Lecture 4 Markov chain and entropy rate

October 17, 2024

Outline

Definition

Two random variables X and Y are **independent**, denoted by $X \perp Y$, if

$$p(x, y) = p(x)p(y)$$

for all x and y (i.e., for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$).

Definition

For $n \geq 3$, random variables X_1, X_2, $\cdots$, X_n are **mutually independent** if

$$p(x_1, \cdots, x_n) = p(x_1) \cdots p(x_n)$$

for all $x_1, x_2, \cdots, x_n$.

Definition

For $n \geq 3$, random variables X_1, X_2, $\cdots$, X_n are **pairwise independent** if X_i and X_j are independent for all $1 \leq i < j \leq n$.

Definition

For random variables X, Y, and Z, X is **independent of** Z **conditioning on** X, denoted $X \perp Z|Y$, if

$$p(x,y,z)p(y) = p(x,y)p(y,z)$$

for all x, y and z, or equivalently,

$$p(x,y,x) = \begin{cases} \frac{p(x,y)p(y,z)}{p(y)} = p(x,y)p(z|y) & \text{if } p(y) > 0 \\ 0 & \text{otherwise} \end{cases}.$$

Proposition

For random variables X, Y and Z, $X \perp Z|Y$ if and only if

$$p(x, y, z) = a(x, y)b(y, z)$$

for all x, y and z such that $p(y) > 0$.

Proof.

The 'only if' part follows immediately from the definition of conditional independence , so we will only prove the 'if' part. Assume

$$p(x, y, z) = a(x, y)b(y, z)$$

for all x, y and z such that $p(y) > 0$. Then for such x, y and z, we have

$$p(x, y) = \sum_z p(x, y, z) = \sum_z a(x, y)b(y, z) = a(x, y) \sum_z b(y, z)$$

and

$$p(y, z) = \sum_x p(x, y, z) = \sum_x a(x, y)b(y, z) = b(y, z) \sum_x a(x, y).$$

Therefore,

$$\begin{aligned} p(x, y)p(y, z) &= \big(a(x, y) \sum_z b(y, z)\big)\big(b(y, z) \sum_x a(x, y)\big) \\ &= \big(\sum_x a(x, y)\big)\big(\sum_z b(y, z)\big)\big(a(x, y)b(y, z)\big) \\ &= p(y)p(x, y, z). \end{aligned}$$

Hence $X \perp Z|Y$, and the proof is accomplished. □

Stochastic process

Definition

A stochastic process is a collection of random variables that arise from the same probability space. It can be mathematically represented by collection

$$\{X_t,\ t \in I\},$$

where X_t denotes the tth random variable in the process, and the index t runs over an index set I which is arbitrary.

In this course, we focus mostly on discrete-time sources; i.e., sources with the countable index set $I = \{1, 2, \ldots\}$. Each such source is denoted by

$$\mathbf{X} := \{X_n\}_{n=1}^{\infty} = \{X_1, X_2, X_3, \cdots\}$$

as an infinite sequence of random variables, where all the random variables take on values from a common generic alphabet $\mathcal{X} \subset \mathbb{R}$.

The source X completely characterized by the sequence of joint cdf's $\{F_{X^n}\}_{n=1}^{\infty}$. When the alphabet $\mathcal{X}$ is finite, the source can be equivalently described by the sequence of joint probability mass function (pmf's):

$$P_{X^n}(a^n) = Pr[X_1 = a_1, X_2 = a_2, \cdots, X_n = a_n]$$

for all $a^n = (a_1, a_2, \ldots, a_n) \in \mathcal{X}^n$, $n = 1, 2, \ldots$.

Memoryless process

The process **X** is said to be memoryless if its random variables are independent and identically distributed (i.i.d.). Here by independence, we mean that any finite sequence $X_{i_1}, X_{i_2}, \ldots, X_{i_n}$ of random variables satisfies

$$Pr[X_{i_1} = x_1, X_{i_2} = x_2, \ldots, X_{i_n} = x_n] = \Pi_{l=1}^{n} Pr[X_{i_l} = x_l].$$

for all $x_l \in \mathcal{X}$, $l = 1, \ldots, n$; we also say that these random variables are mutually independent. Furthermore, the notion of identical distribution means that

$$Pr[X_i = x] = Pr[X_1 = x]$$

for any $x \in \mathcal{X}$ and $i = 1, 2, \cdots$; i.e., all the process' random variables are governed by the same marginal distribution.

Stationary process

The process X is said to be stationary (or strictly stationary) if the probability of every sequence or event is unchanged by a left (time) shift, or equivalently, if any $j = 1, 2, \ldots$, the joint distribution of $(X_1, X_2, \ldots, X_n)$ satisfies

$$Pr[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] = Pr[X_{j+1} = x_1, X_{j+2} = x_2, \ldots, X_{j+n} = x_n]$$

for all $x_l \in \mathcal{X}$, $l = 1, \ldots, n$.

Stationary process

The process X is said to be stationary (or strictly stationary) if the probability of every sequence or event is unchanged by a left (time) shift, or equivalently, if any $j = 1, 2, \ldots$, the joint distribution of $(X_1, X_2, \ldots, X_n)$ satisfies

$$Pr[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] = Pr[X_{j+1} = x_1, X_{j+2} = x_2, \ldots, X_{j+n} = x_n]$$

for all $x_l \in \mathcal{X}$, $l = 1, \ldots, n$.

It is direct to verify that a memoryless source is stationary. Also, for a stationary source, its random variables are identically distributed.

Outline

Markov chain for three random variables

Example

Three random variables X, Y and Z are said to form a Markov chain if

$$P_{X,Y,Z}(x,y,x) = P_X(x) \cdot P_{Y|X}(y|x) \cdot P_{Z|Y}(z|y);$$

i.e., $P_{Z|X,Y}(z|x,y) = P_{Z|Y}(z|y)$. *This is usually denoted by*

$$X \to Y \to Z.$$

$X \to Y \to Z$ is sometimes read as "X and Z are conditionally independent given Y" because it can be shown that the above definition is equivalent to

$$P_{X,Z|Y}(x, z|y) = P_{X|Y}(x|y) \cdot P_{Z|Y}(z|y).$$

Therefore, $X \to Y \to Z$ is equivalent to $Z \to Y \to X$.
Accordingly, the Markovian notation is sometimes expressed as $X \leftrightarrow Y \leftrightarrow Z$.

kth order Markov chain

The sequence of random variables $\{X_n\}_{n=1}^{\infty} = X_1, X_2, X_3, \ldots$ with common finite-alphabet $\mathcal{X}$ is said to form a kth order Markov chain (or kth order Markov source or process) if for all $n > k$, $x_1 \in \mathcal{X}$, $i = 1, \ldots, n$,

$$\begin{aligned} &Pr[X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_1 = x_1] \\ &\qquad = Pr[X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_{n-k} = x_{n-k}]. \end{aligned}$$

Each $x_{n-k}^{n-1} := (x_{n-k}, x_{n-k+1}, \ldots, x_{n-1}) \in \mathcal{X}^k$ is called the state of the Markov chain at time n.

When $k = 1$, then $\{X_n\}_{n=1}^{\infty}$ is called a first-order Markov chain (or just a Markov chain). This means that for any $n > 1$, the random variables $X_1, X_2, \ldots, X_n$ directly satisfy the conditional independence property

$$\Pr[X_i = x_i | X^{i-1} = x^{i-1}] = \Pr[X_i = x_i | X_{i-1} = x_{i-1}]$$

for all $x_i \in \mathcal{X}$, $i = 1, \ldots, n$; this property is denoted by

$$X_1 \to X_2 \to \cdots \to X_n$$

for $n > 2$. We also say the that $X_1 \to X_2 \to \cdots \to X_n$ forms a Markov chain.

Proposition

For random variables X_1, X_2, $\cdots$, X_n, *where* $n \geq 3$,
$X_1 \to X_2 \to \cdots \to X_n$ *forms a Markov chain if and only if*

$$p(x_1, \cdots, x_n)p(x_2)p(x_3)\cdots p(x_{n-1}) = p(x_1, x_2)p(x_2, x_3)\cdots p(x_{n-1}, x_n)$$

for all x_1, x_2, $\cdots$, x_n, *or equivalently,*

$$p(x_1, x_2, \cdots, x_n) =$$
$$\begin{cases} p(x_1, x_2)p(x_3|x_2)\cdots p(x_n|x_{n-1}) & \text{if } p(x_2),\ p(x_3),\ \cdots,\ p(x_{n-1}) > 0 \\ 0 & \text{otherwise} \end{cases}.$$

Proposition

$X_1 \to X_2 \to \cdots \to X_n$ forms a Markov chain if and only if $X_n \to X_{n-1} \to \cdots \to X_1$ forms a Markov chain.

Proposition

$X_1 \to X_2 \to \cdots X_n$ *forms a Markov chain if and only if*

$$
\begin{aligned}
&X_1 \to X_2 \to X_3 \\
&(X_1, X_2) \to X_3 \to X_4 \\
&\qquad \vdots \\
&(X_1, X_2, \cdots X_{n-2}) \to X_{n-1} \to X_n
\end{aligned}
$$

form Markov chains.

Proposition

$X_1 \to X_2 \to \cdots \to X_n$ forms a Markov chain if and only if

$$p(x_1, x_2, \cdots, x_n) = f_1(x_1, x_2) f_2(x_2, x_3) \cdots f_{n-1}(x_{n-1}, x_n)$$

for all $x_1, x_2, \cdots, x_n$ such that $p(x_2), p(x_3), \cdots, p(x_{n-1}) > 0$.

A kth order Markov chain is said to be *time-invariant* or *homogeneous*, if for every $n > k$,

$$Pr[X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_{n-k} = x_{n-k}] \\ = Pr[X_{k+1} = x_{k+1} | X_k = x_k, \ldots, X_1 = x_1].$$

Therefore, a homogeneous first-order Markov chain can be defined through its transition probability:

$$[Pr\{X_2 = x_2 | X_1 = x_1\}]_{|\mathcal{X} \times \mathcal{X}|},$$

and its initial state distribution $P_{X_1}(x)$.

Irreducible Markov chain

A kth order Markov chain is *irreducible* if with some probability, we can go from any state in $\mathcal{X}^k$ to another state in a finite number of steps, i.e., for all $x^k, y^k \in \mathcal{X}^k$ there exists an integer $j \geq 1$ such that

$$Pr\{X_j^{k+j-1} = x^k | X_1^k = y^k\} > 0.$$

In a first-order Markov chain, the period $d(x)$ of state $x \in \mathcal{X}$ is defined by

$$d(x) := \text{gcd}\{n \in \{1, 2, 3, \ldots\} : Pr\{X_{n+1} = x | X_1 = x\} > 0\},$$

where gcd denotes the greatest common divisor; in other words, if the Markov chain starts in state x, then the chain cannot return to state x at any time that is not a multiple of $d(x)$. If $Pr\{X_{n+1} = x | X_1 = x\} = 0$ for all n, we say that state x has infinite period and write $d(x) = \infty$. We also say that state x is *aperiodic* if $d(x) = 1$ and *periodic* if $d(x) > 1$. Furthermore, the first-order Markov chain is called *aperiodic* if all its states are aperiodic. In other words, the first-order Markov chain ia aperiodic if

$$\text{gcd}\{n \in \{1, 2, 3, \ldots\} : Pr\{X_{n+1} = x | X_1 = x\} > 0\} = 1 \ \forall x \in \mathcal{X}.$$

In an irreducible first-order Markov chain, all states have the same period. Hence, if one state in such a chain is aperiodic, then the entire Markov chain is aperiodic.

A distribution $\pi(\cdot)$ on $\mathcal{X}$ is said to be a *stationary* distribution for a homogeneous first-order Markov chain, if for every $y \in \mathcal{X}$,

$$\pi(y) = \sum_{x \in \mathcal{X}} \pi(x) \Pr\{X_2 = y | X_1 = x\}.$$

For a finite-alphabet homogeneous first-order Markov chain, $\pi(\cdot)$ always exists; furthermore, if the Markov chain is irreducible and aperiodic, the stationary distribution $\pi(\cdot)$ is unique, and

$$\lim_{n \to \infty} \Pr\{X_{n+1} = y | X_1 = x\} = \pi(y)$$

for all states x and y in $\mathcal{X}$. If the initial state distribution is equal to stationary distribution, then the homogenous first-order Markov chain becomes a stationary distribution, then the homogenous first-order Markov chain becomes a stationary process.

Example

Consider a two-state Markov chain with probability transition matrix

$$P = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}.$$

The stationary distribution μ can be found by solving the equation $\mu P = \mu$. We have that $\mu = (\mu_1, \mu_2)$, where

$$\mu_1 = \frac{\beta}{\alpha+\beta}, \ \mu_2 = \frac{\alpha}{\alpha+\beta}.$$

So the entropy of the state X_n at time n is

$$H(X_n) = H(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}).$$

Outline

Entropy rate

Definition

The entropy rate of a stochastic process $\{X_i\}$ is defined by

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \cdots, X_n)$$

when the limit exists.

Example

If $X_1, X_2, \ldots$ are i.i.d. random variables. Then

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{H(X_1, X_2, \cdots, X_n)}{n} = \lim_{n\to\infty} \frac{nH(X_1)}{n} = H(X_1).$$

Example

If the random variables $X_1, X_2, \cdots, X_n$ are independent but not identically distributed, then

$$H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i).$$

We choose a sequence of distributions on $X_1, X_2, \cdots$, such that the limit of $\frac{1}{n}\sum H(X_i)$ does not exist. An example of such a sequence is a random binary sequence where $p_i = P(X_i = 1)$ is not constant but a function of i. For example,

$$p_i = \begin{cases} 0.5 & 2k < \log\log i \leq 2k+1 \\ 0 & 2k+1 < \log\log i \leq 2k+2. \end{cases}$$

The running average of the $H(X_i)$ will oscillate between 0 and 1 and will not have a limit. Thus, $H(\mathcal{X})$ is not defined for this process.

We can also define a related quantity for entropy rate:

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_{n-1}, X_{n-2}, \cdots, X_1)$$

Theorem

For a stationary stochastic process, the above two limits exist and are equal.

$$H(\mathcal{X}) = H'(\mathcal{X}).$$

Theorem

For a stationary stochastic process, $H(X_n|X_{n-1},\cdots,X_1)$ is nonincreasing in n and has a limit $H'(\mathcal{X})$.

Proof.

We have

$$\begin{aligned} H(X_{n+1}|X_1,X_2,\cdots,X_n) &\leq H(X_{n+1}|X_n,\cdots,X_2) \\ &= H(X_n|X_{n-1},\cdots,X_1). \end{aligned}$$

where the equality follows from the stationary of the process. Since $H(X_n|X_{n-1},\cdots,X_1)$ is a decreasing sequence of nonnegative numbers, it has a limit, $H'(\mathcal{X})$. □

Lemma

If $a_n \to a$ *as* $n \to \infty$ *and*

$$b_n = \frac{1}{n} \sum_{i=1}^{n} a_i,$$

then $b_n \to a$ *as* $n \to \infty$.

Proof of the theorem.

By the chain rule,

$$\frac{H(X_1, X_2, \ldots, X_n)}{n} = \frac{1}{n} \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1),$$

that is, the entropy rate is the time average of the conditional entropies. But we know that the conditional entropies tend to a limit H'. Hence, their running average has a limit, which is equal to the limit H' of the terms. Thus,

$$H(\mathcal{X}) = \lim \frac{H(X_1, X_2, \cdots, X_n)}{n} = \lim H(X_n | X_{n-1}, \cdots, X_1) = H'(\mathcal{X}).$$

□

For a stationary Markov chain, the entropy rate is given by

$$H(\mathcal{X}) = H'(\mathcal{X}) = \lim H(X_n|X_{n-1}, \cdots, X_1) = \lim H(X_n|X_{n-1}) = H(X_2|X_1),$$

where the third equality follows from

$$\begin{aligned}
& H(X_n|X_{n-1}, \cdots, X_1) \\
= & -\sum_{x_1,\cdots,x_n} p(x_1, x_2, \cdots, x_{n-1}) p(x_n|x_{n-1}, \cdots, x_1) \log p(x_n|x_{n-1}, \cdots, x_1) \\
= & -\sum_{x_1,\cdots,x_n} p(x_1, x_2, \cdots, x_{n-1}) p(x_n|x_{n-1}) \log p(x_n|x_{n-1}) \\
= & -\sum_{x_{n-1},x_n} p(x_n|x_{n-1}) \log p(x_n|x_{n-1}) \sum_{x_1,\cdots,x_{n-2}} p(x_1, x_2, \cdots, x_{n-1}) \\
= & -\sum_{x_{n-1},x_n} p(x_{n-1}) p(x_n|x_{n-1}) \log p(x_n|x_{n-1}) \\
= & -\sum_{x_{n-1},x_n} p(x_{n-1}, x_n) \log p(x_n|x_{n-1}) \\
= & H(X_n|X_{n-1}).
\end{aligned}$$

Recall that the stationary distribution μ is the solution of the equations

$$\mu_j = \sum_i \mu_i P_{ij} \text{ for all } j.$$

We can express the conditional entropy explicitly in the following theorem.

Theorem

Let $\{X_i\}$ be a stationary Markov chain with stationary distribution μ and transition matrix P. Let $X_1 \sim \mu$. Then the entropy rate is

$$H(\mathcal{X}) = -\sum_{ij} \mu_i P_{ij} \log P_{ij}.$$

Proof.

$$H(\mathcal{X}) = H(X_2|X_1) = \sum_i \mu_i (\sum_j -P_{ij} \log P_{ij}).$$

□

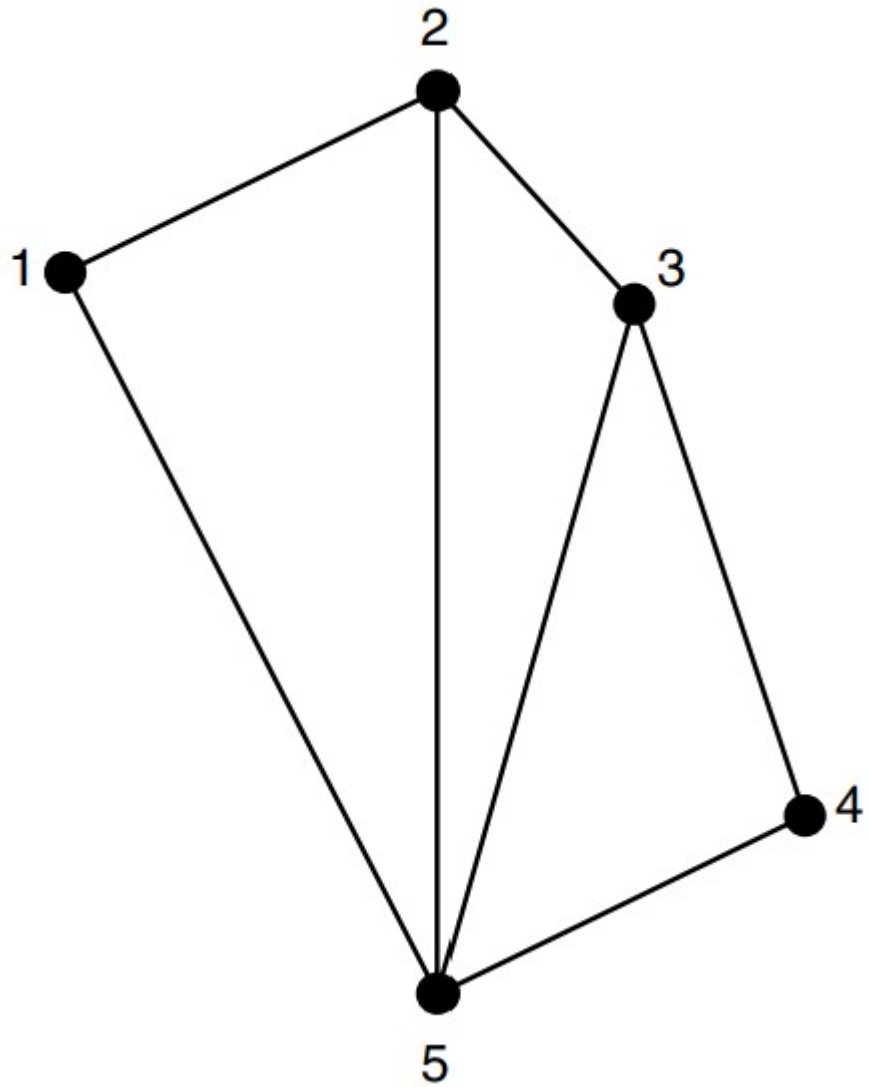
Two-state Markov chain

Example

The entropy rate of the two-state Markov chain is

$$H(\mathcal{X}) = H(X_2|X_1) = \frac{\beta}{\alpha+\beta}H(\alpha) + \frac{\alpha}{\alpha+\beta}H(\beta).$$

- As an example of a stochastic process, let us consider a random walk on a connected graph.
- Consider a graph with m nodes labeled $\{1, 2, \cdots, m\}$ with weight $W_{ij} \geq 0$ on the edge joining node i to node j.
- The graph is assumed to be undirected, so that $W_{ij} = W_{ji}$.
- We set $W_{ij} = 0$ if there is no edge joining the node i and the node j.


2
1
3
4
5

- A particle walks randomly from node to node in this graph.
- The random walk $\{X_n\}$, $X_n \in \{1, 2, \cdots, m\}$, is a sequence of vertices of the graph.
- Given $X_n = i$, the next vertex j is chosen from among the nodes connected to node i with a probability proportional to the weight of the edge connecting i to j.
- Thus, $P_{ij} = W_{ij} / \sum_k W_{ik}$.

- In this case, the stationary distribution has a surprisingly simple form, which we will guess and verify.
- The stationary distribution for this Markov chain assigna probability to node i proportional to the total weight of the edges emanating from node i.
- Let $W_i = \sum_j W_{ij}$ be the total weight of edges emanating from node i, and let

$$W = \sum_{i,j:j>i} W_ij$$

be the sum of weights of all the edges. Then $\sum_i W_i = 2W$.

- We now guess that the stationary distribution is $\mu_i = \frac{W_i}{2W}$.
- We verify that this is the stationary distribution by checking that $\mu P = \mu$. Here

$$\begin{aligned}\sum_i \mu_i P_{ij} &= \sum_i \frac{W_i}{2W}\frac{W_{ij}}{W_i} \\ &= \sum_i \frac{1}{2W} W_{ij} \\ &= \frac{W_j}{2W} \\ &= \mu_j.\end{aligned}$$

- Thus, the stationary probability of state i is proportional to weight of edges emanating from node i.
- This stationary distribution has an interesting property of locality.
- It depends only on the total weight and the weight of edges connected to the node and hence does not change if the weights in some other part of the graph are changed while keeping the total weight constant.

We can now calculate the entropy rate as

$$\begin{aligned}
H(\mathcal{X}) &= H(X_2|X_1) \\
&= -\sum_i \mu_i \sum_j P_{ij} \log P_{ij} \\
&= -\sum_i \frac{W_i}{2W} \sum_j W_{ij} W_i \log \frac{W_{ij}}{W_i} \\
&= -\sum_i \sum_j \frac{W_{ij}}{2W} \log \frac{W_{ij}}{W_i} \\
&= -\sum_i \sum_j \frac{W_{ij}}{2W} \log \frac{W_{ij}}{2W} + \sum_i \sum_j \frac{W_{ij}}{2W} \log \frac{W_i}{2W} \\
&= H(\cdots, \frac{W_{ij}}{2W}, \cdots) - H(\cdots, \frac{W_i}{2W}, \cdots).
\end{aligned}$$

If all the edges have equal weight, the stationary distribution puts weight $E_i/2E$ on node i, where E_i is the number of edges emanating from node i and E is the total number of edges in the graph. In this case, the entropy rate of the random walk is

$$H(\mathcal{X}) = \log(2E) - H(\frac{E_1}{2E}, \frac{E_2}{2E}, \cdots, \frac{E_m}{2E}).$$

Remark

- It is easy to see that a stationary random walk on a graph is **time-reversible**; that is, the probability of any sequence of states is the same forward or backward:

$$\begin{aligned}&\Pr(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) \\ &\qquad = \Pr(X_n = x_1, X_{n-1} = x_2, \cdots, X_1 = x_n).\end{aligned}$$

- Rather surprisingly, the converse is also true; that is, any time-reversible Markov chain can be represented as a random walk on an undirected weighted graph.