Finite-states Markov chains

Yinan Huang

Oct 15, 2020

Contents

1 Introduction

Finites-states Markov chain is a discrete stochastic process $\{X_n | n = 0, 1, 2, ...\}$ defined by

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i', ..., X_0 = i'') = P(X_{n+1} = j | X_n = i). \tag{1}$$

So a discrete stochastic process is a finite-states Markov chain when it has finite states to stay at and probability staying at state j only depends on the last state it stayed at. To characterize a unique Markov chain, what we need is to specify its transition matrix defined by

$$P(X_{n+1} = j | X_n = i) = P_{i,j}. \tag{2}$$

Note that $\sum_j P_{i,j} = 1$. For matrices whose rows sum up to 1, they are called stochastic matrices. Starting from $X_0 = i$ and take n steps, we get the n-steps transition matrix

$$P(X_n = j | X_0 = i) = \sum_{k_1} P(X_n = j | X_{n-1} = k_1, X_0 = i) P(X_{n-1} = k_1, X_0 =) \tag{3}$$

2 Terminology

Some terminology will be useful for the latter discussion.

Definition 2.1. *State i and j are called to **communicate** with each other if i is reachable from j and j is also reachable from i.*

We can see that "communicate" is an equivalent relation. Thus using this relation we can partition all states into disjoint classes. Only states within the same class can move to each other, and there is no way to move from states in one class to states in another class and move back: only one-direction path exists between classes.

Definition 2.2. *A state i is called **recurrent**, if for any state j, state j is reachable to i implies state i is reachable to state j. In other words, for a recurrent state i, for any state j, if $P^n_{j,i} > 0$ for some n, then there exist some m such that $P^m_{i,j} > 0$. If a state is not recurrent, then it is called transient.*

Theorem 2.1. *For a class $\mathbb{C}$, either all its states are recurrent or transient. Thus we can categorize classes into **recurrent class** and **transient class**.*

Proof 2.1. *Suppose $i \in \mathbb{C}$ is recurrent. Consider state $j \in \mathbb{C}$ and any state m. If $m \in \mathbb{C}$, then we have $j \to m$ and $m \to j$. If $m \notin \mathbb{C}$, then there is no way to go to m from j. Since if not, i can also go to m via j, and i is recurrent implies $m \to i$ as well. Therefore $m \in \mathbb{C}$, which is a contradiction. So for state $j \in \mathbb{C}$, it cannot go outside the class if $i \in \mathbb{C}$ is recurrent. This shows j is also recurrent. So all states in $\mathbb{C}$ is recurrent. Consequently, if one state in $\mathbb{C}$ is not recurrent, namely transient, then all states in $\mathbb{C}$ will be transient.*

Definition 2.3. *The **period** of a state i is defined as the greatest common divisor of n such that $P^n_{i,i} > 0$, namely*

$$d(i) \equiv gcd\{n \in \mathbb{N} | P^n_{i,i} > 0\}. \tag{4}$$

If $d(i) = 1$, then we say state i is aperiodic.

Theorem 2.2. *Class $\mathbb{C}$ has the same period. So we say a class $\mathbb{C}$ with period $d(\mathbb{C})$.*

Proof 2.2. *Let $i, j \in \mathbb{C}$ and their period is $d(i)$ and $d(j)$ respectively. Suppose $P^r_{i,j} > 0$ and $P^s_{j,i} > 0$ for some r, s, which is possible since $i \leftrightarrow j$. Then $r + s$ steps can goes from i to j and back to i, which suggests $r + s$ is divisible by $d(i)$. Suppose $P^t_{j,j} > 0$, then $r + s + t$ steps can goes from i to j to j and back to i, which suggests t is also divisible by $d(i)$. Note that t is random, so let $t = d(j)$ and we conclude that $d(j)$ is divisible by $d(i)$. Since roles of i and j are symmetric, so the similiar argument gives $d(i)$ is divisible by $d(j)$. In other words, $d(i) = d(j)$.*

Definition 2.4. *A state is called* ***ergodic****, if it is recurrent and aperiod. A Markov chain consists of only one ergodic class is called* ***ergodic chain****. A Markov chain consists of only one ergodic class and some other transient classes is called* ***ergodic unichain****.*

We are going to show an important property of ergodic chain. It is related with the steady state latter.

Theorem 2.3. *For an ergodic chain with M states, for each state i and j, if $n \geq (M-1)^2 + 1$, then going from i to j in n steps is always possible:*

$$P^n_{i,j} > 0. \tag{5}$$

Note that the theorem says for an ergodic chain, as long as we take enough large steps, we can goes to everywhere we want no matter what our starting point is. Be careful that it must be a ergodic chain, which means unique ergodic class is neccessary to ensure the theorem.

3 Steady states

We are interested in the long-term behavior of a Markov chain. Especially, we are looking forward to find a steady state vector.

Definition 3.1. *A* ***steady distribution*** *or* ***steady-state vector*** $\boldsymbol{\pi}$ *is the distribution that does not change after one step transition. Concretely, if*

$$\boldsymbol{\pi} = \boldsymbol{\pi} P, \quad \sum_i \pi_i = 1. \tag{6}$$

Then $\boldsymbol{\pi}$ is called the steady vector.

Our discussion will focus on the following questions:

- Under what conditions will $\boldsymbol{\pi}$ exist?
- Under what conditions will $\boldsymbol{\pi}$ exist and be unique?
- Under what conditions will the final vector from an arbitrary starting vector converge to $\boldsymbol{\pi}$?

Before proofs, we first give the answers. First, steady vector $\boldsymbol{\pi}$ always exist, and the number of steady vectors is equal to the number of recurrent states (Markov chain must have recurrent states). So steady vector is unique if and only if chain only contains one recurrent class, namely a recurrent unichain. Finally, any arbitray vector will converge to the unique steady vector, if and only if the chain is an ergodic unichain.

3.1 Markov Chain with $P_{i,j} > 0$

We first assume $P_{i,j} > 0$ for any i, j. Though this assumption seems making no sense, but we have seen that for an ergodic chain, $P^n_{i,j} > 0$ if n is larger. So it is useful to discuss a positive transition matrix.

Theorem 3.1. *Let P be the transition matrix of a finite-state Markov chain, and $P_{i,j} > 0$ for any i, j. Let $\alpha = \min_{i,j} P_{i,j}$, then for all states i, j,*

- $\max_i P^{n+1}_{i,j} - \min_i P^{n+1}_{i,j} \leq (\max_l P^n_{l,j} - \min_l P^n_{l,j})(1 - 2\alpha)$.
- $\max_l P^n_{l,j} - \min_l P^n_{l,j} \leq (1 - 2\alpha)^n$.
- $\lim_{n\to\infty} \max_i P^n_{i,j} = \lim_{n\to\infty} \min_i P^n_{i,j} > 0$.

Proof 3.1. *(1) To prove the first inequality, let us consider*

$$\begin{aligned} P^{n+1}_{i,j} &= \sum_k P_{i,k}P^n_{k,j} = P_{i,l_{\min}} \min_l P^n_{l,j} + \sum_{k\neq l_{\min}} P_{i,k}P^n_{k,j} \leq P_{i,l_{\min}} \min_l P^n_{l,j} + \sum_{k\neq l_{\min}} P_{i,k} \max_l P^n_{l,j} \\ &= P_{i,l_{\min}} \min_l P^n_{l,j} + \max_l P^n_{l,j} - P_{i,l_{\min}} \max_l P^n_{l,j} \leq \alpha \left(\min_l P^n_{l,j} - \max_l P^n_{l,j}\right) + \max_l P^n_{l,j} \end{aligned} \tag{7}$$

Let $i = \arg\max_l P^{n+1}_{l,j}$ and we will find

$$\max_i P^{n+1}_{i,j} \leq \alpha \left(\min_l P^n_{l,j} - \max_l P^n_{l,j}\right) + \max_l P^n_{l,j}. \tag{8}$$

Reverse the roles of min *and* max *we also have*

$$\begin{aligned} P^{n+1}_{i,j} &= \sum_k P_{i,k}P^n_{k,j} = P_{i,l_{\max}} \max_l P^n_{l,j} + \sum_{k\neq l_{\max}} P_{i,k}P^n_{k,j} \geq P_{i,l_{\max}} \max_l P^n_{l,j} + \sum_{k\neq l_{\max}} P_{i,k} \min_l P^n_{l,j} \\ &= P_{i,l_{\max}} \max_l P^n_{l,j} + \min_l P^n_{l,j} - P_{i,l_{\max}} \min_l P^n_{l,j} \geq \alpha \left(\max_l P^n_{l,j} - \min_l P^n_{l,j}\right) + \min_l P^n_{l,j} \end{aligned} \tag{9}$$

Let $i = \arg\min_l P^n_{l,j}$, then we will find

$$\min_l P^{n+1}_{l,j} \geq \alpha(\max_l P^n_{l,j} - \min_l P^n_{l,j}) + \min_l P^n_{l,j}. \tag{10}$$

Thus

$$\max_i P^{n+1}_{i,j} - \min_i P^{n+1}_{i,j} \leq (1 - 2\alpha) \left(\max_l P^n_{l,j} - \min_l P^n_{l,j}\right). \tag{11}$$

(2) Note that $\min_l P_{l,j} \geq \alpha$ and $\max_l P_{l,j} \leq 1 - \min_j \max_l P_{l,j} \leq 1 - \alpha$, so by applying conclusion in (1) we find

$$\max_l P^n_{l,j} - \min_l P^n_{l,j} \leq (1 - 2\alpha)^n. \tag{12}$$

(3) Note that $0 \leq \alpha \leq 1/2$, thus $0 \leq 1 - 2\alpha \leq 1$, thus $(1 - 2\alpha)^n \to 0$ as $n \to \infty$. Thus $\lim_{n\to\infty} \max_l P^n_{l,j} = \lim_{n\to\infty} \min_l P^n_{l,j}$*, which suggests* $\lim_{n\to\infty} P_{i,j} = \pi_j$ *is a constant of i.*

3.2 Ergodic Markov Chain

Let us consider ergodic chain, which says $P^n_{i,j} > 0$ if $n \geq (M - 1)^2 + 1$. By defining $h \equiv (M-1)^2+1$, we can see P^h as the positive matrix we discuss previously and apply the conclusion in the last subsection.

Theorem 3.2. *Let P be the transition matrix of a finite-state Markov chain. Then there is a unique steady vector $\boldsymbol{\pi}$ such that any starting vector converges to $\boldsymbol{\pi}$, namely $\lim_{n\to\infty} P^n = e\boldsymbol{\pi}$, where $e \equiv (1,1,1,...)^T$.*

Proof 3.2. *Applying the conclusion in $[P] > 0$, we find*

$$\max_l P^n_{l,j} - \min_l P^n_{l,j} \leq (1-\beta)^{\lfloor \frac{n}{h} \rfloor}, \tag{13}$$

where $\beta \equiv \min_{i,j} P^h_{i,j}$. As $n \to \infty$,

$$\lim_{n\to\infty} \max_i P^n_{i,j} = \lim_{n\to\infty} \min_i P^n_{i,j} \equiv \pi_j. \tag{14}$$

Suppose $\boldsymbol{\mu}$ is the steady state such that $\boldsymbol{\mu} = \boldsymbol{\mu}P$, then it is easy to prove that $\boldsymbol{\pi}$ is just the steady state:

$$\boldsymbol{\mu} = \boldsymbol{\mu}P = \boldsymbol{\mu} \lim_{n\to\infty} P^n = \boldsymbol{\mu}e\boldsymbol{\pi} = \boldsymbol{\pi}. \tag{15}$$

The theorem says if P is an ergodic chain, then the steady state $\boldsymbol{\pi}$ is unique and P exponentially converges to $e\boldsymbol{\pi}$.

3.3 Ergodic Markov Unichain

Unichain means except an ergodic class, there are transient states as well. We will see that ergodic unichain P^n can also converge to $e\boldsymbol{\pi}$. We first note that there is no path going from ergodic class to transient states, so the transition matrix looks like

$$P = \left(\begin{array}{c|c} P_T & P_{TR} \\ \hline 0 & P_R \end{array}\right), \quad P^n = \left(\begin{array}{c|c} P^n_T & ... \\ \hline 0 & P^n_R \end{array}\right) \tag{16}$$

We have already know that $\lim_{n\to\infty} P^n_R = e\boldsymbol{\pi}$. So we will focus on P^n_T and the off-diagonal elements.

Let T be set of transient states and R be set of recurrent states. Let t the number of transient states, then take at most t steps will let us go to R from T, which means for any $i \in T$

$$\sum_{j\in R} P^t_{i,j} > 0. \tag{17}$$

Let

$$\gamma \equiv \max_{i\in T} \sum_{j\in T} P_{i,j} < 1. \tag{18}$$

Theorem 3.3. *Let P be the ergodic unichain, T be the set of transient states and t be the number of transient states. Then*

$$\max_{i\in T} \sum_{j\in T} P^n_{i,j} \leq \gamma^{\lfloor \frac{n}{t} \rfloor}. \tag{19}$$

Proof 3.3. *For each $i \in T$ and $v \in \mathbb{N}$,*

$$\sum_{j\in T} P^{(v+1)t}_{i,j} = \sum_{j\in T}\sum_{k\in T} P^t_{i,k} P^{vt}_{k,j} = \sum_{k\in T} P^t_{i,k} \sum_{j\in T} P^{vt}_{k,j} \leq \gamma \max_{i\in T} \sum_{j\in T} P^{vt}_{i,j}. \tag{20}$$

Let $i = \arg\max_{i\in T} \sum_{j\in T} P^{(v+1)t}_{i,j}$, then

$$\max_{i\in T} \sum_{j\in T} P^{(v+1)t}_{i,j} \leq \gamma \max_{i\in T} \sum_{j\in T} P^{vt}_{i,j}. \tag{21}$$

Note that $\max_{i\in T}\sum_{j\in T} P^t_{i,j} \leq \gamma$, *so*

$$\max_{i\in T}\sum_{j\in T} P^{vt}_{i,j} \leq \gamma^t, \tag{22}$$

which means

$$\max_{i\in T}\sum_{j\in T} P^{n}_{i,j} \leq \gamma^{\lfloor \frac{n}{t} \rfloor}. \tag{23}$$

The theorem suggests for $i, j \in T$, $\lim_{n\to\infty} P^n_{i,j} = 0$. Thus $P^n_T \to 0$ as $n \to \infty$.

Now we need to prove the off-diagonal part also converge to $\boldsymbol{e\pi}$. Let $i \in T$ and $j \in R$, then

$$\begin{aligned}
\left|P^n_{i,j} - \pi_j\right| &= \left|\sum_{k\in T} P^m_{i,k}P^{n-m}_{k,j} + \sum_{k\in R} P^m_{i,k}P^{n-m}_{k,j} - \sum_{k\in T} P^m_{i,k}\pi_j - \sum_{k\in R} P^m_{i,k}\pi_j\right| \\
&= \left|\sum_{k\in T} P^m_{i,k}\left(P^{n-m}_{k,j} - \pi_j\right) + \sum_{k\in R} P^m_{i,k}\left(P^{n-m}_{k,j} - \pi_j\right)\right| \\
&\leq \sum_{k\in T} P^m_{i,k}\left|P^{n-m}_{k,j} - \pi_j\right| + \sum_{k\in R} P^m_{i,k}\left|P^{n-m}_{k,j} - \pi_j\right| \\
&\leq \sum_{k\in T} P^m_{i,k} + \sum_{k\in R} P^m_{i,k}\left|P^{n-m}_{k,j} - \pi_j\right| \leq \gamma^{\lfloor m/t \rfloor} + (1-2\beta)^{\lfloor (n-m)/h \rfloor}
\end{aligned} \tag{24}$$

Let $m = n-1$, then we can see that $\lim_{n\to\infty} P^n_{i,j} = \pi_j$. In summary, we prove that P^n_T decays to 0 and the off-diagonal term also converge to $\boldsymbol{e\pi}$, where $\pi_j \equiv \lim_{n\to\infty} P^n_{i,j}$ with $i, j \in R$. Thus $\lim_{n\to\infty} P^n_{i,j} = \boldsymbol{e\pi}$ for any i, j.

Finally, we note that if there is multiple ergodic classes, for example say 3, then

$$P_R = \left(\begin{array}{c|c|c} P_{R_1} & 0 & 0 \\ \hline 0 & P_{R_2} & 0 \\ \hline 0 & 0 & P_{R_3} \end{array}\right) \tag{25}$$

So we can treat different ergodic classes seperately. It leads to multiple steady vectors, and thus P will not converge to a unique steady vector. The long-term final state depends on the initial vector. We will use eigenstates analysis to further discuss these conclusions in the next section.

4 Eigenvectors Analysis

We are going to use the method of eigenvectors to draw the similar conclusions we get in the last section and see carefully why P^n will converge to $\boldsymbol{e\pi}$ under an ergodic unichain condition.

Definition 4.1. *For transition matrix P, vector $\boldsymbol{v_i}$ such that $\boldsymbol{v_i}P = \lambda_i\boldsymbol{v_i}$ is called the left eigenvector of P; vector $\boldsymbol{u_i}$ such that $P\boldsymbol{u_i} = \lambda_i\boldsymbol{u}_i$ is called the right eigenvector of P.*

Eigenvectors analysis requires an assumption that the eigenvectors of P span the whole vector space. This is equivalent to say we can find M linearly independent eigenvectors. If so, we can diagonalize P by a similar transform U, whose columns are the right eigenvectors:

$$P = U^{-1}\Lambda U, \tag{26}$$

Here Λ is the diagonal matrix with $\Lambda_{i,j} = \delta_{i,j}\lambda_i$. Also we can diagonalize P by a similar transform V related with V, whose rows are the left eigenvectors:

$$P = V\Lambda V^{-1}. \tag{27}$$

So we find $U = V^{-1}$. Then we can write

$$P = V\Lambda U. \tag{28}$$

Note that $Pe = e$, where $e = (1, 1, 1, ...)^T$. Thus a stochastic matrix P always has a eigenvalue $\lambda = 1$. This suggests the existence of steady state $\boldsymbol{\pi} P = \boldsymbol{\pi}$ for any stochastic matrix P. For P^n, we have

$$P^n = V\Lambda^n U. \tag{29}$$

Or in a vector product form we can write

$$P^n = \sum_i \lambda_i^n \boldsymbol{v_i u_i}. \tag{30}$$

Now we can let $n \to \infty$. Then for those $|\lambda_i| < 1$, term $\lambda_i^n \boldsymbol{v_i u_i}$ will exponentially decay to 0. So

$$\lim_{n\to\infty} P^n = \sum_{\{i:|\lambda_i|=1\}} \lambda_i^n \boldsymbol{v_i u_i}. \tag{31}$$

We claim that the number of eigenvectors with $\lambda_i = 1$ is equal to the number of recurrent states, and the number of eigenvectors with $|\lambda_i| = 1$ is equal to the period related with that recurrent class, and $\lambda_n = \text{exp}(i2n\pi/d)$.

Suppose Markov chain is an ergodic unichain, then there is only one unique eigenvalue $\lambda = 1$, related with right eigenvector e and left eigenvector $\boldsymbol{\pi}$. So

$$\lim_{n\to\infty} P^n = e\boldsymbol{\pi}, \tag{32}$$

which implies P^n converges to steady state $\boldsymbol{\pi}$. If the recurrent class is not aperiod (thus not ergodic), then we have multiple eigenvalues $|\lambda_i| = 1$, then P^n will not converge to the unique steady state. If there are multiple ergodic states, then we have multiplie eigenvalues $\lambda_i = 1$, namely multiple steady states.

5 Markov chain with rewards

In this section we are going to consider a Markov chain with rewards. It is a model that assigns rewards to corresponding states, and once we leaves that state, we will obtain the corresponding reward.

5.1 Expected aggregate rewards

To formalize the problem, we define a reward function

$$R(i) \equiv r_i, \tag{33}$$

for all states i. Then we define our aggregate reward as a random variable depending on steps n:

$$R_n \equiv \sum_{i=0}^{n-1} R(X_i). \tag{34}$$

We are interested in the expected aggregate reward given our initial state, namely

$$v_i(n) \equiv E(R_n|X_0 = i) = \sum_{j=0}^{n-1} E(R(X_j)|X_0 = i) = \sum_{j=0}^{n-1}\sum_k r_k P(X_j = k|X_0 = i) = \sum_{j=0}^{n-1}\sum_k r_k P_{i,k}^j. \tag{35}$$

We can write it in more compact form:

$$\boldsymbol{v}(n) = \sum_{j=0}^{n-1} P^j \boldsymbol{r}. \tag{36}$$

Sometimes in our problem we will have a final reward, which is the reward of your destination. In this case our reward can be written as

$$\boldsymbol{v}(n, \boldsymbol{u}) + \sum_{j=0}^{n-1} P^j \boldsymbol{r} + P^n \boldsymbol{u}, \tag{37}$$

where u_i is the final reward vector representing the reward of state i.

5.2 Markov decision theory and dynamic programming

In real world sometimes we need to make decision about our Markov chain (usually about P and $\boldsymbol{r}$). Then the expected aggregate reward is not an constant, but depends on our decision.

Definition 5.1. *For each state i, there are K_i different $(P_{i,:}^{(k)}, r_i^k)$. We have to choose one $(P_{i,:}^{(k_i)}, r_i^{(k_i)})$ for each state i, and all these that decides our transition matrix $P^{(\boldsymbol{k})}$ and $\boldsymbol{r}^{\boldsymbol{k}}$ is called a **decision**. For each time, we can make different decisions, and all these decisions over time is called a **policy**.*

We can see that our expected aggregate reward highly relevalent to our policy. Our goal is to find a "best" policy in terms of maximizing the expected aggregate reward. The following algorithm, called dynamic programming, helps achive this goal.

We first consider the expected aggregate reward in one step, then the maximized reward $v^*(1, \boldsymbol{u})$ is

$$\boldsymbol{v}^*(1, \boldsymbol{u}) \equiv \max_{\boldsymbol{k}} \boldsymbol{v}^{(\boldsymbol{k})}(1, \boldsymbol{u}) = \max_{\boldsymbol{k}} \left\{ \boldsymbol{r}^{(\boldsymbol{k})} + P^{(\boldsymbol{k})} \boldsymbol{u} \right\}, \tag{38}$$

then consider the expected aggregate reward in two steps, we find

$$\begin{aligned} v_i^{(\boldsymbol{k_1}, \boldsymbol{k_2})}(2, \boldsymbol{u}) &= \mathrm{E}\left(R(X_0) + R(X_1) + U(X_1) | X_0 = i\right) \\ &= r_i^{(\boldsymbol{k_2})} + \sum_j \mathrm{E}(R(X_1) + U(X_1) | X_1 = j) P(X_1 = j | X_0 = i) = r_i^{(\boldsymbol{k_1})} + \sum_j v_j^{(\boldsymbol{k_2})}(1, \boldsymbol{u}) P_{i,j}^{(\boldsymbol{k_1})}. \end{aligned} \tag{39}$$

Thus expected aggregate reward vector is

$$\boldsymbol{v}^{(\boldsymbol{k_1}, \boldsymbol{k_2})}(2, \boldsymbol{u}) = \boldsymbol{r}^{(\boldsymbol{k_1})} + P^{(\boldsymbol{k_1})} \boldsymbol{v}^{(\boldsymbol{k_2})}(1, \boldsymbol{u}). \tag{40}$$

Then if we want to maximize $\boldsymbol{v}^{(\boldsymbol{k_1}, \boldsymbol{k_2})}(2, \boldsymbol{u})$ by choosing $(\boldsymbol{k_1}, \boldsymbol{k_2})$, we get

$$\boldsymbol{v}^*(2, \boldsymbol{u}) \equiv \max_{\boldsymbol{k_1}, \boldsymbol{k_2}} v^{(\boldsymbol{k_1}, \boldsymbol{k_2})}(2, \boldsymbol{u}) = \max_{\boldsymbol{k_2}} r^{(\boldsymbol{k_1})} + \max_{\boldsymbol{k_1}} P^{(\boldsymbol{k_1})} \max_{\boldsymbol{k_2}} \boldsymbol{v}^{(\boldsymbol{k_2})}(1, \boldsymbol{u}), \tag{41}$$

where in the last step we use the fact that since $P_{i,j} > 0$, so we can maximize P and $\boldsymbol{v}(1, \boldsymbol{u})$ seperately. So

$$\boldsymbol{v}^*(2, \boldsymbol{u}) = \max_{\boldsymbol{k_1}} r^{(\boldsymbol{k_1})} + \max_{\boldsymbol{k_1}} P^{(\boldsymbol{k_1})} \boldsymbol{v}^*(1, \boldsymbol{u}). \tag{42}$$

So we can see that n-steps expected aggregate reward can be written as the one-step expected aggregate reward with $\boldsymbol{u}$ replaced by $n-1$-step expected aggregate reward:

$$\boldsymbol{v}^*(n, \boldsymbol{u}) = \max_{\boldsymbol{k}} \left\{ r^{(\boldsymbol{k})} + P^{(\boldsymbol{k})} \boldsymbol{v}^*(n-1, \boldsymbol{u}) \right\}, \quad \boldsymbol{v}^*(0, \boldsymbol{u}) \equiv \boldsymbol{u}. \tag{43}$$

This iteration process to maximize the expected aggregate rewards is called **dynamic programming**.

5.3 Shortest path problem

We take shortest path problem as an example of dynamic programming.

Definition 5.2. *The following problem is called a* ***shortest path problem****. Let G be a directed graph with positive length to each edge. Given initial node i, we are asked to find the shortest path going to destination j.*

To model this as Markov decision problem, we can see **each move as a decision with corresponding transition probability 1 and others are 0, and the reward is the negative corresponding length**. To force us to head to destination (say $i = 0$) at the end, we set our final reward as

$$\boldsymbol{u} = \begin{pmatrix} 0 \\ -\infty \\ -\infty \\ \cdots \end{pmatrix} \tag{44}$$

So the shortest path problem is completely restated as a Markov decision problem, which can be handled by applying dynamic programming.