Entropy Rates of a Stochastic Process

Best Achievable Data Compression

Radu Trîmbiţaş

October 2012

1 Entropy Rates of a Stochastic Process

Entropy rates

- The AEP states that $nH(X)$ bits suffice on the average for n i.i.d. RVs
- What for dependent RVs?
- For stationary processes $H(X_1, X_2, \ldots, X_n)$ grows (asymptotically) linearly with n at a rate $H(\mathcal{X})$ – the *entropy rate* of the process
- A *stochastic process* $\{X_i\}_{i \in I}$ is an indexed sequence of random variables, $X_i : S \to \mathcal{X}$ is a RV $\forall i \in I$
- If $I \subseteq \mathbb{N}$, $\{X_1, X_2, \ldots\}$ is a *discrete stochastic process*, called also a *discrete information source*.
- A discrete stochastic process is characterized by the joint probability mass function

$$P((X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n)) = p(x_1, x_2, \ldots, x_n)$$

where $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$.

1.1 Markov chains

Markov chains

Definition 1. 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

$$P(X_1 = x_1, \ldots, X_n = x_n) = P(X_{1+\ell} = x_1, \ldots, X_{n+\ell} = x_n) \tag{1}$$

$\forall n, \ell$ and $\forall x_1, x_2, \ldots, x_n \in \mathcal{X}$.

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

$$\begin{aligned} &P\left(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_1 = x_1\right) \\ &= P(X_{n+1} = x_{n+1} | X_n = x_n), \qquad x_1, x_2, \dots, x_n, x_{n+1} \in \mathcal{X}. \end{aligned} \tag{2}$$

The joint pmf can be written as

$$p(x_1, x_2, \dots, x_n) = p(x_1)p(x_2|x_1)\dots p(x_n|x_{n-1}). \tag{3}$$

Definition 3. A Markov chain is said to be *time invariant* (*time homogeneous*) if the conditional probability $p(x_{n+1}|x_n)$ does not depend on n; that is for $n = 1, 2, \dots$

$$P\left(X_{n+1} = b | X_n = a\right) = P(X_2 = b | X_1 = a), \quad \forall a, b \in \mathcal{X}. \tag{4}$$

This property is assumed unless otherwise stated.

- $\{X_i\}$ Markov chain, X_n is called the *state* at time n
- A time-invariant Markov chain is characterized by its initial state and a *probability transition matrix* $P = [P_{ij}]$, $i, j = 1, \dots, m$, where $P_{ij} = P(X_{n+1} = j | X_n = i)$.
- The Markov chain $\{X_n\}$ is *irreducible* if it is possible to go from any state to another with a probability > 0
- The Markov chain $\{X_n\}$ is *aperiodic* if $\forall$ state a, the possible times to go from a to a have highest common factor = 1.
- Markov chains are often described by a directed graph where the edges are labeled by the probability of going from one state to another.
- $p(x_n)$ - pmf of the random variable at time n

$$p(x_{n+1}) = \sum_{x_n} p(x_n) P_{x_n x_{n+1}} \tag{5}$$

- A distribution on the states such that the distribution at time $n + 1$ is the same as the distribution at time n is called a *stationary distribution* - so called because if the initial state of a Markov chain is drawn according to a stationary distribution, the Markov chain form a stationary process.
- If the finite-state Markov chain is irreducible and periodic, the stationary distribution is unique, and from any starting distribution, the distribution of X_n tends to a stationary distribution as $n \to \infty$.

Example 4. Consider a two-state Markov chain with a probability transition matrix

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

(Figure 1)

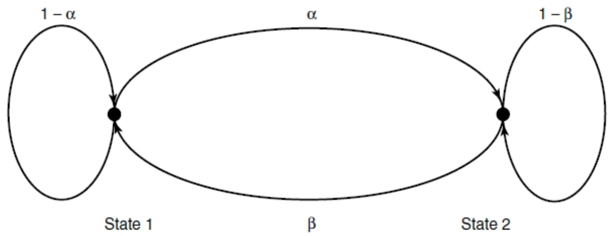

Figure 1: Two-state Markov chain

The stationary probability is the solution of $\mu P = \mu$ or $(I - P^T)\mu^T = 0$. We add the condition $\mu_1 + \mu_2 = 0$.

The solution is

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

Click here for a Maple solution `Markovex1.html`. The entropy of X_n is

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

1.2 Entropy rate

Entropy rate

Definition 5. The *entropy rate* of a stochastic process $\{X_i\}$ is defined by

$$H(\mathcal{X}) = \lim_{n \to \infty} \frac{1}{n} H\left(X_1, \ldots, X_n\right) \tag{6}$$

when the limit exists.

Examples

1. Typewriter - m equally likely output letters; he(she) can produce m^n sequences of length n, all of them equally likely. $H\left(X_1, \ldots, X_n\right) = \log m^n$, and the entropy rate is $H(\mathcal{X}) = \log m$ bits per symbol.

2. $X_1, X_2, \ldots$ i.i.d. RVs

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

3. $X_1, X_2, \ldots$ independent, but not identically distributed RVs

$$H\left(X_1, \ldots, X_n\right) = \sum_{i=1}^{n} H(X_i)$$

It is possible that $\frac{1}{n} \sum H(x_i)$ does not exists

Definition 6.
$$H'(\mathcal{X}) = \lim_{n\to\infty} H\left(X_n|X_{n-1}, X_{n-2}, \dots X_1\right). \tag{7}$$

$H(\mathcal{X})$ is entropy per symbol of the n RVs; $H'(\mathcal{X})$ is the conditional entropy of the last RV given the past.

For stationary processes both limits exist and are equal.

Lemma 7. *For a stationary stochastic process, $H(X_n|X_{n-1}, \dots, X_1)$ is nonincreasing in n and has a limit $H'(\mathcal{X})$.*

Proof.
$$\begin{aligned} H(X_{n+1}|X_1, X_2, \dots, X_n) &\le H(X_{n+1}|X_n, \dots, X_2) && \text{conditioning} \\ &= H(X_n|X_{n-1}, \dots, X_1). && \text{stationarity} \end{aligned}$$

$(H(X_n|X_{n-1}, \dots, X_1))_n$ is decreasing and nonnegative, so it has a limit $H'(\mathcal{X})$. ☐

Lemma 8 (Cesáro). *If $a_n \to a$ and $b_n = \frac{1}{n}\sum_{i=1}^{n} a_i$ then $b_n \to a$.*

Theorem 9. *For a stationary stochastic process $H(\mathcal{X})$ (given by (6)) and $H'(\mathcal{X})$ (given by (7)) exist and*
$$H(\mathcal{X}) = \mathcal{H}'(\mathcal{X}). \tag{8}$$

Proof. By the chain rule,
$$\frac{H\left(X_1, \dots, X_n\right)}{n} = \frac{1}{n}\sum_{i=1}^{n} H(X_i|X_{i-1}, \dots, X_1).$$

But,
$$\begin{aligned} H(\mathcal{X}) &= \lim_{n\to\infty} \frac{H\left(X_1, \dots, X_n\right)}{n} \\ &= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} H(X_i|X_{i-1}, \dots, X_1) \\ &= \lim_{n\to\infty} H(X_n|X_{n-1}, \dots, X_1) && \text{(Lemma 8)} \\ &= H'(\mathcal{X}) && \text{(Lemma 7)} \end{aligned}$$

☐

1.3 Entropy rate for Markov chain

Entropy rate for Markov chain

- For a stationary Markov chain, the entropy rate is given by
$$\begin{aligned} H\left(\mathcal{X}\right) = H'\left(\mathcal{X}\right) = \lim H\left(X_n|X_{n-1}, \dots, X_1\right) &= \lim H\left(X_n|X_{n-1}\right) \\ = H(X_2|X_1), \end{aligned} \tag{9}$$
  where the conditional entropy is calculated using the given stationary distribution.

- The stationary distribution μ is the solution of the equations

$$\mu_j = \sum_i \mu_i P_{ij},\ \forall j.$$

- Expression of conditional entropy:

Theorem 10. *$\{X_i\}$ stationary Markov chain with stationary distribution μ and transition matrix P. Let $X_1 \sim \mu$. then the entropy rate is*

$$H(\mathcal{X}) = -\sum_i \sum_j \mu_i P_{ij} \log P_{ij}. \tag{10}$$

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

Example 11 (Two-state Markov chain). The entropy rate of the two-state Markov chain in Figure 1 is

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

Remark. If the Markov chain is irreducible and aperiodic, it has a unique stationary distribution on the states, and any initial distribution tends to the stationary distribution as $n \to \infty$. In this case, even though the initial distribution is not the stationary distribution, the entropy rate, which is defined in terms of long-term behavior, is $H(\mathcal{X})$, as defined in (9) and (10).

1.4 Functions of Markov chains

Functions of Markov chains

- $X_1, X_2, \ldots, X_n, \ldots$ stationary Markov chain, $Y_i = \phi(X_i)$, $H(\mathcal{Y}) =?$
- in many cases $Y_1, Y_2, \ldots, Y_n, \ldots$ is not a Markov chain, but it is stationary
- lower bound

Lemma 12.

$$H(Y_n|Y_{n-1}, \ldots, Y_2, X_1) \le H(\mathcal{Y}). \tag{11}$$

Proof. For $k = 1, 2, \ldots$

$$\begin{aligned}
&H(Y_n|Y_{n-1}, \ldots, Y_2, X_1) \overset{(a)}{=} H(Y_n|Y_{n-1}, \ldots, Y_2, Y_1, X_1) \\
&\overset{(b)}{=} H(Y_n|Y_{n-1}, \ldots, Y_2, Y_1, X_1, X_0, X_{-1}, \ldots, X_{-k}) \\
&\overset{(c)}{=} H\left(Y_n|Y_{n-1}, \ldots, Y_2, Y_1, X_1, X_0, X_{-1}, \ldots,\right. \\
&\left.X_{-k}, Y_0, \ldots, Y_{-k}\right) \\
&\overset{(d)}{\le} H\left(Y_n|Y_{n-1}, \ldots, Y_2, Y_1, Y_0, \ldots, Y_{-k}\right) \\
&\overset{(e)}{=} H\left(Y_{n+k+1}|Y_{n+k}, \ldots, Y_1\right),
\end{aligned}$$

(a) follows from the fact that $Y_1 = \phi(X_1)$, (b) from the Markovity, (c) from $Y_i = \phi(X_i)$, (d) conditioning reduces entropy, (e) stationarity. □

Proof - continuation. Since inequality is true for all k, in the limit

$$\begin{aligned} H(Y_n|Y_{n-1},\dots,Y_2,X_1) &\le \lim_k H\left(Y_{n+k+1}|Y_{n+k},\dots,Y_1\right) \\ &= H(\mathcal{Y}). \end{aligned}$$

□

Lemma 13.

$$H(Y_n|Y_{n-1},\dots,Y_2,X_1) - H(Y_n|Y_{n-1},\dots,Y_2,Y_1,X_1) \to 0. \tag{12}$$

Proof. Expression of interval length:

$$\begin{aligned} &H(Y_n|Y_{n-1},\dots,Y_2,X_1) - H(Y_n|Y_{n-1},\dots,Y_2,Y_1,X_1) \\ &= I(X_1;Y_n|Y_{n-1},\dots,Y_1). \end{aligned}$$

By properties of mutual information,

$$I(X_1;Y_1,\dots,Y_n) \le H(X_1),$$

and $I(X_1;Y_1,\dots,Y_n)$ increases with n. Thus, $\lim I(X_1;Y_1,\dots,Y_n)$ exists and

$$\lim_{n\to\infty} I(X_1;Y_1,\dots,Y_n) \le H(X_1).$$

□

Proof - continuation. By the chain rule

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

The general term of the series must tend to 0

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

□

The last two lemmas imply

Theorem 14. *$X_1, X_2, \dots, X_n, \dots$ stationary Markov chain, $Y_i = \phi(X_i)$*

$$H(Y_n|Y_{n-1},\dots,Y_1,X_1) \le H(\mathcal{Y}) \le H(Y_n|Y_{n-1},\dots,Y_1) \tag{13}$$

and

$$\lim H(Y_n|Y_{n-1},\dots,Y_1,X_1) = H(\mathcal{Y}) = \lim H(Y_n|Y_{n-1},\dots,Y_1) \tag{14}$$

Hiden Markov models

- We could consider Y_i to be a stochastic function of X_i
- $X_1, X_2, \ldots, X_n, \ldots$ stationary Markov chain, $Y_1, Y_2, \ldots, Y_n, \ldots$ a new process where Y_i is drawn according to $p(y_i|x_i)$, conditionally independent of all the other X_j, $j \neq i$

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

- $Y_1, Y_2, \ldots, Y_n, \ldots$ is called a hidden Markov model (HMM)
- Applied to speech recognition, handwriting recognition, and so on.
- The same argument as for functions of Markov chain works for HMMs.

References

References

[1] Thomas M. Cover, Joy A. Thomas, *Elements of Information Theory*, 2nd edition, Wiley, 2006.

[2] David J.C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003.

[3] Robert M. Gray, *Entropy and Information Theory*, Springer, 2009