Entropy Rate of Stochastic Processes

Timo Mulder
tmamulder@gmail.com

Jorn Peters
jornpeters@gmail.com

February 8, 2015

The entropy rate of independent and identically distributed events can on average be encoded by $H(X)$ bits per source symbol. However, in reality, series of events (or processes) are often randomly distributed and there can be arbitrary dependence between each event. Such processes with arbitrary dependence between variables are called stochastic processes. This report shows how to calculate the entropy rate for such processes, accompanied with some definitions, proofs and brief examples.

1 Introduction

One may be interested in the uncertainty of an event or a series of events. Shannon entropy is a method to compute the uncertainty in an event. This can be used for single events or for several independent and identically distributed events. However, often events or series of events are not independent and identically distributed. In that case the events together form a stochastic process i.e. a process in which multiple outcomes are possible. These processes are omnipresent in all sorts of fields of study.

As it turns out, Shannon entropy cannot directly be used to compute the uncertainty in a stochastic process. However, it is easy to extend such that it can be used. Also when a stochastic process satisfies certain properties, for example if the stochastic process is a Markov process, it is straightforward to compute the entropy of the process. The entropy rate of a stochastic process can be used in various ways. An example of this is given in Section 4.1.

In Section 3 stochastic processes are defined and properties they can possess are discussed. Extra consideration is given to Markov processes as these have various properties that help when computing the entropy rate. In Section 4 the entropy rate for a stochastic process is discussed and defined. The findings based on all of this are reported in Section 5.

2 Preliminaries

The notation $\log(\cdot)$ will refer to the logarithm with base 2. Throughout this paper random variables (RV) are denoted using a capital letter. The set of values a RV can take on is denoted using the same calligraphic capital letter. For example, X is a RV and can take on any value in $\mathcal{X}$.

$H(X)$ is the entropy (Shannon, 2001) of a random variable X defined as

$$H(X) \triangleq -\sum_{x \in \mathcal{X}} P_X(x) \log\left(P_X(x)\right). \tag{1}$$

The entropy of the joint distribution of two or more random variables is notated as $H(X_1, \ldots, X_n)$. The joint entropy (Shannon, 2001) of $X_1, \ldots, X_n$ is defined as

$$H(X_1, \ldots, X_n) \triangleq -\sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_n \in \mathcal{X}_n} P(X_1 = x_1, \ldots, X_n = x_n) \log \big(P(X_1 = x_1, \ldots, X_n = x_n)\big). \tag{2}$$

Furthermore, the conditional entropy rate (Shannon, 2001) $H(Y \mid X)$ is defined as

$$H(Y \mid X) \triangleq \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} P(x, y) \log \left(\frac{P(x)}{P(x, y)} \right). \tag{3}$$

The binary entropy, $h(p)$, is the entropy of event an event with probability p happening. Note that $h(p) = h(1 - p)$. $h(p)$ is defined as

$$h(p) = H(p, 1 - p) = p \log(\frac{1}{p}) + (1 - p) \log(\frac{1}{1 - p}). \tag{4}$$

Finally in this paper $P(X_1 = 1)$ has preference over $P_{X_1}(1)$ or even just $P(1)$ when there is no confusion. P_X may still be used when referred to the probability distribution of X.

3 Stochastic Processes

A **stochastic process**, also called a **random process**, is a set of random variables that model a non deterministic system. In other words the outcome of the system if not known on beforehand and the system can evolve in multiple ways. The uses of stochastic processes are manifold and for example are used in the stochastic analysis of financial markets (Bachelier, 2011) or in models for the simulation of seismic motion during earthquakes (Shinozuka and Deodatis, 1988). If not cited otherwise the rest of this section follows chapter 4 of Cover and Thomas (2012).

A stochastic process $\{X_i\}$ is an indexed collection of random variables. There can be arbitrary dependence between each of the random variables. The stochastic process is characterized by the joint probability mass function

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n), \quad (x_1, x_2, , x_n) \in \mathcal{X}^n. \tag{5}$$

The indexes of the random variables can be seen as discrete time indexes, but are not necessarily time indexes.

A stochastic process is said to be stationary if the joint probability distribution of any subsequence of the sequence of random variables is invariant of shifts in time.

Definition 1 (Stationary Stochastic Process)**.** A stochastic process $\{X_i\}$ is stationary if and only if

$$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n) = P(X_{1+l} = x_1, X_{2+l} = x_2, \ldots X_{n+l} = x_n) \tag{6}$$

for every n and l and every $(x_1, \ldots, x_n) \in \mathcal{X}^n$.

More specifically this means that $P(X_i = \alpha) = P(X_j = \alpha)$ for any i, j and $\alpha \in \mathcal{X}$.

The random variables in a stochastic process can have arbitrary dependence. However, if the random variables that a random variable can depend are restricted to only its direct predecessor the stochastic process is called a **Markov chain** or **markov process**.

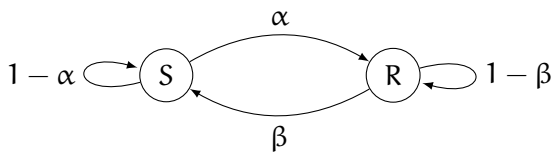


Figure 1: Example of a time invariant Markov process. Here $\mathcal{X} = \{S, R\}$. Independent of the time step the probability for the next state is only dependent on the current state. For example if today is sunny (S) tomorrow will be rainy (R) with probability α.

Definition 2 (Markov Process)**.** A stochastic process $\{X_i\}$ is a Markov process if and only if

$$\begin{aligned} &P(X_i = x_i \mid X_n = x_n, \ldots, X_{i+1} = x_{i+1}, X_{i-1} = x_{i-1}, \ldots, X_1 = x_1) \\ &\quad = P(X_i = x_i \mid X_{i-1} = x_{i-1}) \end{aligned} \tag{7}$$

For every n, i and $(x_1, \ldots, x_n) \in \mathcal{X}^n$.

It follows that in the case of a Markov process the probability mass function of can be written as

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ &\quad = P(X_1 = x_1)P(X_2 = x_2 \mid X_1 = x_1) \cdots P(X_n = x_n \mid X_{n-1} = x_{n-1}). \end{aligned} \tag{8}$$

The conditional probability function $P(X_j = \alpha \mid X_{j-1} = \beta)$ for any j and $\alpha, \beta \in \mathcal{X}$ is the transition probability of moving from state β to a state α from time step $j-1$ to time step j. If this is the same for any j the Markov process is said to be time invariant.

Definition 3 (Time Invariant Markov Process)**.** A Markov process is time invariant if and only if for every $\alpha, \beta \in \mathcal{X}$ and $n = 1, 2, \ldots$

$$P(X_n = \beta \mid X_{n-1} = \alpha) = P(X_2 = \beta \mid X_1 = \alpha) \tag{9}$$

figure 1 shows an example of the state transitions of a time invariant Markov process. If a Markov process is time invariant it can be characterized by its initial state and a *probability transition matrix* $P \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}$ where $P_{ij} = P(X_{n+1} = j \mid X_n = i)$. For example the probability transition matrix for the Markov process shown in figure 1 is

$$P = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}. \tag{10}$$

The probability transition matrix is a stochastic matrix as all rows sum to 1.

If the probability mass function at time t, P_{X_t}, is known then the probability mass function at time $t+1$ is

$$\begin{aligned} P(X_{t+1} = \alpha) &= \sum_{x \in \mathcal{X}} P(X_t = x)P(X_{t+1} = \alpha \mid X_t = x) \\ &= \sum_{x \in \mathcal{X}} P(X_t = x)P_{x\alpha} \end{aligned} \tag{11}$$

If the distribution at time $t+1$ is the same as at time t then the distribution is called the *stationary distribution*. This means that if the initial state is drawn according to a stationary distribution the Markov process will be stationary.

Example 1 (Find the stationary distribution for a stochastic process)**.** For the stochastic process

$\{X_i\}$ in figure 1 let $\mu \in \mathbb{R}^{|\mathcal{X}|}$ be the stationary distribution. Here μ is a row vector. The probability transition matrix P is given in (10). μ is found by solving

$$\mu_i = \sum_{j=1}^{|\mathcal{X}|} \mu_j P_{ji}, \quad \text{for } i = 1, 2, \ldots, |\mathcal{X}| \tag{12}$$

which is equivalent to solving $\mu P = \mu$.

$$\mu P = \mu \Rightarrow \mu P = \mu I \Rightarrow \mu(P - I) = 0 \tag{13}$$

Solve for μ to obtain $\alpha\mu_1 = \beta\mu_2$. As μ is a probability distribution it follows that $\mu_1 + \mu_2 = 1$ μ can be found by solving the following system of equations.

$$\begin{aligned} \alpha\mu_1 - \beta\mu_2 &= 0 \\ \mu_1 + \mu_2 &= 1 \end{aligned} \tag{14}$$

From this $\mu_1 = \frac{\beta}{\alpha+\beta}$ and $\mu_2 = \frac{\alpha}{\alpha+\beta}$ is obtained which form the stationary distribution

$$\mu = \begin{bmatrix} \mu_1 & \mu_2 \end{bmatrix} = \begin{bmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{bmatrix} \tag{15}$$

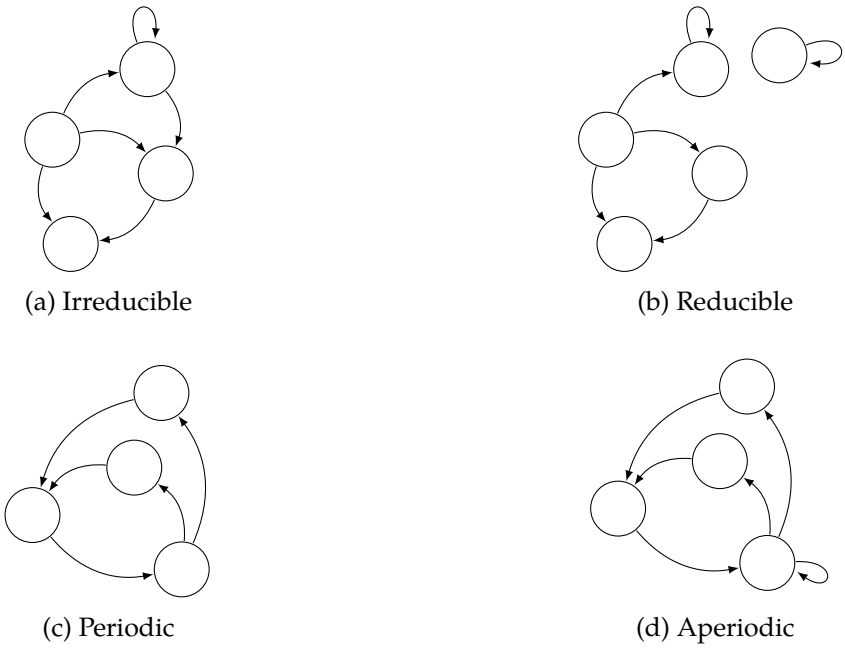
From the Perron-Frobenius theorem (Perron, 1907) it follows that for a time invariant Markov process the stationary distribution always exists as it ensures that every stochastic matrix P has a vector μ such that $\mu = \mu P$.

If from any state in a Markov process any other state can be reached in a finite number of steps with a non-zero probability the Markov chain is said to be irreducible. Let $\mathbb{Y}$ be the set of all cycles in the Markov process. If there is a $k \in \mathbb{N}$ such that $k > 1$ and $\forall c \in \mathbb{Y} : k \mid l(c)$, where $l(c)$ is the length of cycle c, then the Markov process is periodic. Otherwise the Markov process is aperiodic. Both irreducibility and aperiodicity is shown visually in figure 2

Theorem 1 (Markovity and Stationarity)**.** *Let* $\{X_i\}$ *be a Markov process that is both irreducible and aperiodic. Then*

1. $\{X_i\}$ *has a unique stationary distribution* μ*;*
2. *Independent of the initial distribution* P_{X_1}, P_{X_k} *will converge to the stationary distribution as* $k \to \infty$*;*
3. *The Markov process is stationary iff the initial distribution is chosen according to* μ*.*

Table 1 is an illustration of Theorem 1 item 2. The Markov process of figure 1 is initialized with $\alpha = \frac{1}{2}$ and $\beta = \frac{3}{4}$. Then $\mu = [\frac{3}{5}, \frac{2}{5}]$. Table 1 shows that as $k \to \infty$ the distribution converges to μ.



(a) Irreducible (b) Reducible

(c) Periodic (d) Aperiodic

Figure 2: Examples of irreducible, reducible, periodic and aperiodic Markov processes.

$P_{X_k}(\cdot)$	$k = 1$	$k = 2$	$k = 3$	$k = 4$
$P_{X_k}(S)$	1	$\frac{1}{2} = 0.5$	$\frac{5}{8} = 0.625$	$\frac{19}{32} = 0.59375$
$P_{X_k}(R)$	0	$\frac{1}{2} = 0.5$	$\frac{3}{8} = 0.375$	$\frac{13}{32} = 0.40625$
$P_{X_k}(\cdot)$	$k = 5$		$\cdots$	$k = \infty$
$P_{X_k}(S)$	$\frac{77}{128} = 0.6015625$		$\cdots$	$\frac{3}{5} = 0.6$
$P_{X_k}(R)$	$\frac{51}{128} = 0.3984375$		$\cdots$	$\frac{2}{5} = 0.4$

Table 1: Convergence to stationary distribution when $k \to \infty$. (Table taken from Moser (2013))

3.1 Google PageRank and Stationary Distribution

Up to this point stochastic processes are mostly described as processes that have some sort of time index. However, this is not needed at all. Murphy (2012) describes Google's PageRank as a stochastic process in which the the stationary distribution is determined to compute the authority of a web page. This is briefly discussed below to give an example of a different type of system that stochastic processes are used for.

Firstly, in information retrieval the standard process is to build an inverted document index which is a mapping from words to the documents they occur in. When a query is entered into the system this can also be seen as a document. Documents that are similar to this query are likely to be those the user is searching for. However, on the world wide web there is an additional source of information. The idea is that a web page that is linked to more often is more authoritative than a page that is linked to less often. Pages that are more authoritative should rank higher when they match the query. However, to protect for link farms, i.e. thousands of websites that only link to other websites to boost the rank, the authority of the web page with

the outgoing link is taken into account. This results in the following recursive definition for authority of page j.

$$\pi_j = \sum_i A_{ij}\pi_i \tag{16}$$

where A_{ij} is the probability of web page i having a link to web page j. The first idea would be to have a uniform distribution over all outgoing links on a web page. However, according to the Perron-Frobenius theorem (Perron, 1907), to have a unique PageRank all items in A should be strictly positive. Therefore there should be a small probability of having a link from any page to every other page, including itself. By then solving equation (16) for π the authority for each page is found. For big networks, such as the world wide web, this is costly. However, no further detail is given in the present study.

This clearly shows that stochastic processes can also be used when there is no 'time' involved or even 'states'. In this case just the fact of moving between pages is enough to employ the 'power' of stochastic processes.

4 Entropy Rate of a Stochastic Process

In the previous section several varieties of stochastic processes are discussed. It is of interest to compute the uncertainty in these processes. Shannon entropy (Shannon, 2001) can be used to compute the uncertainty in bits for a single random variable. For example the entropy of a random variable in the Markov process $\{X_i\}$ of Figure 1 is

$$H(X_i) = H\left(\frac{\alpha}{\alpha+\beta}, \frac{\beta}{\alpha+\beta}\right) = h\left(\frac{\alpha}{\alpha+\beta}\right). \tag{17}$$

However, this is not the entropy of the stochastic process. If not cited otherwise this section follows chapter five of Moser (2013).

Example 2. Let $\{X_i\}$ be a stochastic process such that all X_i are i.i.d. Recall that the entropy is the average number of bits to encode a single source symbol. As all X_i are i.i.d. each random variable emits symbols according to the same distribution. Therefore the output of each RV can be encoded using $H(X_i)$ bits. If the entropy rate of a stochastic process is the average number of bits used to encode a source symbol it makes sense that for an i.i.d. stochastic process the entropy rate is equal to the entropy of its random variables. That is,

$$H(\{X_i\}) = H(X_i) = H(X_1) \quad \text{if } X_i \text{ are i.i.d.} \tag{18}$$

However, the following example shows that this is not always the case.

Example 3. Let $\{Y_i\}$ be a Markov process with an initial distribution $P_{Y_1}(0) = P_{Y_1}(1) = \frac{1}{2}$. Furthermore let the probability transition matrix P_Y be defined as

$$P_Y \triangleq \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}. \tag{19}$$

That is, if the process is in state 0 it will always go to state 1 in the next time step and after it reached state 1 it will forever stay in state 1. The entropy of the initial state is $H(X_1) = h(\frac{1}{2}) = 1$.

For all states X_i for $i = 2, 3, \ldots$ there is no uncertainty and the entropy is zero. Obviously entropy of this process is not equal to the entropy of one of the random variables as these differ. Also it is not equal to $H(X_1, X_2, \ldots, X_n) = 1$ as on average zero bits are needed to encode the output symbols.

This clearly shows that a new definition is needed for the entropy rate of stochastic processes.

Definition 4 (Entropy Rate for Stochastic Processes)**.** The entropy rate (that is, the entropy rate per source symbol) for a stochastic process $\{X_i\}$ is defined as

$$H(\{X_i\}) \triangleq \lim_{n\to\infty} \frac{H(X_1, X_2, \ldots, X_n)}{n}. \tag{20}$$

If the limit exists.

For $\{X_i\}$ from Example 2 this means that the entropy rate

$$\begin{aligned} H(\{X_i\}) &= \lim_{n\to\infty} \frac{H(X_1, X_2, \ldots, X_n)}{n} \\ &= \lim_{n\to\infty} \frac{H(X_1) + H(X_2) + \cdots + H(X_n)}{n} \\ &= \lim_{n\to\infty} \frac{nH(X_1)}{n} = H(X_1) \end{aligned} \tag{21}$$

Which still complies with (18). For $\{Y_i\}$ from example 3 however the entropy rate is

$$\begin{aligned} H(\{Y_i\}) &= \lim_{n\to\infty} \frac{H(Y_1, Y_2, \ldots, Y_n)}{n} \\ &= \lim_{n\to\infty} \frac{H(Y_1) + H(Y_2) + \cdots + H(Y_n)}{n} \\ &= \lim_{n\to\infty} \frac{H(Y_1)}{n} = 0. \end{aligned} \tag{22}$$

This makes sense as only the initial random variable has any uncertainty. After this there is no uncertainty in the process. As the number of time steps goes to infinity the average number of bits per symbol approaches zero. This is also according to earlier intuition. A different measure for entropy can also be defined.

Definition 5 (Entropy Rate Given Past)**.**

$$H'(\{X_i\}) \triangleq \lim_{n\to\infty} H(X_n \mid X_{n-1}, \ldots, X_1) \tag{23}$$

$H(\{X_i\})$ can be seen as the average entropy rate per source symbol and $H'(\{X_i\})$ is the entropy rate of the last random variable given all random variables in the past. Theorem 2 states that it doesn't matter which entropy rate is used for stationary stochastic processes.

Theorem 2. *For stationary stochastic processes the limit* $H(\{X_i\})$ *always exists and is equal to* $H'(\{X_i\})$*. That is,*

$$H(\{X_i\}) = \lim_{n\to\infty} \frac{H(X_1, \ldots, X_n)}{n} = \lim_{n\to\infty} H(X_n \mid X_{n-1}, \ldots, X_1) = H'(\{X_i\}). \tag{24}$$

Moreover,

1. $H(X_n \mid X_{n-1}, \ldots, X_1)$ *is non increasing in* n*;*
2. $\frac{1}{n}H(X_1, \ldots, X_n)$ *is non increasing in* n*;*
3. $H(X_n \mid X_{n-1}, \ldots, X_1) \leqslant \frac{1}{n}H(X_1, \ldots, X_n)$ *for* $\forall n > 1$.

Proof of Theorem 2, sub 1.

$$\begin{aligned} H(X_n \mid X_{n-1}, \ldots, X_1) &= H(X_{n+1} \mid X_n, \ldots, X_2) && (25)\\ &\leqslant H(X_{n+1} \mid X_n, \ldots, X_2, X_1) && (26)\end{aligned}$$

Where (25) follows from stationarity of the process and the inequality of (26) follows from the fact that conditioning reduces entropy. □

Proof of Theorem 2, sub 3.

$$\begin{aligned} \frac{1}{n}H(X_1, \ldots, X_n) &= \frac{1}{n}\sum_{k=1}^{n} H(X_k \mid X_{k-1}, \ldots, X_1) && (27)\\ &\geqslant \frac{1}{n}\sum_{k=1}^{n} H(X_n \mid X_{n-1}, \ldots, X_1) && (28)\\ &= H(X_n \mid X_{n-1}, \ldots, X_1)\end{aligned}$$

Here (27) follows from the chain rule for entropy and (28) follows from Theorem 2 sub 1. □

Proof of Theorem 2, sub 2.

$$\begin{aligned} H(X_1, \ldots, X_n, X_{n+1}) &= H(X_1, \ldots, X_n) + H(X_{n+1} \mid X_n, \ldots, X_1) && (29)\\ &\leqslant H(X_1, \ldots, X_n) + H(X_n \mid X_{n-1}, \ldots, X_1) && (30)\\ &\leqslant H(X_1, \ldots, X_n) + \frac{1}{n}H(X_1, \ldots, X_n) && (31)\\ &= \frac{n+1}{n}H(X_1, \ldots, X_n)\end{aligned}$$

Where (29) follows from the chain rule, (30) follows from Theorem 2 part 1 and (31) follows from Theorem 2 part 3. From this it follows that

$$\frac{1}{n+1}H(X_1, \ldots, X_n, X_{n+1}) \leqslant \frac{1}{n}H(X_1, \ldots, X_n). \quad (32)$$

□

For the proof of Theorem 2 one more part is needed, namely Cesáro Mean. The proof for this is omitted.

Lemma 1 (Cesáro Mean)**.** *If* $\lim_{n\to\infty} a_n = a$ *and* $b_n \triangleq \frac{1}{n}\sum_{k=1}^{n} a_k$ *then* $\lim_{n\to\infty} b_n = a$.

Proof of Theorem 2. The first part of Theorem 2 states that the limit $H(\{X_i\})$ always exists. Theorem 2 sub 2 states that $\frac{1}{n}H(X_1, \ldots, X_n)$ is non increasing in n, however $H(X_1, \ldots, X_n)$ is the joint entropy of $X_1, \ldots, X_n$ and consequently is lower bounded by zero. It follows that

$\frac{1}{n}H(X_1, \ldots, X_n)$ must converge and that the limit must exist. It remains to show that both limits converge to the same limit.

$$
\begin{aligned}
H(\{X_i\}) &= \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n) \\
&= \lim_{n\to\infty} \overbrace{\frac{1}{n} \sum_{k=1}^{n} \underbrace{H(X_k \mid X_{k-1}, \ldots, X_1)}_{\triangleq a_k}}^{\triangleq b_n} && (33) \\
&= \lim_{n\to\infty} H(X_n \mid X_{n-1}, \ldots, X_1) && (34) \\
&= H'(\{X_i\})
\end{aligned}
$$

Step (33) follows from the entropy chain rule and (34) follows from Lemma 1. □

For a stationary Markov process the entropy rate is particularly easy to compute. Let $\{Z_i\}$ be a stationary Markov chain with stationary distribution μ and probability transition matrix P.

$$
\begin{aligned}
H(\{Z_i\}) &= H'(\{Z_i\}) \\
&= \lim_{n\to\infty} H(Z_n \mid Z_{n-1}, \ldots, Z_1) \\
&= \lim_{n\to\infty} H(Z_n \mid Z_{n-1}) && (35) \\
&= H(Z_2 \mid Z_1) && (36) \\
&= -\sum_{i=1}^{|\mathcal{Z}|} \mu_i \left(\sum_{j=1}^{|\mathcal{Z}|} P_{ij} \log P_{ij} \right) && (37)
\end{aligned}
$$

Step (35) follows from the markovity of the process and (36) follows from the stationarity of the process. This shows that the entropy rate of a Markov process is not dependent on its initial distribution, but only on the transitions between the states and the stationary distribution (Cover and Thomas, 2012).

4.1 Detecting Spam in Blog Comments using Entropy Rate

Calculating the entropy rate of a stochastic process is useful in several applications. One of these applications is the detection of spam messages in blog comments. The following example is a brief illustration on the usefulness of computing the entropy rate in such a case.

Kantchelian et al. (2012) designed a method to detect spam in blog comments. The approach described in the paper is based on the assumption that calculating the entropy rate of comments will give a useful indication of the distinction between spam and ham (non-spam). For this project a Lempel-Ziv-Markov chain algorithm was used, which makes sense, as language itself can be seen as a markov-chain where every word is dependent of its direct predecessor. As spam words might appear often after other spam words, the uncertainty of the string decreases. When this is true, the entropy for the markov process lowers as well. Figure 3 shows that strings containing spam on average have a lower entropy rate. A user was labeled as spam iff one of the comments by that particular user was labeled as spam. Conform the expectations, comments by spam users had a lower entropy rate than comments by non-spam users. The figure also shows that blog posts in general contain less bits per character than e-books, from which can be deducted that blogposts are less informative than texts from e-books.

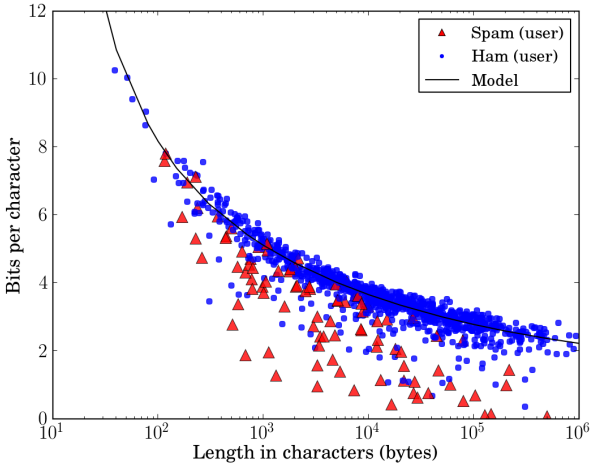


Figure 3: Spam and non-spam (ham) users plotted against a model created using an average entropy rate of a dozen e-books. Each dot represents all contributions of a user in the dataset, concatenated into one string. A user was considered a spam user iff one of the comments by that particular user was labeled as spam. (Figure taken from Kantchelian et al. (2012)).

5 Conclusion

Stochastic processes were introduced as processes that model non deterministic systems, i.e. systems in which multiple outcomes are possible. Stochastic processes are modeled using an indexed series of random variables. These RVs can have arbitrary dependence. It is also possible to have stricter constraints for the stochastic process. In that case the process can, for example, be a Markov process in which the rules for dependence are more strict. These stricter constraints help later when one tries to reason with the model and is is easy to, for example, compute the entropy rate. Other properties that a stochastic process may or may not have were also defined.

The normal Shannon entropy is not applicable to stochastic processes. It was shown that using this definition of entropy results are found that do not agree with the intuition about entropy and uncertainty. Therefore a new definition for the entropy rate of stochastic processes is defined that does comply with the intuition about entropy and uncertainty.

A second definition for entropy of stochastic processes is also given.The first is $H(\{X_i\})$ which is the entropy measured in average number of bits per source symbol, and the second entropy rate is $H'(\{X_i\})$ which is the number of bits used for the last source symbol given all source symbols in the past. It was proved that for a stationary stochastic process is does not matter which one is used as both exist and are equal for stationary stochastic processes.

Finally it was shown that due to its properties it is straightforward to compute the entropy rate for a (stationary) Markov process. This shows that when using stochastic processes to model real world systems it may be of interest to put extra constraints on the model to ease the computations of, for example, the entropy rate of a system.

References

Bachelier, L. (2011). *Louis Bachelier's theory of speculation: the origins of modern finance*. Princeton University Press.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Kantchelian, A., Ma, J., Huang, L., Afroz, S., Joseph, A., and Tygar, J. (2012). Robust detection of comment spam using entropy rate. In *Proceedings of the 5th ACM Workshop on Security and Artificial Intelligence*, pages 59–70. ACM. http://dx.doi.org/10.1145/2381896.2381907.

Moser, S. M. (2013). *Information Theory: Lecture Notes*.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Perron, O. (1907). Zur theorie der matrices. *Mathematische Annalen*, 64(2):248–263. http://dx.doi.org/10.1007/BF01449896.

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55. http://dx.doi.org/10.1145/584091.584093.

Shinozuka, M. and Deodatis, G. (1988). Stochastic process models for earthquake ground motion. *Probabilistic engineering mechanics*, 3(3):114–123. http://dx.doi.org/10.1016/0266-8920(88)90023-9.