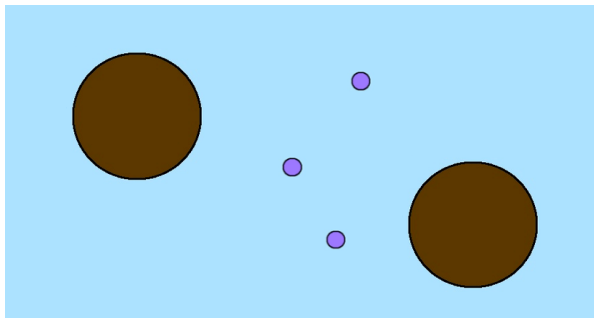
Entropy

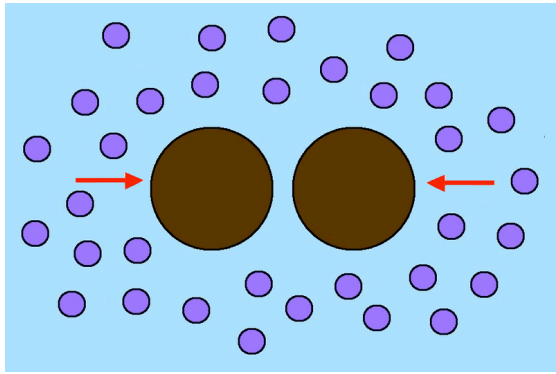
1 (Boltzmann)Physics (micro-macro)
2 (Shannon)Communication (compression)
3 (Bayes) ML – KL Divergence (prior to posterior)

Entropy

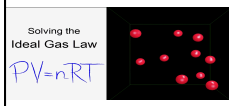
Communication



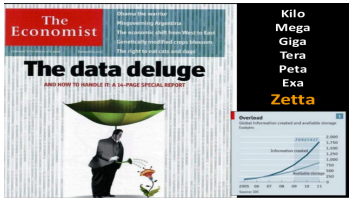
Why?



What is the Hope in Learning?



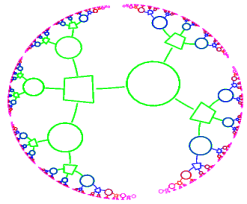
Avogadro constant $\equiv$ 6.02×10^{23} particles mol^{-1}



– Kilo	$= 10^3$	$= 2^{10}$
– Mega	$= 10^6$	$= 2^{20}$
– Giga	$= 10^9$	$= 2^{30}$
– Tera	$= 10^{12}$	$= 2^{40}$
– Peta	$= 10^{15}$	$= 2^{50}$
– Exa	$= 10^{18}$	$= 2^{60}$
– Zetta	$= 10^{21}$	$= 2^{70}$
– Yotta	$= 10^{24}$	$= 2^{80}$

Large number to the rescue!

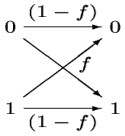
Entropy
Communication



Information theory,
pattern recognition,
and neural networks

Lecture 1:
Introduction to
Information Theory

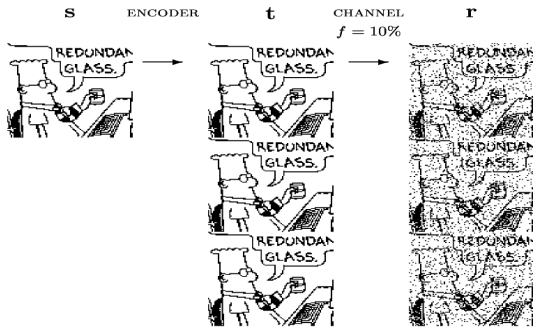
David Mackay



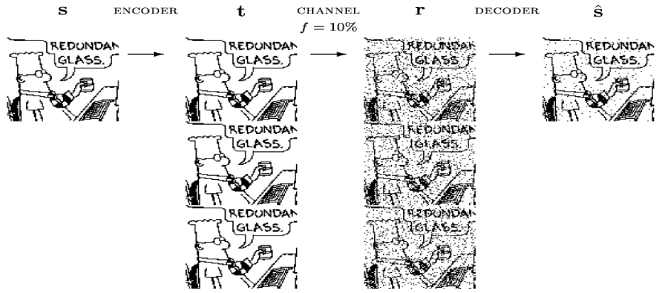
f = 0.1

Error rate 10% mistakes

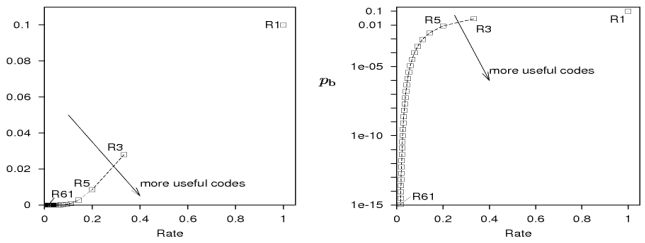
Repetition code 'R3'

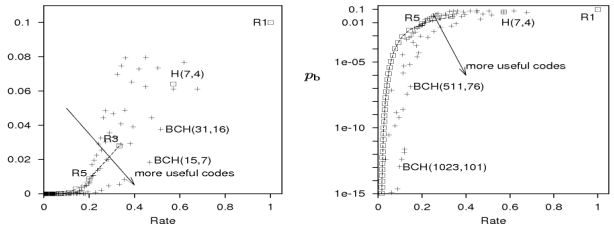


Repetition code 'R3'



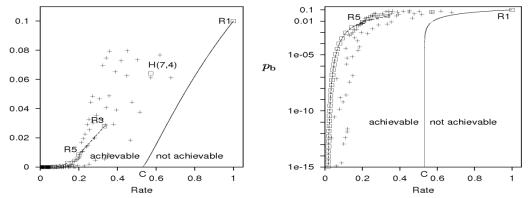
Performance of repetition codes





What's achievable?

Shannon's noisy-channel coding theorem



$$C_{\mathrm{BSC}}(f) = 1 - H_2(f)$$

$$H_2(f) = f \log_2 \frac{1}{f} + (1-f) \log_2 \frac{1}{1-f}$$

If f=1/2 (totally random, no real information), then C=0 (zero rate of transmission)!

Entropy

$$\mathrm{H}[x] = -\sum_x p(x) \log_2 p(x)$$

Important quantity in
- coding theory
- statistical physics
- machine learning

Entropy

- Coding theory: x discrete with 8 possible states; how many bits to transmit the state of x?
- All states equally likely

$$\mathrm{H}[x] = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3 \text{ bits}.$$

Entropy

- In how many ways can N identical objects be allocated M bins?

$$W = \frac{N!}{\prod_i n_i!}$$

$$\mathrm{H} = \frac{1}{N} \ln W \simeq -\lim_{N\to\infty} \sum_i \left(\frac{n_i}{N}\right) \ln \left(\frac{n_i}{N}\right) = -\sum_i p_i \ln p_i$$

- Entropy maximized when

$$\forall i : p_i = \frac{1}{M}$$

Entropy

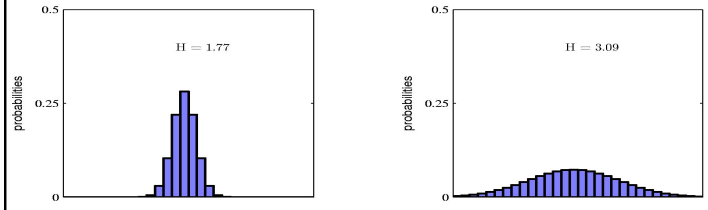
x	a	b	c	d	e	f	g	h
$p(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$
code	0	10	110	1110	111100	111101	111110	111111

$$\begin{aligned} \mathrm{H}[x] &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{4}{64}\log_2\frac{1}{64} \\ &= 2 \text{ bits} \end{aligned}$$

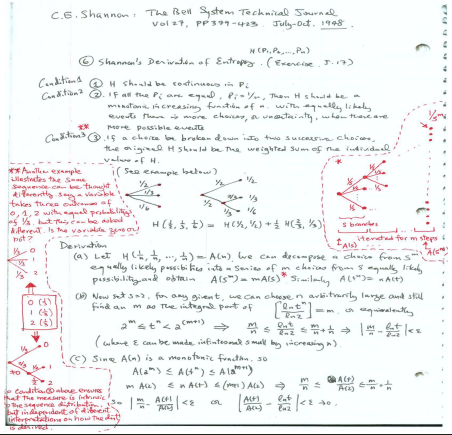
$$\begin{aligned} \text{average code length} &= \frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{16}\times 4 + 4\times\frac{1}{64}\times 6 \\ &= 2 \text{ bits} \end{aligned}$$

If not evenly distributed, the entropy is getting smaller.

Entropy



Learning, in a way, is trying to narrow down the choices of "parameters" (the unknown) using the information obtained in the data set observed, or decreasing the entropy that is consistent with the observation. (also recall that the choice of logic functions for 3-variable problem: from 256 to 8 such functions).



Condition1 (1) H should be continuous in P_i

Condition2 (2). If all the P_i are equal, $P_i = 1/n$, then H should be a monotonic increasing function of n. With equally likely events there is more choices, or uncertainty, when there are more possible events

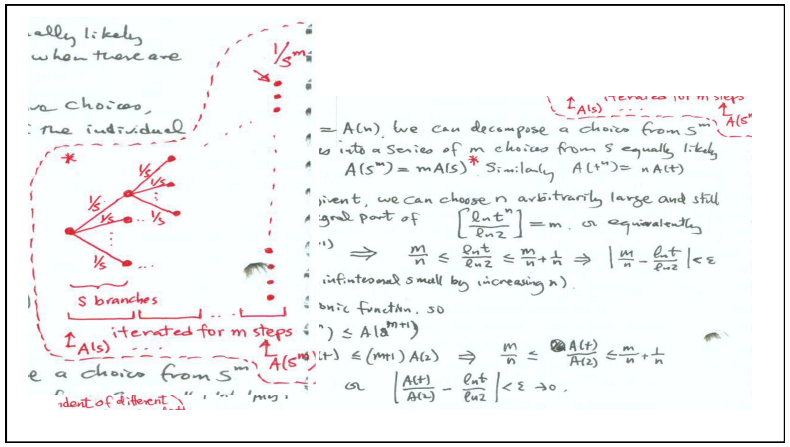
Condition3 (3). If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H.

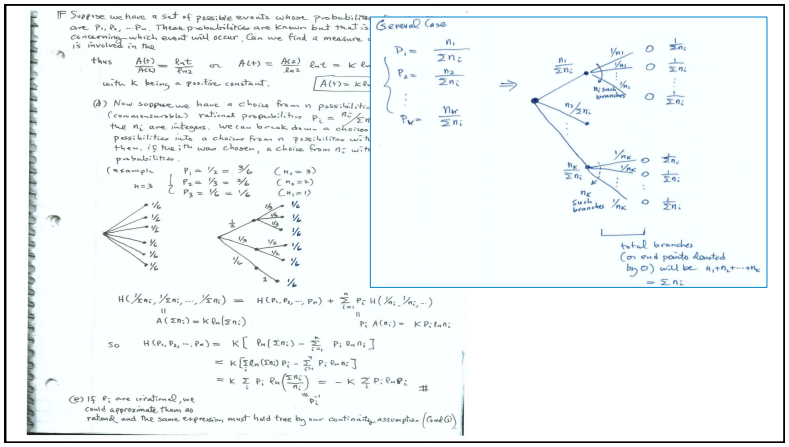
**Another example illustrates the same sequence can be thought differently. Say a variable takes three outcomes of 0, 1, 2 with equal probability of 1/3, but this can be asked different. Is the variable zero or not?

(See example below)

$H(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}) =$

**Another example illustrates the same sequence can be thought differently. Say a variable takes three outcomes of 0, 1, 2 with equal probability of 1/3. but this can be asked different. Is the variable zero or not?





Entropy in ML?

KL Divergence

Foundation for Approximation

KL Divergence

Foundation for Approximation

The relative entropy *or* Kullback–Leibler divergence between two probability distributions $P(x)$ and $Q(x)$ that are defined over the same alphabet $\mathcal{A}_X$ is

$$D_{\mathrm{KL}}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}. \tag{2.45}$$

The relative entropy satisfies *Gibbs' inequality*

$$D_{\mathrm{KL}}(P||Q) \geq 0 \tag{2.46}$$

with equality only if $P = Q$. Note that in general the relative entropy is not symmetric under interchange of the distributions P and Q: in general $D_{\mathrm{KL}}(P||Q) \neq D_{\mathrm{KL}}(Q||P)$, so D_{KL}, although it is sometimes called the 'KL distance', is not strictly a distance. The relative entropy is important in pattern recognition and neural networks, as well as in information theory.

2.7 Jensen's inequality for convex functions

The words 'convex $\smile$' and 'concave $\frown$' may be pronounced 'convex-smile' and 'concave-frown'. This terminology has useful redundancy: while one may forget which way up 'convex' and 'concave' are, it is harder to confuse a smile with a frown.

Convex $\smile$ functions. A function $f(x)$ is *convex* $\smile$ over (a, b) if every chord of the function lies above the function, as shown in figure 2.10; that is, for all $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2). \tag{2.47}$$

A function f is *strictly convex* $\smile$ if, for all $x_1, x_2 \in (a, b)$, the equality holds only for $\lambda = 0$ and $\lambda = 1$.

Similar definitions apply to concave $\frown$ and strictly concave $\frown$ functions.

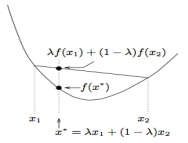


Figure 2.10. Definition of convexity.

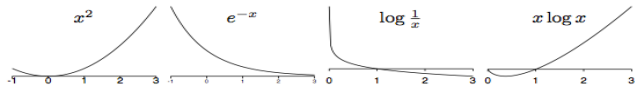
Jensen's inequality. If f is a convex $\smile$ function and x is a random variable then:

$$\mathcal{E}\left[f(x)\right] \geq f(\mathcal{E}[x]), \tag{2.48}$$

where $\mathcal{E}$ denotes expectation. If f is strictly convex $\smile$ and $\mathcal{E}\left[f(x)\right] = f(\mathcal{E}[x])$, then the random variable x is a constant.

Some strictly convex $\smile$ functions are

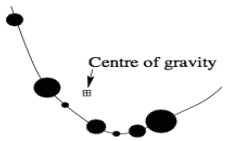
- x^2, e^x and e^{-x} for all x;
- $\log(1/x)$ and $x \log x$ for $x > 0$.



Proof (more of a argument for it)

A physical version of Jensen's inequality runs as follows.

If a collection of masses p_i are placed on a convex $\smile$ curve $f(x)$ at locations $(x_i, f(x_i))$, then the centre of gravity of those masses, which is at $(\mathcal{E}[x], \mathcal{E}\left[f(x)\right])$, lies above the curve.



The relative entropy *or* Kullback–Leibler divergence between two probability distributions $P(x)$ and $Q(x)$ that are defined over the same alphabet $\mathcal{A}_X$ is

$$D_{\rm KL}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}. \tag{2.45}$$

The relative entropy satisfies *Gibbs' inequality*

$$D_{\rm KL}(P||Q) \geq 0 \tag{2.46}$$

1st

$$D_{\rm KL}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}. \tag{2.94}$$

We prove Gibbs' inequality using Jensen's inequality. Let $f(u) = \log 1/u$ and $u = \frac{Q(x)}{P(x)}$. Then

$$\begin{aligned} D_{\rm KL}(P||Q) &= \mathcal{E}[f(Q(x)/P(x))] &(2.95)\\ &\geq f\left(\sum_x P(x)\frac{Q(x)}{P(x)}\right) = \log\left(\frac{1}{\sum_x Q(x)}\right) = 0, &(2.96)\end{aligned}$$

with equality only if $u = \frac{Q(x)}{P(x)}$ is a constant, that is, if $Q(x) = P(x)$. □

The relative entropy *or* Kullback–Leibler divergence between two probability distributions $P(x)$ and $Q(x)$ that are defined over the same alphabet $\mathcal{A}_X$ is

$$D_{\rm KL}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}. \tag{2.45}$$

The relative entropy satisfies *Gibbs' inequality*

$$D_{\rm KL}(P||Q) \geq 0 \tag{2.46}$$

2nd

Second solution. In the above proof the expectations were with respect to the probability distribution $P(x)$. A second solution method uses Jensen's inequality with $Q(x)$ instead. We define $f(u) = u \log u$ and let $u = \frac{P(x)}{Q(x)}$. Then

$$\begin{aligned} D_{\rm KL}(P||Q) &= \sum_x Q(x)\frac{P(x)}{Q(x)} \log \frac{P(x)}{Q(x)} = \sum_x Q(x) f\left(\frac{P(x)}{Q(x)}\right) &(2.97)\\ &\geq f\left(\sum_x Q(x)\frac{P(x)}{Q(x)}\right) = f(1) = 0, &(2.98)\end{aligned}$$

with equality only if $u = \frac{P(x)}{Q(x)}$ is a constant, that is, if $Q(x) = P(x)$. □

8.2.1 The Kullback-Leibler Divergence KL($q|p$)

The Kullback-Leibler divergence KL($q|p$) measures the 'difference' between distributions q and p[68].

Definition 8.11 (KL divergence). For two distributions $q(x)$ and $p(x)$

$$\mathrm{KL}(q|p) \equiv \langle \log q(x) - \log p(x) \rangle_{q(x)} \geq 0 \tag{8.2.28}$$

where $\langle f(x) \rangle_{r(x)}$ denotes average of the function $f(x)$ with respect to the distribution $r(x)$.

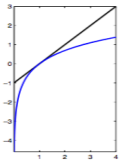
3rd Proof (Chap8 Barber)

The KL divergence is ≥ 0

The KL divergence is widely used and it is therefore important to understand why the divergence is positive.

To see this, consider the following linear bound on the function $\log(x)$

$$\log(x) \leq x - 1 \tag{8.2.29}$$



as plotted in the figure on the right. Replacing x by $p(x)/q(x)$ in the above bound

$$\frac{p(x)}{q(x)} - 1 \geq \log \frac{p(x)}{q(x)} \tag{8.2.30}$$

Since probabilities are non-negative, we can multiply both sides by $q(x)$ to obtain

$$p(x) - q(x) \geq q(x) \log p(x) - q(x) \log q(x) \tag{8.2.31}$$

We now integrate (or sum in the case of discrete variables) both sides. Using $\int p(x)dx = 1, \int q(x)dx = 1$,

$$1 - 1 \geq \langle \log p(x) - \log q(x) \rangle_{q(x)} \tag{8.2.32}$$

Rearranging gives

$$\langle \log q(x) - \log p(x) \rangle_{q(x)} \equiv \mathrm{KL}(q|p) \geq 0 \tag{8.2.33}$$

The KL divergence is zero if and only if the two distributions are exactly the same.

How to use KL Divergence

$$\mathrm{KL}(p\|q) = -\int p(\mathbf{x}) \ln q(\mathbf{x})\,\mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right)$$

$$= -\int p(\mathbf{x}) \ln \left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\}\,\mathrm{d}\mathbf{x}$$

This step is an approximation that gets better and better as N gets larger and larger.

$$\mathrm{KL}(p\|q) \simeq \frac{1}{N}\sum_{n=1}^{N}\{-\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x}_n)\}$$

$$\mathrm{KL}(p\|q) \geqslant 0$$

Howto "Try my best!"
Variational Method (Mean Field Approximation) etc …