Stochastic processes

A stochastic (or random) process \( \{X_i\} \) is an indexed sequence of random variables. The dependencies among the random variables can be arbitrary. The process is characterized by the joint probability mass functions

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

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

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, ie

\[
\Pr\{X_1 = x_1, \ldots, X_n = x_n\} = \Pr\{X_{1+l} = x_1, \ldots, X_{n+l} = x_n\}
\]

for every \( n \) and shift \( l \) and for all \( x_1, \ldots, x_n \in \mathcal{X} \).
Markov chains

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

$$Pr\{X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_1 = x_1\} = Pr\{X_{n+1} = x_{n+1} | X_n = x_n\}$$

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

For a Markov chain we can write the joint probability mass function as

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

A Markov chain is said to be *time invariant* if the conditional probability $p(x_{n+1} | x_n)$ does not depend on $n$, ie

$$Pr\{X_{n+1} = b | X_n = a\} = Pr\{X_2 = b | X_1 = a\}$$

for all $a, b \in \mathcal{X}$ and $n = 1, 2, \ldots$. 
Markov chains, cont.

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 and a probability transition matrix \( P = [P_{ij}] \), where \( P_{ij} = Pr\{X_{n+1} = j|X_n = i\} \) (assuming that \( \mathcal{X} = \{1, 2, \ldots, m\} \)).

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, the Markov chain is said to be irreducible. If the largest common factor of the lengths of different paths from a state to itself is 1, the Markov chain is said to be aperiodic.

If the probability mass function 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 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*. The stationary distribution is so called because if the initial state of the distribution is drawn according to a stationary distribution, the Markov chain forms a stationary process.

If a finite-state Markov chain is irreducible and aperiodic, the stationary distribution is unique, and from any starting distribution, the distribution of $X_n$ tends to the stationary distribution as $n \to \infty$. 
Entropy rate

The entropy rate (or just 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. It can also be defined as

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

when the limit exists.

For stationary processes, both limits exist and are equal

\[
H(\mathcal{X}) = H'(\mathcal{X})
\]

For a stationary Markov chain, we have

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

\[
= \lim_{n \to \infty} H(X_n | X_{n-1}) = H(X_2 | X_1)
\]
For a Markov chain, the stationary distribution $\mu$ is the solution to the equation system

$$\mu_j = \sum_i \mu_i P_{ij}, \quad j = 1, \ldots, m$$

Given a stationary Markov chain $\{X_i\}$ with stationary distribution $\mu$ and transition matrix $P$. Let $X_1 \sim \mu$. The entropy rate of the Markov chain is then given by

$$H(X) = -\sum_{ij} \mu_i P_{ij} \log P_{ij} = \sum_i \mu_i (-\sum_j P_{ij} \log P_{ij})$$

ie a weighted sum of the entropies for each state.
Functions of Markov chains

Let $X_1, X_2, \ldots$ be a stationary markov chain and let $Y_i = \phi(X_i)$ be a process where each term is a function of the corresponding state in the Markov chain. What is the entropy rate $H(Y)$? If $Y_i$ also forms a Markov chain this is easy to calculate, but this is not generally the case. We can however find upper and lower bounds.

Since the Markov chain is stationary, so is $Y_1, Y_2, \ldots$. Thus, the entropy rate $H(Y)$ is well defined. We might try to compute $H(Y_n|Y_{n-1} \ldots Y_1)$ for each $n$ to find the entropy rate, however the convergence might be arbitrarily slow.

We already know that $H(Y_n|Y_{n-1} \ldots Y_1)$ is an upper bound on $H(Y)$, since it converges monotonically to the entropy rate. For a lower bound, we can use $H(Y_n|Y_{n-1}, \ldots, Y_1, X_1)$. 
Functions of Markov chains, cont.

We have
\[ H(Y_n|Y_{n-1}, \ldots, Y_2, X_1) \leq H(Y) \]

Proof: For \( k = 1, 2, \ldots \) we have
\[
H(Y_n|Y_{n-1}, \ldots, Y_2, X_1) = H(Y_n|Y_{n-1}, \ldots, Y_2, Y_1, X_1) \\
= H(Y_n|Y_{n-1}, \ldots, Y_2, Y_1, X_1, X_0, \ldots, X_{-k}) \\
= H(Y_n|Y_{n-1}, \ldots, Y_2, Y_1, X_1, X_0, \ldots, X_{-k}, Y_0, \ldots, Y_{-k}) \\
\leq H(Y_n|Y_{n-1}, \ldots, Y_2, Y_1, Y_0, \ldots, Y_{-k}) \\
= H(Y_{n+k+1}|Y_{n+k}, \ldots, Y_2, Y_1)
\]

Since this is true for all \( k \) it will be true in the limit too. Thus
\[
H(Y_n|Y_{n-1}, \ldots, Y_1, X_1) \leq \lim_{k \to \infty} H(Y_{n+k+1}|Y_{n+k}, \ldots, Y_2, Y_1) = H(Y)
\]
The interval between the upper and lower bounds tends to 0, ie

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

Proof: The difference can be rewritten as

\[ H(Y_n|Y_{n-1}, \ldots, Y_1) - H(Y_n|Y_{n-1}, \ldots, Y_1, X_1) = I(X_1; Y_n|Y_{n-1}, \ldots, Y_1) \]

By the properties of mutual information

\[ I(X_1; Y_1, Y_2, \ldots, Y_n) \leq H(X_1) \]

Since \( I(X_1; Y_1, Y_2, \ldots, Y_n) \) increases with \( n \), the limit exists and

\[ \lim_{n \to \infty} I(X_1; Y_1, Y_2, \ldots, Y_n) \leq H(X_1) \]
The chain rule for mutual information gives us

\[ H(X_1) \geq \lim_{n \to \infty} I(X_1; Y_1, Y_2, \ldots, Y_n) \]

\[ = \lim_{n \to \infty} \sum_{i=1}^{n} I(X_1; Y_i|Y_{i-1}, \ldots, Y_1) \]

\[ = \sum_{i=1}^{\infty} I(X_1; Y_i|Y_{i-1}, \ldots, Y_1) \]

Since this infinite sum is finite and all terms are non-negative, the terms must tend to 0, which means that

\[ \lim_{n \to \infty} I(X_1; Y_n|Y_{n-1}, \ldots, Y_1) = 0 \]
Functions of Markov chains, cont.

We have thus shown that

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

and that

$$\lim_{n \to \infty} H(Y_n|Y_{n-1}, \ldots, Y_1, X_1) = H(Y) = \lim_{n \to \infty} H(Y_n|Y_{n-1}, \ldots, Y_1)$$

This result can also be generalized to the case when $Y_i$ is a random function of $X_i$, called a hidden Markov model.
Higher order Markov sources

A Markov chain is a process where the dependency of an outcome at time $n$ only reaches back one step in time

$$p(x_n| x_{n-1}, x_{n-2}, x_{n-3}, \ldots) = p(x_n|x_{n-1})$$

In some applications, particularly in data compression, we might want to have source models where the dependency reaches back longer in time. A Markov source of order $k$ is a random process where the following holds

$$p(x_n| x_{n-1}, x_{n-2}, x_{n-3}, \ldots) = p(x_n|x_{n-1}, \ldots, x_{n-k})$$

A Markov source of order 0 is a memoryless process. A Markov source of order 1 is a Markov chain. Given $X_n$, a stationary Markov source of order $k$. Form a new source $S_n = (X_{n-k+1}, X_{n-k+2}, \ldots, X_n)$. This new process is a Markov chain. The entropy rate is

$$H(S_n| S_{n-1}) = H(X_n| X_{n-1}, \ldots, X_{n-k})$$
Differential entropy

A continuous random variable $X$ has the probability density function $f(x)$. Let the support set $S$ be the set of $x$ where $f(x) > 0$. The differential entropy $h(X)$ of the variable is defined as

$$h(X) = -\int_S f(x) \log f(x) \, dx$$

Unlike the entropy for a discrete variable, the differential entropy can be both positive and negative.
Some common distributions

Normal distribution (gaussian distribution)

\[ f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-m)^2}{2\sigma^2}}, \quad h(X) = \frac{1}{2} \log 2\pi e\sigma^2 \]

Laplace distribution

\[ f(x) = \frac{1}{\sqrt{2\sigma}} e^{-\frac{\sqrt{2}|x-m|}{\sigma}}, \quad h(X) = \frac{1}{2} \log 2e^2\sigma^2 \]

Uniform distribution

\[ f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}, \quad h(X) = \log(b - a) = \frac{1}{2} \log 12\sigma^2 \]
AEP for continuous variables

Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. random variables drawn according to the density $f(x)$. Then

$$-\frac{1}{n} \log f(X_1, X_2, \ldots, X_n) \to E[-\log f(X)] = h(X)$$

in probability.

For $\epsilon > 0$ and any $n$ we define the typical set $A_{\epsilon}^{(n)}$ with respect to $f(x)$ as follows

$$A_{\epsilon}^{(n)} = \{(x_1, \ldots, x_n) \in S^n : \left| -\frac{1}{n} \log f(x_1, \ldots, x_n) - h(X) \right| \leq \epsilon \}$$

The volume $\text{Vol}(A)$ of any set $A$ is defined as

$$\text{Vol}(A) = \int_A d_{x_1} d_{x_2} \ldots d_{x_n}$$
The typical set $A_\epsilon^{(n)}$ has the following properties

1. $Pr\{A_\epsilon^{(n)}\} > 1 - \epsilon$ for sufficiently large $n$
2. $\text{Vol}(A_\epsilon^{(n)}) \leq 2^{n(h(X)+\epsilon)}$ for all $n$
3. $\text{Vol}(A_\epsilon^{(n)}) \geq (1 - \epsilon)2^{n(h(X)-\epsilon)}$ for sufficiently large $n$

This indicates that the typical set contains almost all probability and has a volume of approximately $2^{nh(X)}$. 
Quantization

Suppose we do uniform quantization of a continuous random variable $X$, i.e., we divide the range of $X$ into bins of length $\Delta$. Assuming that $f(x)$ is continuous in each bin, there exists a value $x_i$ within each bin such that

$$f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)dx$$

Consider the quantized variable $X^\Delta$ defined by

$$X^\Delta = x_i \text{ if } i\Delta \leq X < (i+1)\Delta$$

The probability $p(x_i) = p_i$ that $X^\Delta = x_i$ is

$$p_i = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = f(x_i)\Delta$$
The entropy of the quantized variable is

\[ H(X^\Delta) = - \sum_i p_i \log p_i \]

\[ = - \sum_i \Delta f(x_i) \log(\Delta f(x_i)) \]

\[ = - \sum_i \Delta f(x_i) \log f(x_i) - \sum_i \Delta f(x_i) \log \Delta \]

\[ \approx - \int_{-\infty}^{\infty} f(x) \log f(x) \, dx - \log \Delta \]

\[ = h(X) - \log \Delta \]
Differential entropy, cont.

Two random variables $X$ and $Y$ with joint density function $f(x, y)$ and conditional density functions $f(x|y)$ and $f(y|x)$. The joint differential entropy is defined as

$$h(X, Y) = -\int f(x, y) \log f(x, y) \, dxdy$$

The conditional differential entropy is defined as

$$h(X|Y) = -\int f(x, y) \log f(x|y) \, dxdy$$

We have

$$h(X, Y) = h(X) + h(Y|X) = h(Y) + h(X|Y)$$

which can be generalized to (chain rule)

$$h(X_1, X_2, \ldots, X_n) = h(X_1) + h(X_2|X_1) + \ldots + h(X_n|X_1, X_2, \ldots, X_{n-1})$$
Relative entropy

The relative entropy (Kullback-Leibler distance) between two densities $f$ and $g$ is defined by

$$D(f||g) = \int f \log \frac{f}{g}$$

The relative entropy is finite only if the support set of $f$ is contained in the support set for $g$. 
Mutual information

The mutual information between $X$ and $Y$ is defined as

$$I(X; Y) = \int f(x, y) \log \frac{f(x, y)}{f(x)f(y)} \, dx\,dy$$

which gives

$$I(X; Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) = h(X) + h(Y) - h(X, Y)$$

and

$$I(X; Y) = D(f(x, y)||f(x)f(y))$$

Given two uniformly quantized versions of $X$ and $Y$

$$I(X^\Delta; Y^\Delta) = H(X^\Delta) - H(X^\Delta|Y^\Delta)$$

$$\approx h(X) - \log \Delta - (h(X|Y) - \log \Delta)$$

$$= I(X; Y)$$
Properties

\[ D(f \| g) \geq 0 \]
with equality iff \( f \) and \( g \) are equal almost everywhere.

\[ I(X; Y) \geq 0 \]
with equality iff \( X \) and \( Y \) are independent.

\[ h(X \mid Y) \leq h(X) \]
with equality iff \( X \) and \( Y \) are independent.

\[ h(X + c) = h(X) \]
\[ h(aX) = h(X) + \log |a| \]
\[ h(AX) = h(X) + \log |\det(A)| \]
Differential entropy, cont.

The gaussian distribution is the distribution that maximizes the differential entropy, for a given covariance matrix.

I.e., for the one-dimensional case, the differential entropy for a variable $X$ with variance $\sigma^2$ satisfies the inequality

$$h(X) \leq \frac{1}{2} \log 2\pi e\sigma^2$$

with equality iff $X$ is gaussian.

For the general case, the differential entropy for a random $n$-dimensional vector $X$ with covariance matrix $K$ satisfies the inequality

$$h(X) \leq \frac{1}{2} \log(2\pi e)^n |K|$$

with equality iff $X$ is a multivariate gaussian vector.