Elements of Information Theory

George Voutsadakis[1]

[1] Mathematics and Computer Science
Lake Superior State University

LSSU Math 500

1. Entropy Rates of a Stochastic Process
    - Markov Chains
    - Entropy Rate
    - Example: Entropy Rate of a Random Walk on a Weighted Graph
    - Second Law of Thermodynamics
    - Functions of Markov Chains

Subsection 1

Markov Chains

Stochastic Processes

- A **stochastic process** $\{X_i\}$ is an indexed sequence of random variables.
- There may be an arbitrary dependence among the random variables.
- The process is characterized by the joint probability mass functions, given, for all $n = 1, 2, \ldots$ and all $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$, by

$$\Pr\{(X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n)\} = p(x_1, x_2, \ldots, x_n).$$

Stationary Stochastic Processes

Definition

A stochastic process is said to be **stationary** if the joint distribution of any subset of the sequence of random variables is invariant with respect to shifts in the time index. I.e., $\{X_i\}$ is stationary if

$$\begin{aligned}\Pr\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\} \\ = \Pr\{X_{1+\ell} = x_1, X_{2+\ell} = x_2, \ldots, X_{n+\ell} = x_n\},\end{aligned}$$

for every n, every shift ℓ and all $x_1, x_2, \ldots, x_n \in \mathcal{X}$.

Markov Chains

- A simple example of a stochastic process is one in which each random variable:
  - Depends only on the one preceding it;
  - Is conditionally independent of all the other preceding random variables.

Definition

A discrete stochastic process $X_1, X_2, \ldots$ is said to be a **Markov chain** or a **Markov process** if, for $n = 1, 2, \ldots$,

$$\begin{gathered}\Pr(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1) \\ = \Pr(X_{n+1} = x_{n+1} | X_n = x_n),\end{gathered}$$

for all $x_1, x_2, \ldots, x_n, x_{n+1} \in \mathcal{X}$.

- In this case, the joint probability mass function of the random variables can be written as

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

Time Invariant Markov Chains

Definition

A Markov chain is said to be **time invariant** if the conditional probability $p(x_{n+1}|x_n)$ does not depend on n, i.e., for $n = 1, 2, \ldots$,

$$\Pr\{X_{n+1} = b|X_n = a\} = \Pr\{X_2 = b|X_1 = a\}, \text{ for all } a, b \in \mathcal{X}.$$

- We will assume that the Markov chain is time invariant unless otherwise stated.
- If $\{X_i\}$ is a Markov chain, X_n is called the **state** at time n.
- A time invariant Markov chain is characterized by:
  - Its initial state;
  - A **probability transition matrix** $P = [P_{ij}]$, $i, j \in \{1, 2, \ldots, m\}$, where $P_{ij} = \Pr\{X_{n+1} = j|X_n = i\}$.

Irreducibility and Aperiodicity

- A Markov chain is called **irreducible** if it is possible to go with positive probability from any state of the Markov chain to any other state in a finite number of steps.
- A Markov chain is called **aperiodic** if the largest common factor of the lengths of different paths from a state to itself is 1.

Stationary Distribution

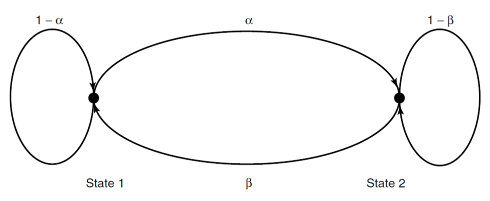
- If the probability mass function of the random variable at time n is $p(x_n)$, the probability mass function at time $n+1$ is

$$p(x_{n+1}) = \sum_{x_n} p(x_n) P_{x_n x_{n+1}}.$$

- A **stationary distribution** is a distribution on the states, such that the distribution at time $n+1$ is the same as the distribution at time n.
- If the initial state of a Markov chain is drawn according to a stationary distribution, the Markov chain forms a stationary process.
- If the finite-state Markov chain is irreducible and aperiodic, then:
  - The stationary distribution is unique;
  - From any starting distribution, the distribution of X_n tends to the stationary distribution as $n \to \infty$.

Example

- Consider a two-state Markov chain with a probability transition matrix



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

Let the stationary distribution be represented by a vector μ whose components are the stationary probabilities of States 1 and 2.

Then the stationary probability can be found by solving the equation $\mu P = \mu$ or, more simply, by balancing probabilities.

For the stationary distribution, the net probability flow across any cut set in the state transition graph is zero.

Applying this to the figure we obtain $\mu_1 \alpha = \mu_2 \beta$.

Example (Cont'd)

- We found stationary distribution

$$\mu_1\alpha = \mu_2\beta.$$

Since $\mu_1 + \mu_2 = 1$, the stationary distribution is

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

If the Markov chain has an initial state drawn according to the stationary distribution, the resulting process will be stationary.

The entropy of the state X_n at time n is

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

This is not the rate at which entropy grows for $H(X_1, X_2, \ldots, X_n)$.

The dependence among the X_i's will take a steady toll.

Subsection 2

Entropy Rate

Entropy of a Stochastic Process

Definition

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

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

when the limit exists.

Example (**Typewriter**): Consider the case of a typewriter that has m equally likely output letters. The typewriter can produce m^n sequences of length n, all of them equally likely. Hence

$$H(X_1, X_2, \ldots, X_n) = \log m^n.$$

The entropy rate is $H(\mathcal{X}) = \log m$ bits per symbol.

Example ($X_1, X_2, \ldots$ i.i.d. Random Variables)

- Suppose $X_1, X_2, \ldots$ are i.i.d. random variables.

  Then

$$\begin{aligned} H(\mathcal{X}) &= \lim \tfrac{H(X_1,X_2,\ldots,X_n)}{n} \\ &= \lim \tfrac{nH(X_1)}{n} \\ &= H(X_1). \end{aligned}$$

  This is what one would expect for the entropy rate per symbol.

Process with Undefined Entropy

- Suppose $X_1, X_2, \ldots$ are independent but not identically distributed random variables.

  Then
  $$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i),$$
  but the $H(X_i)$'s are not all equal.

  We can choose a sequence of distributions on $X_1, X_2, \ldots$, 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;
  - p_i is chosen carefully so that the limit $H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n)$ does not exist.

Process with Undefined Entropy (Cont'd)

- Let

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

  for $k = 0, 1, 2, \ldots$.

  Then there are arbitrarily long stretches where $H(X_i) = 1$, followed by exponentially longer segments where $H(X_i) = 0$.

  Hence, the running average of the $H(X_i)$ will oscillate between 0 and 1 and will not have a limit.

  Thus, $H(X)$ is not defined for this process.

Another Definition for Entropy for a Process

- We define a related quantity for entropy rate,

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

  when the limit exists.
- The two quantities $H(\mathcal{X})$ and $H'(\mathcal{X})$ correspond to two different notions of entropy rate.
  - The first is the per symbol entropy of the n random variables.
  - The second is the conditional entropy of the last random variable given the past.

$H'(\mathcal{X})$ for a Stationary Process

Theorem

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

- We have

$$\begin{array}{rcl} H(X_{n+1}|X_1,X_2,\ldots,X_n) & \leq & H(X_{n+1}|X_n,\ldots,X_2) \\ & & \text{(conditioning reduces entropy)} \\ & = & H(X_n|X_{n-1},\ldots,X_1). \\ & & \text{(stationarity of the process)} \end{array}$$

Since $H(X_n|X_{n-1},\ldots,X_1)$ is a decreasing sequence of nonnegative numbers, it has a limit $H'(\mathcal{X})$.

Cesáro Mean

Theorem (Cesáro Mean)

If $a_n \to a$ and $b_n = \frac{1}{n}\sum_{i=1}^n a_i$, then $b_n \to a$.

Informal Outline: Since most of the terms in the sequence $\{a_k\}$ are eventually close to a, then b_n, which is the average of the first n terms, is also eventually close to a.

Formal Proof: Let $\epsilon > 0$. Since $a_n \to a$, there exists a number $N(\epsilon)$, such that $|a_n - a| \leq \epsilon$, for all $n \geq N(\epsilon)$. Hence, for all $n \geq N(\epsilon)$,

$$\begin{aligned} |b_n - a| &= |\tfrac{1}{n}\textstyle\sum_{i=1}^n (a_i - a)| \leq \tfrac{1}{n}\sum_{i=1}^n |a_i - a| \\ &\leq \tfrac{1}{n}\textstyle\sum_{i=1}^{N(\epsilon)} |a_i - a| + \frac{n-N(\epsilon)}{n}\epsilon \\ &\leq \tfrac{1}{n}\textstyle\sum_{i=1}^{N(\epsilon)} |a_i - a| + \epsilon. \end{aligned}$$

The first term goes to 0 as $n \to \infty$. Hence, we can make $|b_n - a| \leq 2\epsilon$ by taking n large enough. So $b_n \to a$ as $n \to \infty$.

$H(\mathcal{X})$ and $H'(\mathcal{X})$ for a Stationary Process

Theorem

For a stationary stochastic process, the limits $H(\mathcal{X})$ and $H'(\mathcal{X})$ exist and are equal:

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

- By the chain rule,

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

So the entropy rate is the time average of the conditional entropies.

But we showed that the conditional entropies tend to a limit H'.

$H(\mathcal{X})$ and $H'(\mathcal{X})$ for a Stationary Process (Cont'd)

- By the Cesáro Mean Theorem, the running average of the conditional entropies has a limit.

  Moreover, this limit is equal to the limit H' of the terms.

  Thus, by a previous theorem,

$$\begin{aligned} H(\mathcal{X}) &= \lim \frac{H(X_1, X_2, \ldots, X_n)}{n} \\ &= \lim \frac{1}{n} \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1) \\ &= \lim H(X_n | X_{n-1}, \ldots, X_1) \\ &= H'(\mathcal{X}). \end{aligned}$$

The Case of Markov Chains

- For a stationary Markov chain, the entropy rate, when the conditional entropy is calculated using the given stationary distribution, is given by
$$\begin{aligned} H(\mathcal{X}) &= H'(\mathcal{X}) \\ &= \lim H(X_n|X_{n-1},\ldots,X_1) \\ &= \lim H(X_n|X_{n-1}) \\ &= H(X_2|X_1). \end{aligned}$$
- Recall that the stationary distribution μ is the solution of the equations
$$\mu_j = \sum_i \mu_i P_{ij}, \quad \text{for all } j.$$

Entropy Rate of a Stationary Markov Chain

Theorem

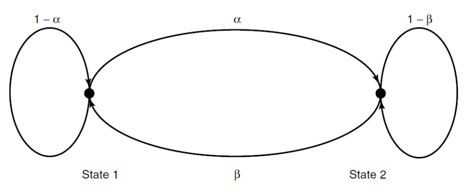
Let $\{X_i\}$ be a stationary Markov chain with stationary distribution μ and transition matrix P. Let $X_1 \sim \mu$. Then the entropy rate is

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

- We have
$$\begin{array}{rcl} H(\mathcal{X}) & = & H(X_2 | X_1) \\ & = & \sum_i \mu_i (\sum_j -P_{ij} \log P_{ij}). \end{array}$$

Example

- We look at the following two-state Markov chain.



Its entropy rate is

$$\begin{aligned} H(\mathcal{X}) &= H(X_2|X_1) \\ &= \tfrac{\beta}{\alpha+\beta} H(\alpha) + \tfrac{\alpha}{\alpha+\beta} H(\beta). \end{aligned}$$

Subsection 3

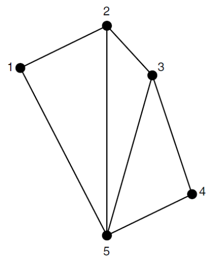
Example: Entropy Rate of a Random Walk on a Weighted Graph

Random Walk on a Weighted Graph

- We consider a random walk on a connected graph with m nodes labeled $\{1, 2, \ldots, m\}$, with weight $W_{ij} \geq 0$ on the edge joining node i to node j.

  The graph is undirected, so that $W_{ij} = W_{ji}$.

  We set $W_{ij} = 0$ if there is no edge joining nodes i and j.



  A particle walks randomly from node to node in this graph.

  The random walk $\{X_n\}$ 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:

$$P_{ij} = \frac{W_{ij}}{\sum_k W_{ik}}.$$

The Stationary Distribution: Guessing

- The stationary distribution for this Markov chain assigns 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.
- Let $W = \displaystyle\sum_{i,j:j>i} W_{ij}$ be the sum of the weights of all the edges.
- Then $\sum_i W_i = 2W$.
- We now guess that the stationary distribution is

$$\mu_i = \frac{W_i}{2W}.$$

The Stationary Distribution: Verification

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

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

- Thus, the stationary probability of state i is proportional to the 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.
  - Hence, it does not change if the weights in some other part of the graph are changed while keeping the total weight constant.

The Entropy Rate of the Random Walk

- We can now calculate the entropy rate as

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

The Entropy Rate in a Special Case

- Suppose all the edges have equal weight.
- Then the stationary distribution puts weight $\frac{E_i}{2E}$ on node i, where:
  - E_i is the number of edges emanating from node i;
  - E is the total number of edges in the graph.
- In this case, the entropy rate of the random walk is

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

Random Walk on a Chessboard

- Let a king move at random on an 8×8 chessboard.

  The king has eight moves in the interior, five moves at the edges, and three moves at the corners.

  Using this and the preceding results, the stationary probabilities are, respectively,

  $$\frac{8}{420}, \quad \frac{5}{420}, \quad \frac{3}{420}.$$

  The entropy rate is $0.92 \log 8$.

  The factor of 0.92 is due to edge effects.

  We would have an entropy rate of $\log 8$ on an infinite chessboard.
- Similarly, we can find the entropy rate of rooks ($\log 14$ bits, since the rook always has 14 possible moves), bishops, and queens.

Random Walks and Time Reversibility

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

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

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

Subsection 4

Second Law of Thermodynamics

Second Law of Thermodynamics and Entropy

- The second law of thermodynamics states that the entropy of an isolated system is nondecreasing.
- In statistical thermodynamics, entropy is often defined as the log of the number of microstates in the system.
- This corresponds exactly to our notion of entropy if all the states are equally likely.
- We model the isolated system as a Markov chain with transitions obeying the physical laws governing the system.
- Implicit in this assumption is the notion of an overall state of the system and the fact that knowing the present state, the future of the system is independent of the past.
- We will find that the entropy does not always increase.
- However, relative entropy always decreases.

Relative Entropy $D(\mu_n \| \mu'_n)$ Decreases

- Let μ_n and μ'_n be two probability distributions on the state space of a Markov chain at time n.

  Let μ_{n+1} and μ'_{n+1} be the corresponding distributions at time $n+1$.

  Let the corresponding joint mass functions be denoted by p and q.

  Let $r(\cdot|\cdot)$ be the probability transition function of the Markov chain.

  Then we have

$$\begin{array}{rcl} p(x_n, x_{n+1}) & = & p(x_n) r(x_{n+1}|x_n); \\ q(x_n, x_{n+1}) & = & q(x_n) r(x_{n+1}|x_n). \end{array}$$

  By the chain rule for relative entropy, we have two expansions

$$\begin{array}{l} D(p(x_n, x_{n+1}) \| q(x_n, x_{n+1})) \\ = D(p(x_n) \| q(x_n)) + D(p(x_{n+1}|x_n) \| q(x_{n+1}|x_n)) \\ \ = D(p(x_{n+1}) \| q(x_{n+1})) + D(p(x_n|x_{n+1}) \| q(x_n|x_{n+1})). \end{array}$$

Relative Entropy $D(\mu_n \| \mu'_n)$ Decreases (Cont'd)

- We obtained

$$\begin{aligned}&D(p(x_n, x_{n+1}) \| q(x_n, x_{n+1}))\\ &= D(p(x_n) \| q(x_n)) + D(p(x_{n+1}|x_n) \| q(x_{n+1}|x_n))\\ &= D(p(x_{n+1}) \| q(x_{n+1})) + D(p(x_n|x_{n+1}) \| q(x_n|x_{n+1})).\end{aligned}$$

  Now both p and q are derived from the Markov chain.

  So both $p(x_{n+1}|x_n)$ and $q(x_{n+1}|x_n)$ are equal to $r(x_{n+1}|x_n)$.

  Hence $D(p(x_{n+1}|x_n) \| q(x_{n+1}|x_n)) = 0$.

  But $D(p(x_n|x_{n+1}) \| q(x_n|x_{n+1})) \geq 0$.

  Therefore, $D(p(x_n) \| q(x_n)) \geq D(p(x_{n+1}) \| q(x_{n+1}))$.

  Equivalently

$$D(\mu_n \| \mu'_n) \geq D(\mu_{n+1} \| \mu'_{n+1}).$$

  So the distance between the probability mass functions is decreasing with time n for any Markov chain.

Relative Entropy $D(\mu_n\|\mu)$ Decreases

- In the preceding case, let μ'_n be any stationary distribution μ.

  Then the distribution μ'_{n+1} at the next time is also equal to μ.

  Hence,

  $$D(\mu_n\|\mu) \geq D(\mu_{n+1}\|\mu).$$

  This implies that any state distribution gets closer and closer to each stationary distribution as time passes.
- The sequence $D(\mu_n\|\mu)$ is a monotonically nonincreasing nonnegative sequence and must therefore have a limit.
- The limit is zero if the stationary distribution is unique, but this is more difficult to prove.

Entropy Does Not Increase In General

- In general, the fact that the relative entropy decreases does not imply that the entropy increases.

  Example: Consider any Markov chain with a nonuniform stationary distribution.

  Suppose we start this Markov chain from the uniform distribution, which already is the maximum entropy distribution.

  The distribution will tend to the stationary distribution.

  The limiting distribution has lower entropy than the uniform one.

  So the entropy decreases with time.

Entropy Increases if Stationary Distribution is Uniform

- In the special case where the stationary distribution is the uniform distribution, entropy increases with time.

  In fact, we can express the relative entropy as

$$\begin{aligned} D(\mu_n \| \mu) &= \sum \mu_n(x) \log \frac{\mu_n(x)}{u(x)} \\ &= \sum \mu_n(x) \log |\mathcal{X}| - \sum \mu_n(x) \log \mu_n(x) \\ &= \log |\mathcal{X}| - H(\mu_n) \\ &= \log |\mathcal{X}| - H(X_n). \end{aligned}$$

  In this case the monotonic decrease in relative entropy implies a monotonic increase in entropy.

Double Stochasticity and Uniform Stationary Distribution

Definition

A probability transition matrix $[P_{ij}]$, $P_{ij} = \Pr\{X_{n+1} = j | X_n = i\}$, is called **doubly stochastic** if

$$\sum_i P_{ij} = 1,\ j = 1, 2, \ldots, \quad \sum_j P_{ij} = 1,\ i = 1, 2, \ldots.$$

Claim: The uniform distribution is a stationary distribution of P if and only if the probability transition matrix is doubly stochastic.

Double Stochasticity and Uniform Distribution (Cont'd)

- Assume, first, that P is doubly stochastic.

  Then, for $\mu(i) = \frac{1}{N}$, $i = 1, \ldots, N$, we have

  $$(\mu P)(j) = \sum_{i=1}^{N} \mu(i) \sum_{i=1}^{N} P_{ij} = \frac{1}{N} \sum_{i=1}^{N} P_{ij} = \frac{1}{N} = \mu(j).$$

  So the uniform distribution is a stationary distribution.

  Suppose, conversely, that P is not doubly stochastic.

  Then, there exists a k, $1 \leq k \leq N$, such that $\sum_{i=1}^{N} P_{ik} \neq 1$.

  Then

  $$\sum_{i=1}^{N} \frac{1}{N} P_{ik} = \frac{1}{N} \sum_{i=1}^{N} P_{ik} \neq \frac{1}{N}.$$

  Thus, the uniform distribution cannot be stationary.

The Conditional Entropy $H(X_n|X_1)$ for Stationary X

- Claim: The conditional uncertainty $H(X_n|X_1)$ of a Markov process increases with n.

  Proof 1: Using the properties of entropy, we get

  $$\begin{aligned} H(X_n|X_1) &\geq H(X_n|X_1,X_2) \quad \text{(conditioning reduces entropy)} \\ &= H(X_n|X_2) \quad \text{(by Markovity)} \\ &= H(X_{n-1}|X_1) \quad \text{(by stationarity).} \end{aligned}$$

  Proof 2: Applying the data-processing inequality to the Markov chain $X_1 \to X_{n-1} \to X_n$, we have $I(X_1;X_{n-1}) \geq I(X_1;X_n)$.

  Expanding mutual information in terms of entropies,

  $$H(X_{n-1}) - H(X_{n-1}|X_1) \geq H(X_n) - H(X_n|X_1).$$

  By stationarity, $H(X_{n-1}) = H(X_n)$. Hence, $H(X_{n-1}|X_1) \leq H(X_n|X_1)$.
- These techniques can also be used to show that $H(X_0|X_n)$ is increasing in n for any Markov chain.

Shuffles Increase Entropy

- If T is a shuffle (permutation) of a deck of cards and X is the initial (random) position of the cards in the deck, and if the choice of the shuffle T is independent of X, then

  $$H(TX) \geq H(X),$$

  where TX is the permutation of the deck induced by the shuffle T on the initial permutation X.

  We have

  $$\begin{aligned} H(TX) &\geq H(TX|T) \quad \text{(conditioning reduces entropy)} \\ &= H(T^{-1}TX|T) \quad \text{(given } T\text{, shuffle can be reversed)} \\ &= H(X|T) \\ &= H(X). \end{aligned}$$

Subsection 5

Functions of Markov Chains

Functions of Markov Chains

- Let $X_1, X_2, \ldots, X_n, \ldots$ be a stationary Markov chain.
- Let $Y_i = \phi(X_i)$ be a process each term of which is a function of the corresponding state in the Markov chain.
- We would like to compute the entropy rate $H(\mathcal{Y})$.
- It would simplify matters greatly if $Y_1, Y_2, \ldots, Y_n$ also formed a Markov chain, but in many cases, this is not true.
- Since the Markov chain is stationary, so is $Y_1, Y_2, \ldots, Y_n$.
- So the entropy rate is well defined.
- However, if we wish to compute $H(\mathcal{Y})$, we might try to compute $H(Y_n | Y_{n-1}, \ldots, Y_1)$ for each n and find the limit.
- Since the convergence can be arbitrarily slow, we will never know how close we are to the limit.

A Lower Bound for the Entropy

Lemma

$H(Y_n|Y_{n-1},\ldots,Y_2,X_1)\le H(\mathcal{Y})$.

- We have for $k=1,2,\ldots$,

$$\begin{aligned}
&H(Y_n|Y_{n-1},\ldots,Y_2,X_1)\\
&=H(Y_n|Y_{n-1},\ldots,Y_2,Y_1,X_1)\\
&\qquad\text{(since } Y_1 \text{ is a function of } X_1\text{)}\\
&=H(Y_n|Y_{n-1},\ldots,Y_1,X_1,X_0,X_{-1},\ldots,X_{-k})\\
&\qquad\text{(by the Markovity of } X\text{)}\\
&=H(Y_n|Y_{n-1},\ldots,Y_1,X_1,X_0,X_{-1},\ldots,X_{-k},Y_0,\ldots,Y_{-k})\\
&\qquad\text{(since } Y_i \text{ is a function of } X_i\text{)}\\
&\le H(Y_n|Y_{n-1},\ldots,Y_1,Y_0,\ldots,Y_{-k})\\
&\qquad\text{(since conditioning reduces entropy)}\\
&=H(Y_{n+k+1}|Y_{n+k},\ldots,Y_1).\quad\text{(by stationarity)}
\end{aligned}$$

A Lower Bound for the Entropy (Cont'd)

- We found

$$H(Y_n|Y_{n-1},\dots,Y_2,X_1)\le H(Y_{n+k+1}|Y_{n+k},\dots,Y_1).$$

This inequality is true for all k.

So it is true in the limit.

Thus, we obtain

$$\begin{array}{rcl} H(Y_n|Y_{n-1},\dots,Y_1,X_1) & \le & \lim_k H(Y_{n+k+1}|Y_{n+k},\dots,Y_1) \\ & = & H(\mathcal{Y}). \end{array}$$

The Interval Between the Upper and the Lower Bounds

Lemma

$H(Y_n|Y_{n-1},\ldots,Y_1)-H(Y_n|Y_{n-1},\ldots,Y_1,X_1)\to 0.$

- The interval length can be rewritten as

$$\begin{aligned}H(Y_n|Y_{n-1},\ldots,Y_1)-H(Y_n|Y_{n-1},\ldots,Y_1,X_1)\\=I(X_1;Y_n|Y_{n-1},\ldots,Y_1).\end{aligned}$$

By the properties of mutual information;

  - $I(X_1;Y_1,Y_2,\ldots,Y_n)\leq H(X_1)$;
  - $I(X_1;Y_1,Y_2,\ldots,Y_n)$ increases with n.

Thus, $\lim I(X_1;Y_1,Y_2,\ldots,Y_n)$ exists and

$$\lim_{n\to\infty} I(X_1;Y_1,Y_2,\ldots,Y_n)\leq H(X_1).$$

The Upper and the Lower Bounds (Cont'd)

- By the Chain Rule, we get

$$\begin{aligned} H(X_1) &\geq \lim_{n\to\infty} I(X_1; Y_1, Y_2, \ldots, Y_n) \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} I(X_1; Y_i | Y_{i-1}, \ldots, Y_1) \\ &= \sum_{i=1}^{\infty} I(X_1; Y_i | Y_{i-1}, \ldots, Y_1). \end{aligned}$$

This infinite sum is finite and the terms are nonnegative.

Hence, the terms must tend to 0.

I.e.,

$$\lim I(X_1; Y_n | Y_{n-1}, \ldots, Y_1) = 0.$$

Bounds for the Entropy

Theorem

If $X_1, X_2, \ldots, X_n$ form a stationary Markov chain, and $Y_i = \phi(X_i)$, then

$$H(Y_n|Y_{n-1}, \ldots, Y_1, X_1) \leq H(\mathcal{Y}) \leq H(Y_n|Y_{n-1}, \ldots, Y_1)$$

and

$$\lim H(Y_n|Y_{n-1}, \ldots, Y_1, X_1) = H(\mathcal{Y}) = \lim H(Y_n|Y_{n-1}, \ldots, Y_1).$$

- By the preceding two lemmas.

Hidden Markov Models

- In general, we could also consider the case where Y_i is a stochastic function (as opposed to a deterministic function) of X_i.
- Consider a Markov process $X_1, X_2, \ldots, X_n$.
- Define a process $Y_1, Y_2, \ldots, Y_n$, where each Y_i is drawn according to $p(y_i|x_i)$, conditionally independent of all the other $X_j, j \neq i$.
- That is,

$$p(x_n, y_n) = p(x_1) \prod_{i=1}^{n-1} p(x_{i+1}|x_i) \prod_{i=1}^{n} p(y_i|x_i).$$

- Such a process, called a **hidden Markov model** (**HMM**).
- The same argument used above for functions of a Markov chain carry over to hidden Markov models.
- So we can lower bound the entropy rate of an HMM by conditioning it on the underlying Markov state.