Random variables in Communication network

1. Dependent Random variables in Communication

We use Random vectors with independent components can be used to represent information in Communication network. However, in real world, it is not so as each component may be correlated to the other in some way and if we know this correlation we can perform better compression.To understand the correlation, we must know the models of dependence. Another application of dependent random variable is when a sequence of bits is transmitted through a noisy channel, then the sequence received at receiver, in the form of a random vector, depends on the transmitted random vector. If they were independent, then we could never be able to decode the received back to the original message with low probability.

1.1. Forward and inverse probability

Probability calculations fall into two categories: *forward probability* and *inverse probability*. **Forward probability** involves some generative model that describes a process that results in some data, the task then becomes to compute the probability distribution or expectation of some quantity that depends on the data. Entropy calculation falls into this category.
Inverse probability problems also involve some generative model of a process (In this discussion, this process is communication), but here instead of calculating the probability distribution of some quantity as a result of the process, we calculate the probability of one or more unobserved variables in the process, given the observed variables. This concept will be used ahead in the modelling of a decoder to estimate the transmitted sequence, given a received sequence in communication. This involves the use of Baye's Theorem, and can be used in predictions. It can also be used in data compression.

1.1.1 Inferences in Probability

Inference in communication involves deciphering the message transmitted through a noisy channel upon receiving it at the receivers end. Inferences can be correctly made in communication networks using the Bayes' theorem.

1.2. Types of Entropy

The following Entropy will be used ahead. Entropy can be said as the expected information content carried by a Random variable.
Entropy:

$$H(X) = \sum_{x \in Supp(P_X)} P_X(x).log\frac{1}{P_X(x)}$$

Joint Entropy:

$$H(X,Y) = \sum_{x,y \in Supp(P_{X,Y})} P_{X,Y}(x,y).log\frac{1}{P_{X,Y}(x,y)}$$

Here the entropy represents the expected information carried by X and Y simultaneously.
Conditional Entropy:

$$H(X|Y) = \sum_{y \in Supp(P_Y)} P_Y(y).H(X|Y=y)$$
$$H(X|Y=y) = \sum_{x \in Supp(P_{X|Y})} P_{X|Y}(x|y).H(x|y)$$

This represents the expected information carried by X, given that we know the information in Y. It can also be described as the uncertainty that remains in X after we know Y, as information is the increases as uncertainty increases. We can see that the conditional entropy depends on the conditional probability of X given Y. For independent X and Y:

$$H(X|Y) = H(X)$$

This can be explained by the fact that since they are independent, knowing Y we still cannot say anything about X.
Information Content and chain rule: Information content of X is represented as:

$$\frac{1}{p_X(X=x_i)}$$

And chain rules of probability states that for two random variables:

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

Using this in the expression for entropy, we can derive the chain rule for Entropy:

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

Mutual induction:

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

This is used to represent the reduction in information content after we have received Y, or vice versa. This concept is used to define the **channel capacity** or the maximum amount of information that can be sent through a channel. This can also express the average information content expressed by one about the other.

1.3. Gibbs Inequality

Also known as relative entropy or Kullback-Leibler divergence between two probability distribution $P(x)$ and $Q(x)$ defined over the same alphabet A_X, it is a very important concept in information theory. It can loosely be used to define the "distance" between two probability distributions. However it is not strictly distance. It is not symmetric, that is, relative entropy between P and Q is not the same as the relative entropy between Q and P.

$$D(P||Q) = \sum_x P(x).log\frac{P(x)}{Q(x)}$$

2. Communication over a noisy Channel

Channels, or the medium through which information is transmitted, are usually noisy. The need arises develop methods for error-free communication.For this purpose, we will elaborate on noisy error channel and channel coding. Channel coding is used to make the noisy channel behave close to a noiseless channel. The Channel Code is made such that noisy signal received can be decoded. As we have stated before, the measure of the information transmitted can be expressed by the mutual information between the transmitted and received signal.

2.1. Noisy Channels

Conditional probability between the transmitted signal and the received signal can be used to characterize the noisy channel.

Discrete Noiseless Channel(Q): It is characterized by an input alphabet A_X an output alphabet A_Y (Alphabet is the set of values that the message can be mapped to) , and a set of conditional probability distributions $P(y|x)$, one for each $x \in A_X$.

$$Q_{j|i} = P(y = b_j|x = a_i)$$

We can make this into a matrix such that each column is a probability vector, and then obtain the P_Y probability vector by multiplying Q matrix with P_X probability vector. Some models of noisy channel are: **Binary Symmetric Channel:** A_X=0,1. A_Y=0,1

$$P(y = 0|x = 0) = 1 - p;\ P(y = 1|x = 0) = p$$
$$P(y = 0|x = 1) = p;\ P(y = 1|x = 1) = 1 - p$$

Here p is said to be the probability of the transmitted bit to be flipped.

Binary Erasure Channel: A_X=0,1. $A_Y = 0, \epsilon, 1$

$$P(y = 0|x = 0) = 1 - p;\ P(y = 1|x = 0) = 0$$
$$P(y = \epsilon|x = 1) = p;\ P(y = \epsilon|x = 1) = p$$
$$P(y = 1|x = 0) = 0;\ P(y = 1|x = 1) = 1 - p$$

Here p is said to be the probability of the transmitted bit getting erased. ϵ represent the erasure of the bit. The bit is not flipped here, but erased with some probability. The probability of getting erased or flipped in the above noisy channels is conditionally dependent on the transmitted bit.

2.2. Estimating the input given the output

To estimate the transmitted symbol x from the received signal y, we can use the **Bayes' theorem** given y:

$$\begin{aligned} P(x|y) &= \frac{P(y|x).P(x)}{P(y)} \\ &= \frac{P(y|x).P(x)}{\sum_{x_i \in X} P(y|x_i)P(x_i)} \end{aligned}$$

A decent estimate here is: $\underset{x}{argmax}(P(x|y))$

2.3. Information conveyed by a Channel

To measure the amount of information the output conveys about particular input X, we use mutual Information:

$$I(X;Y) \equiv H(X) - H(Y|X)$$

Note that:$I(X;Y) \leq min(H(Y), H(X))$. Intuitively it means that no more than the conveyed information can be received. **Maximising mutual information**

Maximising mutual information conveyed by the channel would mean to maximise the amount of information transferred in communication. We define the channel capacity of a channel Q to be the maximum mutual information over P_X

$$C(Q) = \underset{P_X}{max} I(X;Y)$$

We have control over P_X and can maximising it by optimising the input coding. P_X at channel capacity is called *optimal input distribution*. The rate can be increased only up to the channel capacity if we want to have arbitrarily small probability of error. This is the converse of the **Shannon's Channel Coding theorem.**

2.4. Optimal Decoder for noisy channel coding

Since the transmitted sequence or codeword might have some error due to it being transmitted though a noisy channel. As a result, the received codeword is not the same as the transmitted one. Therefore, a need arises to infer the transmitted message from the received codeword, to enable communication. This is done by the use of a decoder. For a channel, an optimal decoder is one which minimises the probability of error. It decodes an output y as the input x that has maximum posterior probability P(x—y)

$$\begin{aligned} P(x|y) &= \frac{P(y|x).P(x)}{\sum_{x_i \in X} P(y|x_i)P(x_i)} \\ \hat{x}_{optimal} &= argmax P(x|y) \end{aligned}$$

If the prior distribution on x is uniform, then the optimal decoder is called *maximum likelihood decoder* i.e., the estimation is such that the output is such that P(y—x) has the maximum likelihood.

2.5. Gaussian Random variable to model Noise

Noise can also be modelled using Gaussian random variable. Noise can be defined as the sum of infinitely many independent random variable, as there may be several disturbances in the channel. According to *Central Limit theorem*, sum of infinitely many independent random variable results is a Gaussian random variable. The resultant sum can be said as the random variable representing the noise in the channel.

3. Entropy rates of a Stochastic Process

A stochastic process, also called a random process, is a set of random variables that model a non deterministic system. Many time-varying signals are random in nature, such as noises, image and audio: usually unknown to the distant receiver. Random process (or stochastic process) presents the mathematical model of these random signals. If the random variables are dependent or in particular, if the random variables form a stationary process, we will show, just as in the i.i.d.case, that the entropy $H(X1, X2, \ldots, Xn)$ grows (asymptotically) linearly with n at a rate $H(X)$, which we will call the entropy rate of the process.

An indexed sequence of random variables is called a stochastic process X_i. There can be arbitrary dependencies among random variables in general. The joint probability mass functions characterise the process.
$Pr\{(X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n)\} = p(x_1, x_2, \ldots, x_n), (x_1, x_2, \ldots, x_n) \in X_n \quad for \quad n = 1, 2, \ldots$

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; that is,
$\mathbf{P}r\{X_1 = x_1, ..., X_n = x_n\} = Pr\{X_1 + l = x_1, ..., X_n + l = x_n\}$ for every n and every shift l and $\forall x_1, x_2, \ldots, x_n \in X$.

3.1. Markov chains

An example of a stochastic process with dependency is one in which each random variable is conditionally independent of all the other previous random variables.

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

$\mathbf{P}r(X_{n+1} = x_{n+1}|X_n = x_n, X_{n-1} = x_{n-1}, ..., X_1 = x_1) = Pr(X_{n+1} = x_{n+1}|X_n = x_n)$

$$\forall \quad x_1, x_2, ..., x_n, x_{n+1} \in X.$$

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

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

It is assumed that the Markov chains are time invariant unless otherwise stated.

Definition The Markov chain is said to be time invariant if the conditional probability $p(x_n + 1|x_n)$ does not depend on n; that is, for n = 1, 2, . . . ,

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

$$\forall \quad a, b \in X.$$

State and Transition matrix:
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, given by $P = [P_{ij}], \quad for \quad i, j \in 1, 2, \ldots, m,$
where $P_ij = Pr\{X_{n+1} = j|X_n = i\}$.

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 **aperiodic**.

Example: Consider a two-state Markov chain with a probability transition matrix given by:

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

Let's say the stationary distribution is represented as a vector μ with the stationary probabilities of states 1 and 2 as its components. Then, by solving the equation $\mu P = \mu$ or, more simply, by balancing probabilities, the stationary probability can be obtained.

$$\mu P = \mu \implies \mu P = \mu I \implies \mu(P - I) = 0$$

The net probability flow across any cut set in the state transition graph is zero for the stationary distribution.
As μ is a probability distribution it follows that $\mu_1 + \mu_2 = 1$. Solving for μ, using the above and the following equation, $\mu_1 \alpha = \mu_2 \beta$. The stationary distribution is given by:

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

The resulting process will be stationary if the Markov chain has an initial state drawn according to the stationary distribution. The entropy of the state X_n is:

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

3.2. Entropy Rate

Since a stochastic process defined by a Markov chain that is irreducible, aperiodic and positive recurrent has a stationary distribution, the entropy rate is independent of the initial distribution. is the asymptotic distribution of the chain.

Definition: The entropy of a stochastic process Xi is defined by

$$H(X) = \lim_{n \to \infty} \frac{1}{n} H(X_1, H_2, \dots H_3)$$

when the limit exists.

We can also define a related quantity for entropy rate:

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

when the limit exists.

$H(X)$ and $H'(X)$ are two different notions of entropy rate. The first is the entropy of the n random variables per symbol, and the second is the conditional entropy of the last random variable given the past.
We now show that both limits exist and are identical for stationary processes.

Entropy of a stationary Markov Chain:
For a stationary Markov chain, the entropy rate is given by

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

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

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

Proof

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

3.3. Functions of Markov chains

Let $X1, X2, \dots, Xn, \dots$ be a stationary Markov chain, and let $Y_i = \varphi(X_i)$ be a process each term of which is a function of the corresponding state in the Markov chain. To compute $H(Y)$, we might compute $H(Y_n | Y_{n-1}, ..., Y_1)$ for each n and find the limit. Upper and lower bounds converging to the limit from above and below might be advantageous computationally. When the difference between the upper and lower bounds is small, we can stop the computation and get a solid estimate of the limit.
We know that $H(Y_n | Y_{n-1}, ..., Y_2, Y_1)$ converges to $H(Y)$ i.e, $H(Y_n | Y_{n-1}, ..., Y_2, Y_1) \leq H(Y)$

The lemma shows that the interval between the upper and the lower bounds decreases in length.

Lemma

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

Proof
The LHS can also be written as:

$$H(Y_n | Y_{n-1}, ..., Y_1) - H(Y_n | Y_{n-1}, ..., Y_1, X_1)$$
$$= I(X_1; Y_n | Y_{n-1}, ..., Y_1)$$

This mutual information is always less than or equal to $H(X_1)$ i.e,

$$I(X_1; Y_1, Y_2, ..., Y_n) \leq H(X_1)$$

As n tends to infinity, $\lim I(X_1; Y_1, Y_2, ..., Y_n)$ exists.
From the chain rule,

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

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

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

As n tends to infinity, this sum of infinite terms is finite and the terms are non-negative, the terms must tend to 0; which proves the lemma, that is,

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

From this lemma, we have the following theorem:

Theorem:
If $X_1, X_2, ..., X_n$ form a stationary Markov chain, and $Y_i = \varphi(X_i)$, then

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

and

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

similarly,

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

3.4. Use of stochastic processes in communication

We are all aware that all communication systems both generate and are affected by noise. Since to implement any circuit or device whether in Analog or Digital domain, we need it's Mathematical Model, without which we cannot implement anything. It provides us with a method or metric to use in real-world circumstances. The same may be said for noise. Random process (also known as stochastic process) is an area of mathematics that can develop or apply mathematical models for noise in communication systems.

Because of its time fluctuating character, noise is modelled as a random variable, which is then modelled as a random process. Because it closely reflects the probability distribution of the noise that occurs in communication systems, Gaussian Noise is the most famous model for noise in communication systems today.

So in this way the random process or also known as Stochastic Process is useful in Communication Systems.