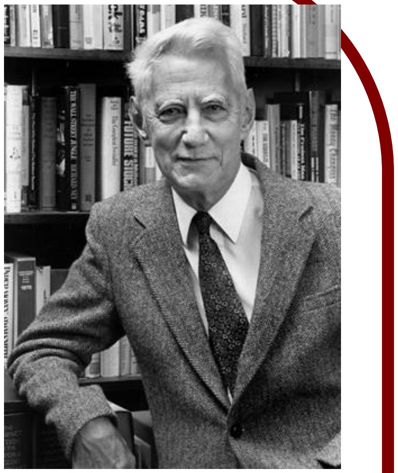
ECSE-626

Statistical Computer Vision

Information Theory

Introduction

- Founded by Claude Shannon in 1948.



www.bell-labs.com/news/2001/february/26/1.html

- Theory for transmission of information over a channel.
- Information theory is based on probability theory.

Introduction

- Some consider information theory to be a subset of communication theory.
- Information theory asks:
  - What is the ultimate data compression? (answer: entropy)
  - What is the ultimate transmission rate of communication? (answer: channel capacity)

Introduction

- Fundamental contributions in:
  - Electrical engineering (communication theory)
  - Statistical physics (thermodynamics)
  - Computer science (complexity)
  - Statistical inference (Occam’s razor)
  - Probability and statistics (error rates for hypothesis testing and estimation)

Introduction

- Communication Theory:
  - Early 40’s, thought that increasing transmission rate of information over communication channel increased the probability of error.
  - Shannon proved that this was not true if communication rate was below channel capacity (which can be computed from channel noise characteristics).

Introduction

- Communication Theory:
  - Shannon argued that random processes (music, speech) have an irreducible complexity below which the signal cannot be compressed – *entropy*.
  - If entropy of source below capacity of channel, then asymptotically error free communication possible.

Introduction

- Progress in IC and code design based on his theory.
- Example of application: the use of error codes on compact disks.

Entropy

The entropy of a random variable X with a probability mass function $p(x)$ is defined:

$$H(X) = -\sum p(x) \log_2 p(x).$$

The entropy is a measure of the average uncertainty in the random variable.

It is the number of bits (or information) on average required to describe the random variable.

Entropy

- Example 1: Consider a random variable which has a uniform distribution over 32 outcomes.
- To identify an outcome, we need a label that takes on 32 different values – 5-bit strings suffice.
- The entropy of this random variable is:

$$H(x) = -\sum_{i=1}^{32} p(i)\log p(i) = -\sum_{i=1}^{32} \frac{1}{32}\log\frac{1}{32} = 5 \text{ bits.}$$

This agrees with the number of bits needed to describe X.

Entropy

- Example 2: Suppose that the probability of winning for 8 horses participating in a race are:

$$\left(\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{16},\frac{1}{64},\frac{1}{64},\frac{1}{64},\frac{1}{64}\right).$$

The entropy of the horse race is:

$$H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} \ldots = 2 \text{ bits.}$$

Entropy

- To send a message indicating which horse won, one can send the index of the winning horse – 3 bits for 8 horses.
- Because the win probabilities are not uniform, it makes sense to use shorter descriptions for the more probable horses and longer descriptions for the less probable ones.
- In this way, we can achieve lower average description length.

Entropy

- For example. We could you the following descriptor strings:

  0, 10, 110, 1110, 111100, 111101, 111110, 111111.

- The average description length is 2 bits (= entropy) , as opposed to 3 bits for the uniform code.
- The entropy gives the lower bound for the average descriptor length.

Entropy

- Entropy can also be interpreted as the expected value of log(1/p(X)) where X is drawn according to the probability mass function p(x).

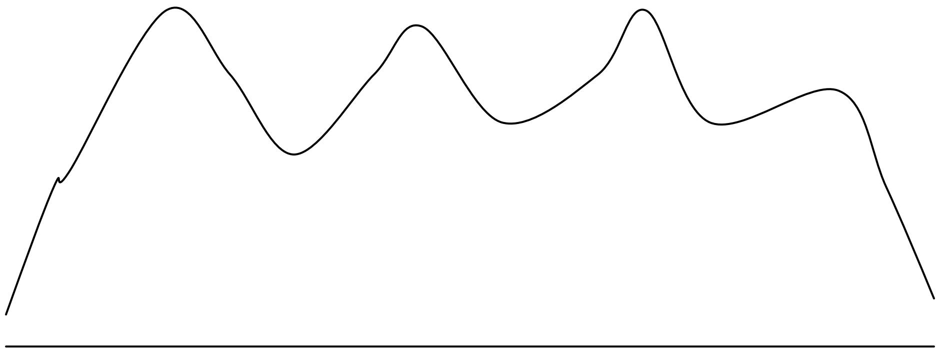
$$H(X) = E_p \log \frac{1}{p(X)}.$$

Entropy

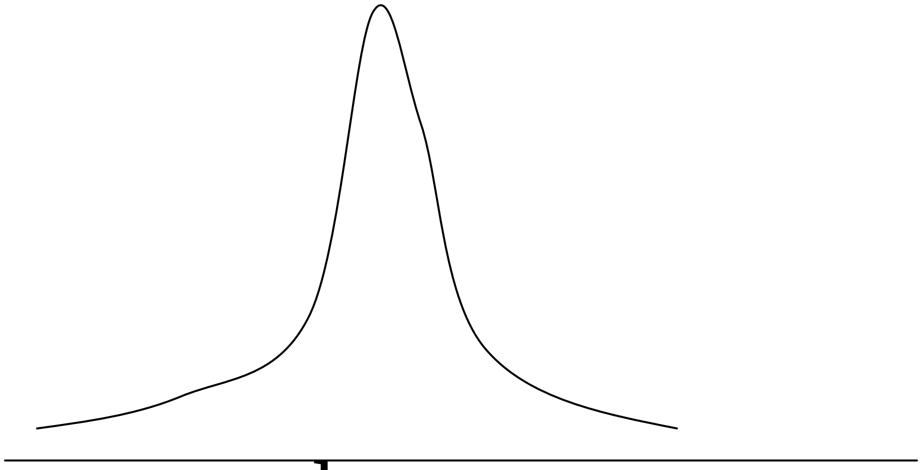
- Why is this useful to computer vision?

- Entropy permits us to *quantify* the uncertainty in a distribution.

Entropy



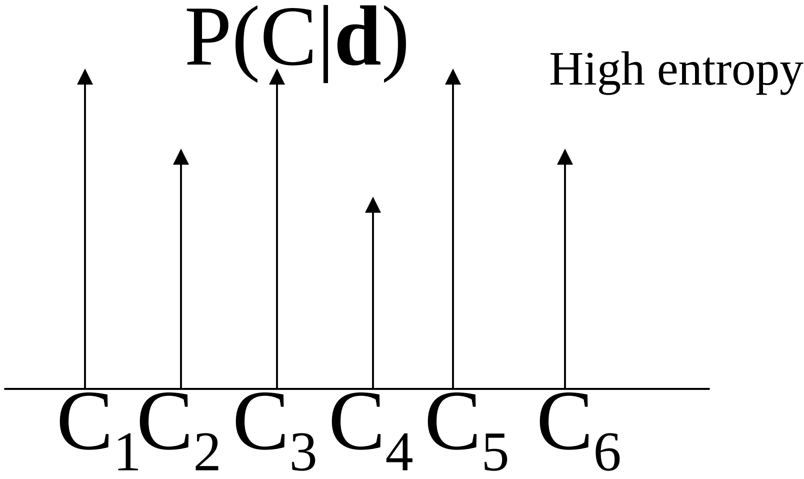
high entropy

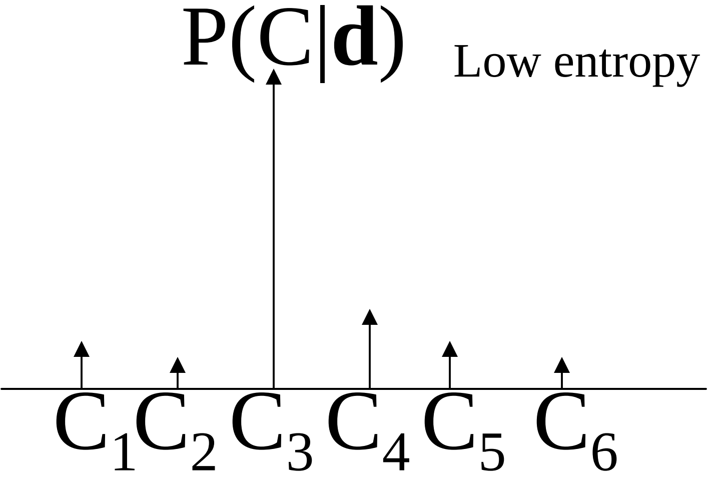


low entropy

Entropy

- In terms of Bayesian inference, this gives us a measure of confidence in the posterior distribution.
- For example, in the case of object recognition:





Entropy

- This gives us a measure of the validity of choosing a MAP solution.
- In cases of low entropy, it makes sense.
- In cases of high entropy, it might not…

Joint Entropy

The joint entropy $H(X,Y)$ of a pair of discrete random variables (X,Y) with a joint distribution $p(x,y)$ is defined as:

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y).$$

Conditional Entropy

- The *conditional entropy* of a random variable given another is defined as the expected value of the entropies of the conditional distributions, averaged over the conditioning variable.

Conditional Entropy

The conditional entropy H(Y|X) is defined as:

$$H(Y\,|\,X) = \sum_{x\in X} p(x)H(Y\,|\,X=x)$$

$$= -\sum_{x\in X} p(x)\sum_{y\in Y} p(y\,|\,x)\log p(y\,|\,x)$$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y\,|\,x)$$

$$= -E_{p(x,\,y)}\log p(Y\,|\,X)$$

Entropy Chain Rule

$$H(X,Y) = H(X) + H(Y \mid X)$$

Proof?

Entropy Chain Rule

$$H(X,Y) = H(X) + H(Y \mid X)$$

Proof:

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y)$$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x)p(y \mid x)$$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y \mid x)$$

$$= -\sum_{x\in X} p(x)\log p(x) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y \mid x)$$

$$= H(X) + H(Y \mid X)$$

Relative Entropy

- The *relative entropy* is a measure of the distance between two distributions.
- Relative entropy $D(p||q)$ is a measure of the inefficiency of assuming that the distribution is q when the true distribution is p.
- It is also referred to as the *Kullback Leibler* (KL) distance between distributions.

Relative Entropy

The Kullback Leibler distance between 2 probability mass functions p(x) and q(x) is defined as:

$$D(p\,||\,q) = \sum_{x\in X} p(x)\log\frac{p(x)}{q(x)} = Ep\log\frac{p(X)}{q(X)}.$$

The relative entropy is always non-negative and is zero iff p=q.

It is not a true distance between 2 distributions since it is not symmetric.

Mutual Information

- *Mutual information* is a measure of the amount of information that one random variable contains about another.
- It is the reduction in uncertainty of one random variable due to the knowledge of the other.

Mutual Information

- Consider two random variables *X* and *Y* with:
  - a joint probability mass function *p*(*x*,*y*),
  - marginal probability mass functions *p*(*x*) and *p*(*y*).

The *mutual information I*(*X*;*Y*) is the relative entropy between the joint distribution and the product distribution *p*(*x*)*p*(*y*), i.e.

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}.$$

Mutual Information

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

$$= E_{p(x,y)} \log \frac{p(X,Y)}{p(X)p(Y)}.$$

The mutual information $I(X;Y)$ is a measure of the dependence between two random variables.

It is symmetric in X and Y and always non-negative.

Entropy and Mutual Information

$$I(X;Y) = H(X) - H(X \mid Y)$$

Proof?

Entropy and Mutual Information

$$I(X;Y) = H(X) - H(X \mid Y)$$

Proof:

$$\begin{aligned}
I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y) \log \frac{p(x \mid y)}{p(x)} \\
&= -\sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x \mid y) \\
&= -\sum_{x} p(x) \log p(x) - \left(-\sum_{x,y} p(x,y) \log p(x \mid y) \right) \\
&= H(X) - H(X \mid Y).
\end{aligned}$$

Entropy and Mutual Information

- This implies that the mutual information is the reduction in the uncertainty of X due to the knowledge of Y.
- It follows that: $I(X;Y) = H(Y) - H(Y|X)$.
- Note that: $I(X;Y) = H(X) + H(Y) - H(X,Y)$.
- Also, $I(X;X) = H(X) – H(X|X) = H(X)$

  (Entropy is also known as *self-information*.)

Entropy and Mutual Information

Summary:

$$I(X;Y) = H(X) - H(X|Y),$$

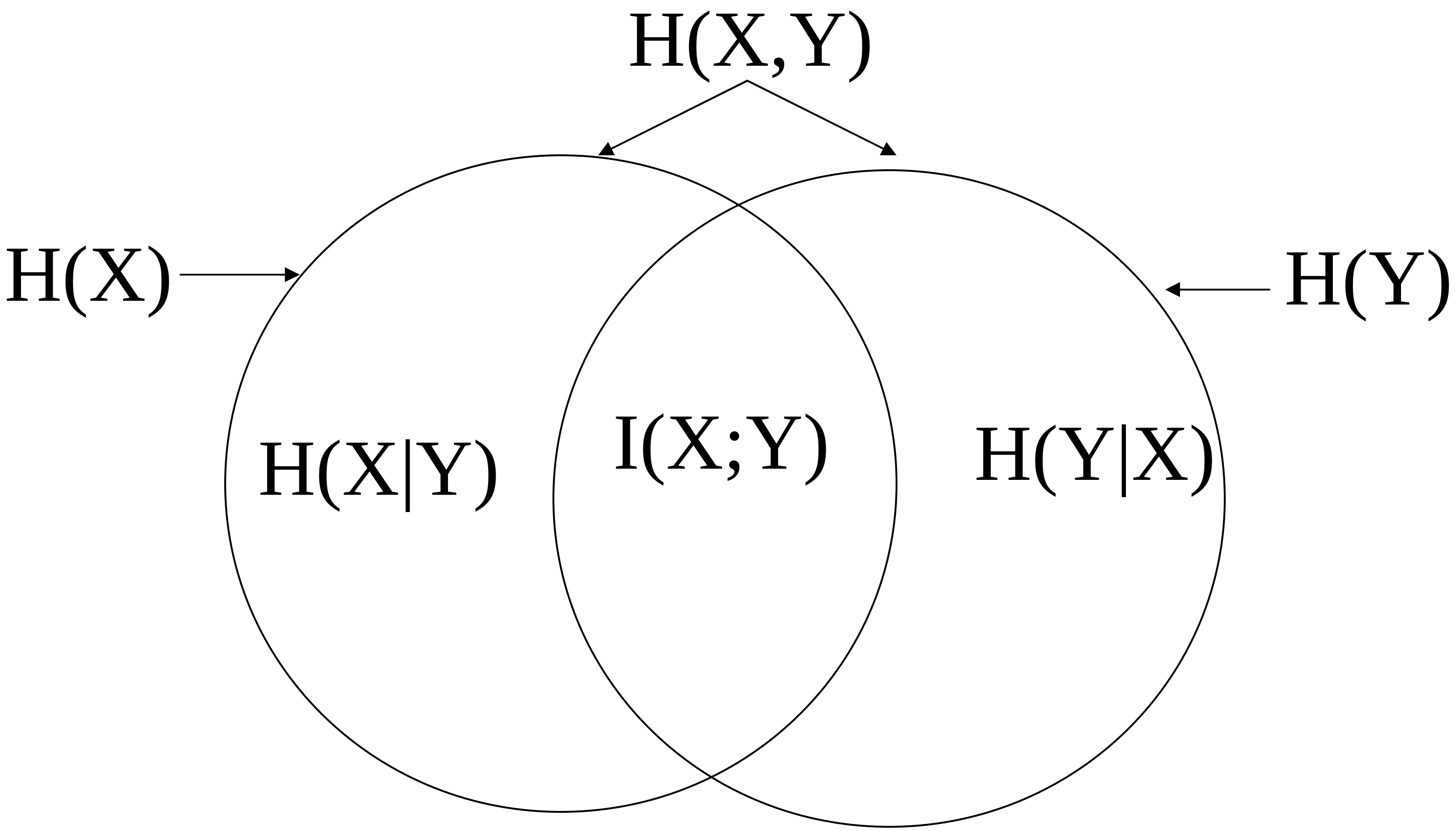
$$I(X;Y) = H(Y) - H(Y|X),$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y),$$

$$I(X;Y) - I(Y;X),$$

$$I(X;X) = H(X).$$

Entropy and Mutual Information



Chain Rule for Entropy

Let $X_1, X_2, \dots, X_n$ be drawn according to $p(x_1, x_2, \dots, x_n)$.

$$H(X_1, X_2, \dots, X_n) = \sum_{i=1}^{n} H(X_i \mid X_{i-1}, \dots, X_1)$$

Conditional Mutual Information

The conditional mutual information of random variables X and Y given Z us defined by:

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$$

$$= E_{p(x,y,z)} \log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}.$$

It is defined as the reduction in the uncertainty of X due to the knowledge of Y when Z is given.

Chain Rule for Information

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y \mid X_{i-1}, X_{i-2}, \ldots, X_1).$$

Entropy Facts

$H(X) \leq \log|X|$, where $|X|$ denotes the number of elements in the range of X, with equality iff X has a uniform distribution over X.

Conditioning reduces entropy:

$H(X|Y) \leq H(X)$

with equality iff X and Y are independent.

Entropy Facts

Independence bound on entropy:

Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, x_2, \ldots, x_n)$.

$$H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

with equality iff X_i are independent.

Data Processing Inequality

Random variables X, Y, Z are said to form a Markov chain in that order, which we will denote $X \rightarrow Y \rightarrow Z,$ if the conditional distribution of Z depends only on Y and is conditionally independent of X.

We now have an important and useful theorem demonstrating that no clever manipulation of the data can improve inferences that can be made from the data.

Data Processing Inequality

In other words, no clever processing of Y, deterministic or random, can increase the information that Ycontains about X.

Theorem: If $X \rightarrow Y \rightarrow Z$, then $I(X;Y) \geq I(X;Z)$.

Data Processing Inequality

By Markovity, I(X;Y|Z) = 0 and we also have that:

$$I(X;Y\mid Z)\leq I(X;Y)$$

This implies that the dependence of X and Y is decreased or stays the same with the observation of a "downstream" random variable Z.

Second Law of Thermodynamics

- One of the basic laws of physics, the second law of thermodynamics, states that the entropy of an isolated system is non-decreasing.
- We model the isolated system as a Markov chain with transitions obeying the physical laws governing the system.
- The entropy doesn’t always increase but the relative entropy always decreases.

Second Law of Thermodynamics

1. Relative entropy $D(\mu_n||\mu_n')$ decreases with time.

μ_n and μ_n' are 2 probability distributions on the state space of a Markov chain at time n.

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

Second Law of Thermodynamics

2. Relative entropy $D(\mu_n||\mu)$ between a distribution μ_n on the states at time n and a stationary distribution μ decreases with time.

Any state distribution gets closer and closer to each stationary distribution as time passes.

Second Law of Thermodynamics

3. Entropy increases if the stationary distribution is uniform.

4. The conditional entropy $H(X_n|X_1)$ increases with n for a stationary Markov process.

5. Shuffles increase entropy.

Fano’s Inequality

- Suppose we know a random variable Y and we wish to guess the value of a correlated random variable X.
- Fano’s inequality relates the probability of error in guessing the random variable X to its conditional entropy H(X|Y).

Fano’s Inequality

- The conditional entropy of a random variable X given another random variable Y is zero iff X is a function of Y.
- We can estimate X from Y with zero probability of error iff H(X|Y)=0.
- By extension, we expect to be able to estimate X with a low probability of error only if the conditional entropy H(X|Y) is small.

Fano's Inequality

- Suppose we want to estimate a random variable X with a distribution $p(x)$.
- We observe a random variable Y which is related to X by the conditional distribution $p(y|x)$.

Fano’s Inequality

From Y, we calculate a function $g(Y)=\hat{X}$, an estimate of X.

We wish to bound the probability that $\hat{X} \neq X$.

We observe that $X \rightarrow Y \rightarrow \hat{X}$ forms a Markov chain.

The probability of error can be defined as:

$P_e=\Pr\{\hat{X} \neq X\}$.

Fano's inequality:

$H(P_e)+P_e \log(|X|-1) \geq H(X|Y)$.

Fano’s Inequality

This inequality can be weakened to:

$$P_e \geq \frac{H(X \mid Y) - 1}{\log(|X|)}$$

Stochastic Processes

- A *stochastic process* is an indexed sequence of random variables.
- There can be arbitrary dependence among the random variables.
- A stochastic process is said to be *stationary* if the joint distribution of any subset of the sequence is invariant to shifts in the time index.

Stochastic Processes

Stationary process:

$\Pr\{X_1=x_1,X_2=x_2,\ldots,X_n=x_n\}=\Pr\{X_{1+m}=x_1,X_{2+m}=x_2,\ldots,X_{n+m}=x_n\}$ for all shifts m and for all $x_1, x_2, \ldots, x_n \in X$.

A Markov chain is said to be time-invariant if the conditional probability $\Pr\{X_{n+1}=b|X_n=a\}=\Pr\{X_2=b|X_1=a\}$. for all a, b $\in \mathcal{X}$.

Stochastic Processes

- More terminology…
- A Markov chain is said to be *irreducible* if it is possible to go from any state of the Markov chain to any other state in a finite number of states with positive probability.

Stochastic Processes

- *Stationary distribution*: A distribution on the states such that the distribution at time n+1 is the same as the distribution at time n.

Entropy Rate

The entropy rate of a stochastic process {X_i} is defined by

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

when the limit exists. It states how the entropy of the sequence grows with n.

In addition, a related quantity can be defined:

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

Entropy Rate

For a stationary Markov chain, the entropy rate is given by

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

$$= \lim_{n\to\infty} \frac{1}{n} H(X_n \mid X_{n-1}) = H(X_2 \mid X_1).$$

Entropy Rate

Let {Xi} be a stationary Markov chain with stationary distribution μ and transition matrix P.

Then the entropy rate is:

$$H(|X|)=-\sum_{i,j} \mu P_{ij} \log P_{ij}$$