# Differential entropy

A continuous random variable $X$ has the probability density function $f(x)$. The *differential entropy* $h(X)$ of the variable is defined as

$$h(X) = -\int_{-\infty}^{\infty} f(x) \cdot \log f(x)\ dx$$

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

Translation and scaling

$$h(X + c) = h(X)$$

$$h(aX) = h(X) + \log |a|$$

# Common distributions

Normal distribution (gaussian distribution)

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

Laplace distribution

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

Uniform distribution

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

# Differential entropy, cont.

The gaussian distribution is the distribution that maximizes the differential entropy, for a given variance. Ie, 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 if $X$ is gaussian.

If we instead only consider distributions with finite support, the differential entropy is maximized (for a given support) by the uniform distribution.

# Quantization

Suppose we do uniform quantization of a continuous random variable $X$. The quantized variable $\hat{X}$ is a discrete variable. The probability $p(x_i)$ for the outcome $x_i$ is approximately $\Delta \cdot f(x_i)$, where $\Delta$ is the step size of the quantizer. The entropy of the quantized variable is

$$\begin{aligned}
H(\hat{X}) &= -\sum_i p(x_i) \cdot \log p(x_i) \\
&\approx -\sum_i \Delta f(x_i) \cdot \log(\Delta f(x_i)) \\
&= -\sum_i \Delta f(x_i) \cdot \log f(x_i) - \sum_i \Delta f(x_i) \cdot \log \Delta \\
&\approx -\int_{-\infty}^{\infty} f(x) \cdot \log f(x)\ dx - \log \Delta \int_{-\infty}^{\infty} f(x)\ dx \\
&= h(X) - \log \Delta
\end{aligned}$$

# 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

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

The conditional differential entropy is defined as

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

Conditioning reduces the differential entropy

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

We have

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

# Differential entropy, cont.

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

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

which gives

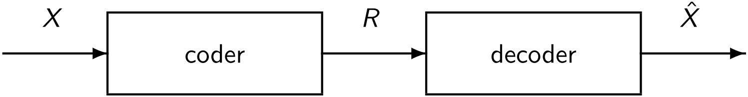
$$I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) = h(X) + h(Y) - h(X,Y)$$

We have that $I(X;Y) \geq 0$ with equality iff $X$ and $Y$ are independent.

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

$$\begin{aligned} I(\hat{X};\hat{Y}) &= H(\hat{X}) - H(\hat{X}|\hat{Y}) \\ &\approx h(X) - \log\Delta - (h(X|Y) - \log\Delta) \\ &= I(X;Y) \end{aligned}$$

# Coding with distortion




If we remove the demand that the original signal $X$ and the decoded signal $\hat{X}$ should be the same, we can get a much lower rate $R$. The downside is of course that we get some kind of distortion.

# Distortion

There are many distortion measures to use. When the signal alphabet is the real numbers, the most common measure is the *mean square error*. Given an original sequence $x_i, i = 1, \ldots, n$ and the corresponding decoded sequence $\hat{x}_i, i = 1, \ldots, n$ the distortion is then

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x}_i)^2$$

If we have a random signal model, with original signal $X_i$ and decoded signal $\hat{X}_i$, the distortion is then

$$E\{(X_i - \hat{X}_i)^2\} = \int_{x,\hat{x}} f(x,\hat{x})(x - \hat{x})^2 dx d\hat{x}$$

# Rate-distortion function

The *rate-distortion function* $R(D)$ gives the theoretical lowest rate $R$ (in bits/sample) that we can ever achieve, on the condition that the resulting distortion is not larger than $D$.

For a memoryless stationary continuous random source $X_i$, the rate-distortion function is given by

$$R(D) = \min_{f(\hat{x}|x): E\{(X_i - \hat{X}_i)^2\} \leq D} I(X_i; \hat{X}_i)$$

The minimization is performed over all conditional density functions $f(\hat{x}|x)$ for which the joint density function $f(x, \hat{x}) = f(x) \cdot f(\hat{x}|x)$ satisfies the distortion constraint.

Note that we don't have a deterministic mapping from $x$ to $\hat{x}$.

# Gaussian source

If the source is a memoryless gaussian source with zero mean and variance $\sigma^2$, the rate-distortion function is

$$R(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D} & 0 \leq D \leq \sigma^2 \\ 0 & D > \sigma^2 \end{cases}$$

Short proof:
If $D > \sigma^2$ we choose $\hat{X}_i = 0$ with probability 1, giving us $I(X;\hat{X}) = 0$ and thus $R(D) = 0$.
If $D \leq \sigma^2$ we have

$$\begin{aligned} I(X;\hat{X}) &= h(X) - h(X|\hat{X}) = h(X) - h(X - \hat{X}|\hat{X}) \\ &\geq h(X) - h(X - \hat{X}) \geq h(X) - h(\mathcal{N}(0, E\{(X-\hat{X})^2\})) \\ &= \frac{1}{2}\log 2\pi e\sigma^2 - \frac{1}{2}\log 2\pi e E\{(X-\hat{X})^2\} \\ &\geq \frac{1}{2}\log 2\pi e\sigma^2 - \frac{1}{2}\log 2\pi e D = \frac{1}{2}\log\frac{\sigma^2}{D} \end{aligned}$$

# Gaussian source, cont.

We have thus shown that

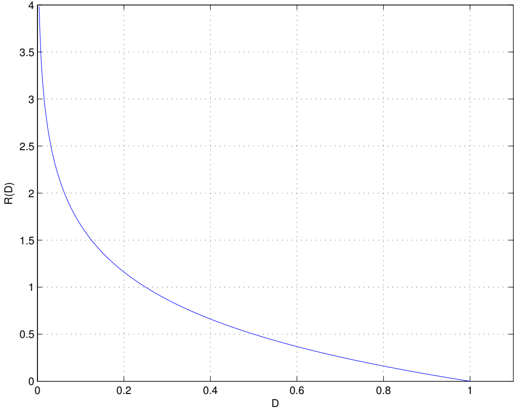
$$R(D) \geq \frac{1}{2} \log \frac{\sigma^2}{D}$$

Now we find a distribution that achieves the bound.
Suppose we choose $\hat{X} \sim \mathcal{N}(0, \sigma^2 - D)$ and $Z \sim \mathcal{N}(0, D)$ such that $\hat{X}$ and $Z$ are independent and $X = \hat{X} + Z$. For this distribution we get

$$I(X; \hat{X}) = \frac{1}{2} \log \frac{\sigma^2}{D}$$

and $E\{(X - \hat{X})^2\} = D$.

# Gaussian source




$R(D)$ for a memoryless gaussian source with variance 1. As $D$ tends towards 0, $R(D)$ tends towards infinity.

# Multiple independent gaussian sources

Suppose we have $m$ mutually independent memoryless gaussian sources with zero mean and variances $\sigma_i^2$. Each source has a rate-distortion function $R_i(D_i)$. We want to find the rate-distortion function for all sources at once, ie given a total maximum allowed distortion $D = \sum_{i=1}^m D_i$, what is the lowest total rate $R = \sum_{i=1}^m R_i$?

The problem of finding the rate-distortion function is reduced to the following optimization

$$R(D) = \min_{\sum D_i = D} \sum_{i=1}^m \max\{\frac{1}{2} \log \frac{\sigma_i^2}{D_i}, 0\}$$

to find the optimal allotment of bits to each component.

Lagrange optimization gives that, if possible, we should choose the same distortion for each component. The distortion for component $i$ can never be larger than the variance $\sigma_i^2$ though.

# Multiple independent gaussian sources

The rate-distortion function is thus given by.

$$R(D) = \sum_{i=1}^{m} \frac{1}{2} \log \frac{\sigma_i^2}{D_i}$$

where

$$D_i = \begin{cases} \lambda & , \ \lambda < \sigma_i^2 \\ \sigma_i^2 & , \ \lambda \geq \sigma_i^2 \end{cases}$$

and $\lambda$ is chosen so that $\sum_{i=1}^{m} D_i = D$.

This is often referred to as “reverse water-filling”. We choose a constant $\lambda$ and only describe those components that have a variance larger than $\lambda$. No bits are used for the components that have a variance less than $\lambda$.

# Multivariate gaussian source

Suppose we have an $m$-dimensional multivariate gaussian source $\mathbf{X}$ with zero means and covariance matrix $C$.

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^m |C|}} \exp(-\frac{1}{2}\mathbf{x}^T C^{-1} \mathbf{x})$$

The rate-distortion function is found by doing reverse water-filling on the eigenvalues $s_i$ of $C$

$$R(D) = \sum_{i=1}^{m} \frac{1}{2} \log \frac{s_i}{D_i}$$

where

$$D_i = \begin{cases} \lambda & , \ \lambda < s_i \\ s_i & , \ \lambda \geq s_i \end{cases}$$

and $\lambda$ is chosen so that $\sum_{i=1}^{m} D_i = D$.

# Gaussian source with memory

For gaussian sources with memory, we do reverse water-filling on the spectrum. Each frequency can be seen as an independent gaussian process.

The auto-correlation function of the source is

$$R_{XX}(k) = E\{X_i \cdot X_{i+k}\}$$

and the power spectral density is the Fourier transform of the auto correlation function

$$\Phi(\theta) = \mathcal{F}\{R_{XX}(k)\} = \sum_{k=-\infty}^{\infty} R_{XX}(k) \cdot e^{-j2\pi\theta k}$$

# Gaussian source with memory

The rate-distortion function is then given by.

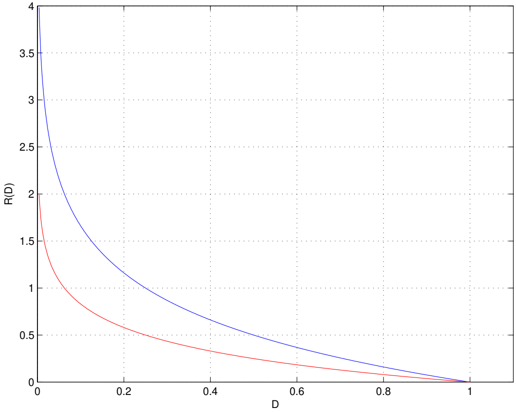
$$R(D) = \int_{-1/2}^{1/2} \max\{\frac{1}{2}\log\frac{\Phi(\theta)}{\lambda}, 0\}\ d\theta$$

where

$$D = \int_{-1/2}^{1/2} \min\{\lambda, \Phi(\theta)\}\ d\theta$$

The integration can of course be done over any interval of size 1, since the power spectral density is a periodic function.

# Gaussian sources




$R(D)$ for an ideally bandlimited gaussian source (red), compared to the $R(D)$ for a memoryless/white gaussian source (blue). Both sources have variance 1.

# Non-gaussian sources

For other distributions, the rate-distortion function can be hard to calculate. However, there are upper and lower bounds.

Given a stationary memoryless random source $X$ with variance $\sigma^2$, the rate-distortion function is bounded by

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

For a gaussian source, both bounds are the same.

For a laplacian source we get

$$\frac{1}{2}\log\frac{\sigma^2}{D} - \frac{1}{2}\log\frac{\pi}{e} \;\leq\; R(D) \;\leq\; \frac{1}{2}\log\frac{\sigma^2}{D}$$

# Real coder

How far from the theoretical rate-distortion are we if we do practical coding?
Suppose we have a memoryless gaussian signal. The signal is quantized with a uniform quantizer and the quantized signal is then source coded.
For uniform quantization, the distortion is approximately

$$D \approx \frac{\Delta^2}{12}$$

Under the assumption that we do a perfect entropy coding of the quantized signal, the data rate is

$$\begin{aligned} R &= H(\hat{X}) \approx h(X) - \log \Delta \approx h(X) - \log \sqrt{12D} \\ &= \frac{1}{2} \log 2\pi e \sigma^2 - \log \sqrt{12D} = \frac{1}{2} \log \frac{\pi e \sigma^2}{6D} \\ &= \frac{1}{2} \log \frac{\sigma^2}{D} + \frac{1}{2} \log \frac{\pi e}{6} \approx \frac{1}{2} \log \frac{\sigma^2}{D} + 0.2546 \end{aligned}$$

# Discrete sources

For discrete alphabets, the mean square error might not be a suitable distortion measure. A common distortion measure is the *Hamming distorsion*, defined by

$$d_H(x, \hat{x}) = \begin{cases} 0 & \text{if } \ x = \hat{x} \\ 1 & \text{if } \ x \neq \hat{x} \end{cases}$$

Given an original sequence $x_i, i = 1, \ldots, n$ and the corresponding decoded sequence $\hat{x}_i, i = 1, \ldots, n$ the distortion is then

$$\frac{1}{n} \sum_{i=1}^{n} d_H(x_i, \hat{x}_i)$$

The Hamming distortion between the two sequences is thus the relative proportion of positions in which they differ.

# Rate-distortion function

For a memoryless stationary discrete random source $X_i$ and using the Hamming distortion measure, the rate-distortion function is given by

$$R(D) = \min_{p(\hat{x}|x):\sum_{x,\hat{x}} p(x)\cdot p(\hat{x}|x)\cdot d_H(x,\hat{x})\leq D} I(X_i;\hat{X}_i)$$

The minimization is performed over all conditional probability distributions $p(\hat{x}|x)$ for which the joint probability distribution $p(x,\hat{x}) = p(x)\cdot p(\hat{x}|x)$ satisfies the distortion constraint.

# Bernoulli source

Given a Bernoulli source (ie a memoryless binary source with probabilities $p$ and $1-p$ for the two outcomes) and using Hamming distortion as the distortion measure, the rate-distortion function is given by

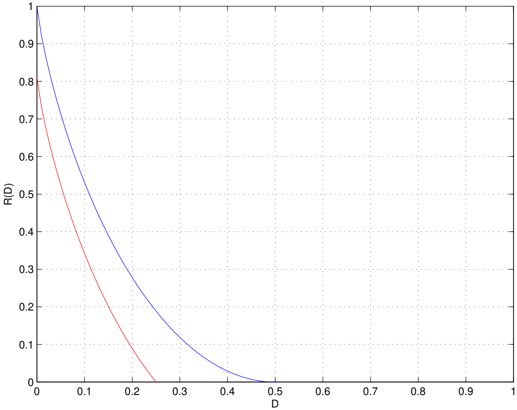
$$R(D) = \begin{cases} H_b(p) - H_b(D) & \text{if } 0 \leq D \leq \min\{p, 1-p\} \\ 0 & \text{if } D > \min\{p, 1-p\} \end{cases}$$

where $H_b(q)$ is the binary entropy function

$$H_b(q) = -q \cdot \log q - (1-q) \cdot \log(1-q)$$

Note that if we require $D = 0$, the lowest possible rate is equal to the entropy rate of the source.

# Bernoulli sources




$R(D)$ for Bernoulli sources with $p = 0.5$ (blue) and $p = 0.75$ (red).

# Real coder

Suppose we have a Bernoulli source. Assume, without loss of generality, that $p \geq 1 - p$, ie $p \geq 0.5$.
Let the coder keep a fraction $0 \leq k \leq 1$ of symbols. Code the symbols that are kept with a perfect source coder and discard the rest.
The decoder will decode the symbols that the coder kept and set the rest to 0 (the most probable value). On average, the fraction of incorrectly decoded symbols will be $(1-k)(1-p)$, which is equal to the distortion $D$, ie

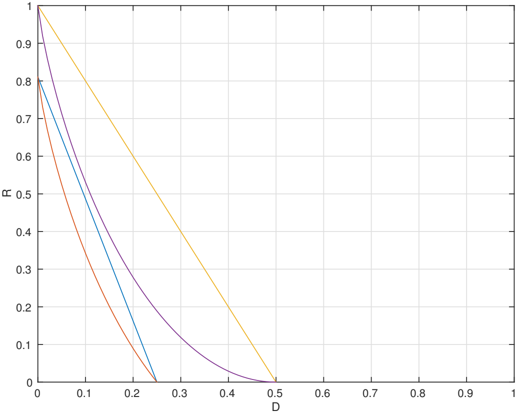
$$(1-k)(1-p) = D \quad \Rightarrow \quad k = 1 - \frac{D}{1-p}$$

The rate of the coder, assuming that the source coder achieves the entropy bound is

$$R = k \cdot H_b(p) = \left(1 - \frac{D}{1-p}\right) \cdot H_b(p)$$

which is a straight line between $(0, H_b(p))$ and $(1-p, 0)$.

# Real coder




Performance of our real coder compared with the rate-distortion function for $p = 0.5$ (yellow/magenta) and $p = 0.75$ (blue/red).