite-horizon discounted cost Markov Decision Problem".

To solve the problem, we first define History $\tilde{F}$ at time $k$ as a sequential states and actions up to time $k - 1$, with $x_k$ (state at time k) that we can observe: $\tilde{F} = \{x_0, u_0, x_1, u_1, \cdots, x_{k-1}, u_{k-1}, x_k\}$ where x is state and u is action. Then we define the policy as $\mu_k(\cdot)$. $\mu_k(\cdot)$ can be deterministic or randomized policy. Therefore, the action at time k is $u_k = \mu_k(\tilde{F}_k)$. Actually, it is sufficient to restrict attention to policies of the form: $u_k = \mu(x_k)$, which means we do not need all the history, only need current state to get optimal action.

9.2 Stationary policy

**Definition 9.2.1.** If policy $\mu_k(\cdot)$ does not depend on time $k$, then $u_k = \mu(x_k)$, and $\mu$ is called "stationary policy".

The above definition means the policy does not vary with the time. Generally speaking, infinite horizon problem has optimal policy with such stationary. Finite horizon problem has policy vary with time to the terminal time. And we will focus on infinite horizon problem. For a non-stationary policy $\mu = (\mu_0, \mu_1, \cdots)$. Therefore, for a stationary policy $\mu = (\mu, \mu, \cdots)$. The "." means this is a vector.

**Lemma 9.2.2.** For a stationary policy, with discount factor $\alpha \in [0, 1)$, $J_\mu(i) = \lim_{N \to \infty} \mathbb{E}\left[\sum_{k=0}^{N} \alpha^k C(x_k, u_k) | x_0 = i\right]$ is always upper bounded:

$$J_\mu(i) \leq \max_{i,u} \frac{C(i,u)}{1 - \alpha}$$
Proof.

\[ J_\mu(i) = \lim_{N \to \infty} \mathbb{E}\left[ \sum_{k=0}^{N} \alpha^k C(x_k, u_k) | x_0 = i \right] \]

because of stationary policy and assume \( x_k = j \):

\[ J_\mu(i) \leq \mathbb{E}\left[ \sum_{k=0}^{\infty} \alpha^k \max_{j,u} C(j, u) | x_0 = i \right] \]

\( \max_{j,u} C(j, u) \) is constant, therefore:

\[ J_\mu(i) \leq \max_{j,u} C(j, u) \sum_{k=0}^{\infty} \alpha_k \]

by the formula of summation for geometric sequence:

\[ J_\mu(i) \leq \max_{j,u} C(j, u) \left( \frac{1 - \alpha^k}{1 - \alpha} \right) \]

because \( \alpha \in [0, 1) \), \( \alpha^k \to 0 \), as \( k \to \infty \). Thus:

\[ J_\mu(i) \leq \max_{j,u} C(j, u) \frac{1}{1 - \alpha} \]

It is intuitive that as \( N \to \infty \), the optimal policy \( \mu^*_0 \) at time 0 loses its dependence on time N. Thus, \( \mu^*_0 \) is optimal for subsequent time instances. Therefore, for the problem considered here the optimal policy \( \mu^* \) is Markovian and Stationary. i.e., \( \mu^* = (\mu^*, \mu^*, \cdots) \)

### 9.3 Performance of a Stationary Deterministic Policy

For a given stationary deterministic policy, we care about the performance of this policy. Again, let \( J_\mu(i) = \lim_{N \to \infty} \mathbb{E}\left[ \sum_{k=0}^{N} \alpha^k C(x_k, u_k) | x_0 = i \right] \). Solving a MDP means we look for whole horizon and solve the optimal policy we need to apply every time instance for whatever the initial state given, so we need \( J_\mu(i) \) to be well-defined for any initial state. Since \( \sum_{k=0}^{N} \alpha^k C(x_k, u_k) \) is monotonically increasing, and \( J_\mu(i) \) upper bounded by \( \max_{i,u} \frac{C(i, u)}{1 - \alpha} \), then the limit of \( J_\mu(i) \) exists. Therefore, \( J_\mu(i) \) is well-defined for any initial state \( i \).

#### 9.3.1 Unique solution of cost \( J_\mu \)

**Definition 9.3.1.** There exists a unique solution for cost \( J_\mu \):

\[ J_\mu = (I - \alpha P_\mu)^{-1} C_\mu \]

**Proof.**

**Lemma 9.3.2.** For cost at state \( i \): \( J_\mu(i) = C(i, \mu(i)) + \alpha \sum_j P_{ij}(\mu(i)) J_\mu(j), \quad \forall i \).
Proof.

\[ J_\mu(i) = \lim_{N \to \infty} \mathbb{E} \left[ \sum_{k=0}^{N} \alpha^k C(x_k, u_k) | x_0 = i \right] \]

by \( u_k = \mu(x_k) \):

\[ = \lim_{N \to \infty} \mathbb{E} \left[ \sum_{k=0}^{N} \alpha^k C(x_k, \mu(x_k)) | x_0 = i \right] \]

split case for initial state:

\[ = \lim_{N \to \infty} \mathbb{E} \left[ C(i, \mu(i)) + \sum_{k=1}^{N} \alpha^k C(x_k, \mu(x_k)) | x_0 = i \right] \]

assume \( C(i, \mu(i)) \) deterministic, then \( C(i, \mu(i)) \) is a constant:

\[ \begin{align*}
&= \lim_{N \to \infty} C(i, \mu(i)) + \mathbb{E} \left[ \sum_{k=1}^{\infty} \alpha^k C(x_k, \mu(x_k)) | x_0 = i \right] \\
&= C(i, \mu(i)) + \alpha \mathbb{E} \left[ \sum_{k=1}^{\infty} \alpha^{k-1} C(x_k, \mu(x_k)) | x_0 = i \right] \\
\end{align*} \]

assume \( x_1 = j \):

\[ \begin{align*}
&= C(i, \mu(i)) + \alpha \sum_{j} P_{ij}(\mu(i)) \mathbb{E} \left[ \sum_{k=0}^{\infty} \alpha^k C(x_k, \mu(x_k)) | x_0 = j \right] \\
&= C(i, \mu(i)) + \alpha \sum_{j} P_{ij}(\mu(i)) J_\mu(j) \\
\end{align*} \]

lemma has been proved. \( \square \)

According to lemma 9.3.2, and let \( x = \{1, 2, \cdots, x\} \), we can define the vector form of \( J_\mu(i), C_\mu(i) \) and \( P_\mu \) as following:

\[
J_\mu = \begin{bmatrix} J_\mu(1) \\ J_\mu(2) \\ \vdots \\ J_\mu(x) \end{bmatrix}, \quad C_\mu = \begin{bmatrix} C(1, \mu(1)) \\ C(2, \mu(2)) \\ \vdots \\ C(x, \mu(x)) \end{bmatrix}, \quad P_\mu = [P_{ij}(\mu(i))] 
\]

Therefore:

\[ J_\mu = C_\mu + \alpha \alpha P_\mu J_\mu \iff (I - \alpha P_\mu) C_\mu = C_\mu \iff J_\mu = (I - \alpha P_\mu)^{-1} C_\mu \]

\[ \begin{align*}
\text{remark 9.3.3.} & \text{ The unique solution of } J_\mu \text{ is only exists, if } I - \alpha P_\mu \text{ is non-singular } \iff I - \alpha P_\mu \text{ does not have eigenvalue equal to } 0 \iff P_\mu \text{ stochastic matrix has eigenvalue } |\lambda_i(P_\mu)| \leq 1, \forall i \text{ (since } \alpha < 1). 
\end{align*} \]
Proof. Let B and A both matrix, and $B = A - aI$. The eigenvector of A is u: $Au = \lambda u$. Then:

$$Bu = (A - aI)u = Au - au = \lambda u - au = (\lambda(A) - a)u = \lambda(B)$$

Therefore, $\lambda(B) = \lambda(A) - a$, apply this to $I - \alpha P$, we can see if $|\lambda_i(P_\mu)| \leq 1$, the eigenvalue of $I - \alpha P_\mu$ will always larger than 0, then we can get unique solution for $J_\mu$.

Therefore, according to remark 9.3.3, in order to show $I - \alpha P_\mu$ is non-singular, we only need to show $|\lambda_i(P_\mu)| \leq 1$, $\forall i$.

**Lemma 9.3.4.** For a stochastic matrix A $(P_\mu)$, all the eigenvalues lie in [-1,1].

Proof. In order to prove Lemma 9.3.4, we first need to introduce some definition of norms.

### 9.3.1.1 Norms

**Definition 9.3.5.** Let $y \in \mathbb{R}^n$, a norm $|| \cdot ||$ is a function from $\mathbb{R}^n$ to $\mathbb{R}^+$ satisfying:

1. $||y|| = 0 \iff y = 0$
2. $||ay|| = ||a|| \cdot ||y||$, $\forall a \in \mathbb{R}$
3. $||x + y|| \leq ||x|| + ||y||$, $\forall x, y \in \mathbb{R}$ (triangle inequality)

**Definition 9.3.6.** Class of P-norms:

$$||y||_P = (|y_1|^P + \cdots + |y_n|^P)^{1/P} = P\sqrt[1/P]{|y_1|^P + \cdots + |y_n|^P}, \quad 1 \leq P \leq \infty$$

**Definition 9.3.7.** The $\infty$-norm of y is defined as:

$$||y||_\infty = \max_i |y_i|$$

**Definition 9.3.8.** Treating $A \in \mathbb{R}^{n \times n}$ as vector in $\mathbb{R}^{n^2}$ (normal way is take column and stack), and combine with Euclidean norm, one can define the Frobenius form: $||A||_F = \sqrt{\sum_{i,j} A_{ij}^2}$

**Definition 9.3.9.** By treating $A \in \mathbb{R}^{n \times n}$ as an operator. The operator $P$-norm (induce $P$-norm) is defined to be:

$$||A||_P = \sup_{y \neq 0} \frac{||Ay||_P}{||y||_P} = \max_{||y||_P = 1} ||Ay||_P, \quad 1 \leq P \leq \infty$$

### 9.3.1.2 Proof 1

Proof. There is a well-known result in linear algebra states that:

$$\max_i |\lambda_i(A)| \leq ||A||_P, \forall P \geq 1$$

Thus, using $P = \infty$, according to definition 9.3.9:

$$||A||_P = \max_{||y||_P = 1} ||Ay||_P$$
Let:
\[ V = Ay = \begin{bmatrix}
\sum_j A_{1j}y_j \\
\sum_j A_{2j}y_j \\
\vdots \\
\sum_j A_{nj}y_j
\end{bmatrix} \]

Then, according to definition 9.3.7:
\[ ||V||_\infty = \max_i |V_i| = \max_i |\sum_j A_{ij}y_j| \]

Thus:
\[ ||A||_\infty = \max_{||y||_\infty = 1} \max_i |\sum_j A_{ij}y_j| \]

In order to maximize the \(|\sum_j A_{ij}y_j|\), When \(A_{ij} > 0\), \(y_j = 1\). When \(A_{ij} < 0\), \(y_j = -1\). Therefore:
\[ ||A||_\infty = \max_i |\sum_j A_{ij}| = \text{maximum row sum of } A. \]

Apply above result to \(P\mu\) for \(P = \infty\):
\[ |\lambda_i(P\mu)| \leq ||P\mu||_\infty = 1, \forall i \]

Thus, we can prove that \(J\mu\) has unique solution:
\[ J\mu = (I - \alpha P\mu)^{-1}C\mu. \]

### 9.3.1.3 Proof 2: Contraction Mapping Theorem

**Theorem 9.3.10.** Contraction Mapping Theorem: Let \(T\) be a mapping from \(\mathbb{R}^n\) to \(\mathbb{R}^n\). Assume that \(T\) satisfy a "contraction property", i.e, \(\exists a \in [0, 1)\) such that
\[ ||T(x) - T(y)|| \leq a||x - y||, \forall x, y \in \mathbb{R}^n \]
where ||·|| can be any norm (e.g ||·||_P for \(1 \leq P \leq \infty\)).

Then:

(a) \(\exists a\) unique fixed point \(x^*\), i.e, \(x^* = T(x^*)\)

(b) Starting with any \(x_0 \in \mathbb{R}^n\), the recursion (used for policy derivation)
\[ x_{k+1} = T(x_k) \quad k \to \infty \to x^* \]

Based on theorem 9.3.10, we can start the proof.

**Proof.**
\[ J\mu = \overline{C}\mu + \alpha P\mu J\mu \quad (1) \]

Set:
\[ T\mu(J\mu) = \overline{C}\mu + \alpha P\mu J\mu \]
Let (1) as a fixed point equation:

\[ J_\mu = T_\mu(J_\mu) \]

we can observe that:

\[ T_\mu(x) - T_\mu(y) = aP_\mu(x - y) \iff \|T_\mu(x) - T_\mu(y)\|_\infty = \max_i |T_\mu(x) - T_\mu(y)| \]

\[ = a \max_i |P_\mu(x - y)| \]

\[ = a \max_i \left| \sum_j P_{ij}(x_j - y_j) \right|_i \]

\[ \leq a \max_i \left( \sum_j P_{ij}|x_j - y_j| \right)_i \]

\[ \leq a \|x - y\|_\infty \]

Therefore, \( T_\mu \) is a contraction mapping, \( J_\mu = T_\mu(J_\mu) \) has a unique solution.

Another way is iteratively solve for \( J_\mu \) by the iteration:

\[ J^{(k+1)} = T_\mu(J^{(k)}) \text{ for any arbitrary } J^{(0)} \]

Because as \( k \to \infty \), \( J^{(k)} \to J_\mu \).