Lecture 5: Information theory

Data Science and Information Theory, ED127 course (2025)

Florent Leclercq

www.florent-leclercq.eu

Institut d'Astrophysique de Paris
CNRS & Sorbonne Université

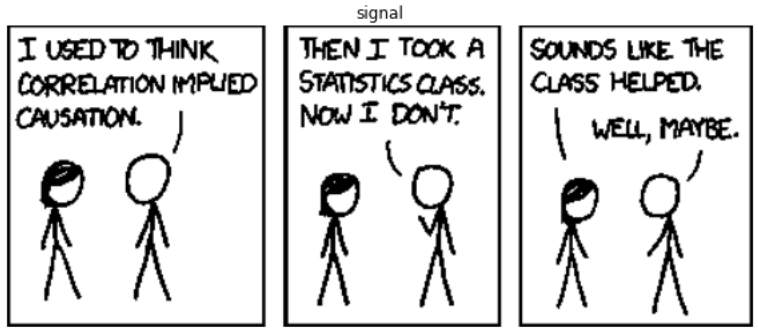
8 April 2025

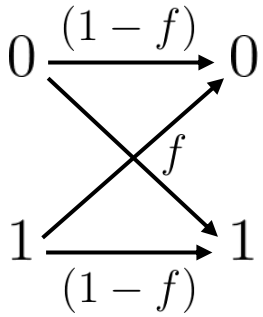
Prince William Sound, Alaska

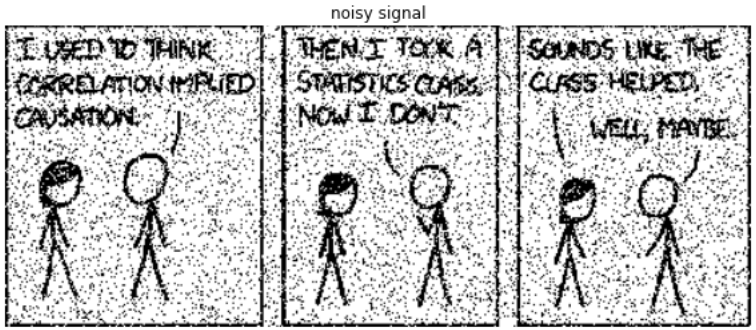
05 NOISY-CODING CHANNEL

THE NOISY BINARY SYMMETRIC CHANNEL

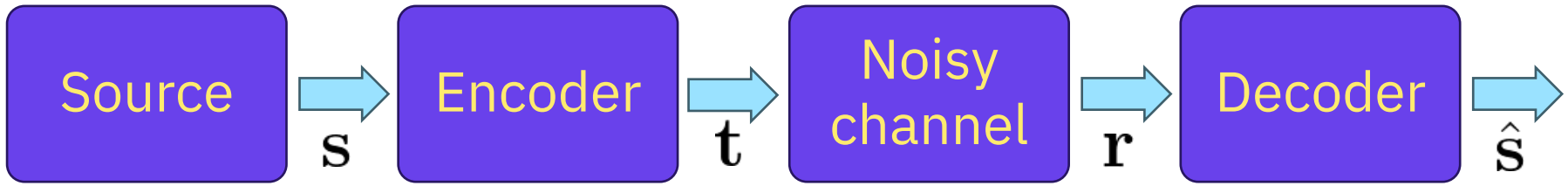
The noisy binary symmetric channel







https://xkcd.com/552/



Rate of information transfer: $R = \frac{\#\mathbf{s}}{\#\mathbf{t}} = \frac{K}{N}$

The R3 code

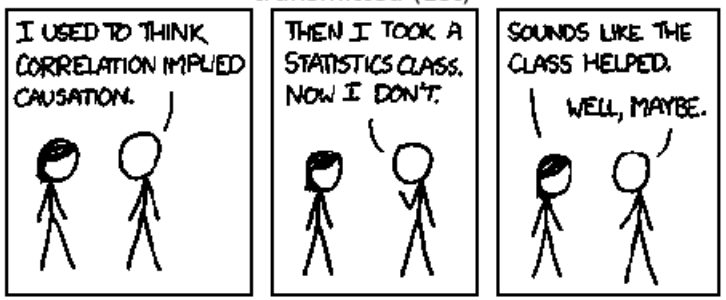
$\mathbf{s}$	0	0	1	0	1	1	0
$\mathbf{t}$	000	000	111	000	111	111	000
$\mathbf{n}$	000	001	000	000	101	000	000
$\mathbf{r}$	000	001	111	000	010	111	000
$\hat{\mathbf{s}}$	0	0	1	0	0	1	0
corrected errors		$\star$					
undetected errors					$\star$		

Rate of information transfer: $R[R_3] = \frac{1}{3}$

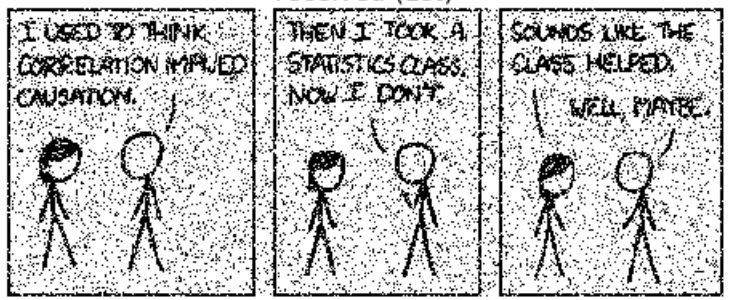
MacKay (2003), chap. 1

The R3 code: example

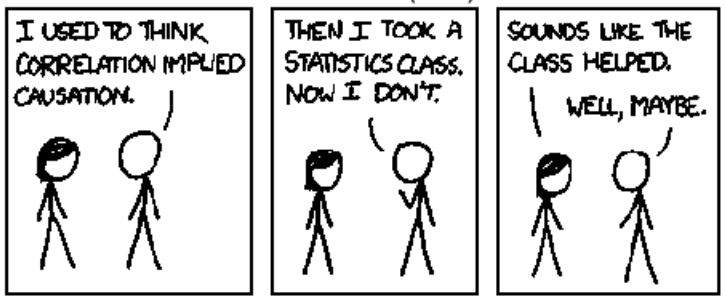
transmitted (1st)



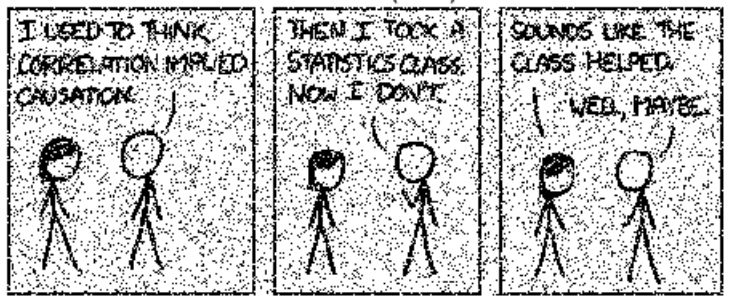
received (1st)



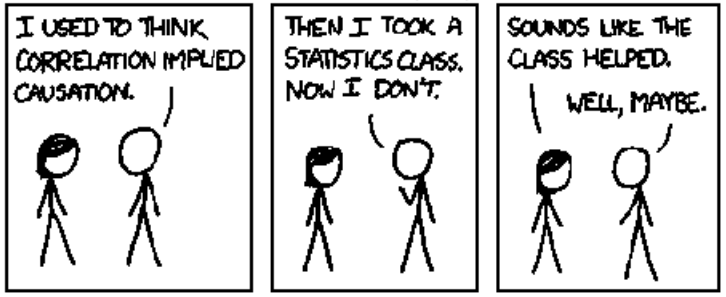
transmitted (2nd)



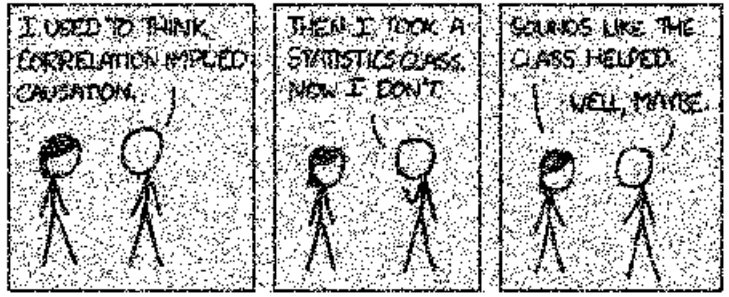
received (2nd)



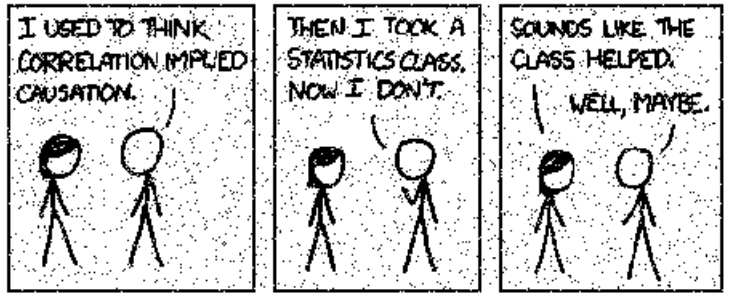
transmitted (3rd)



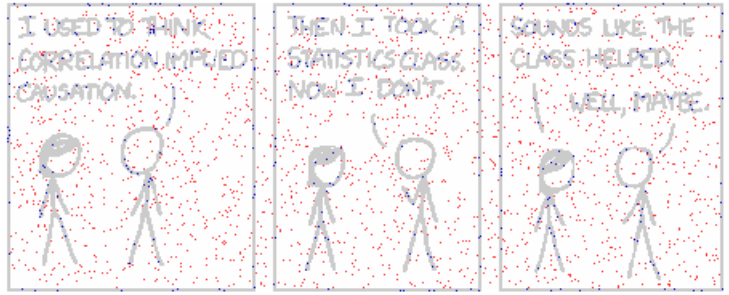
received (3rd)



decoded

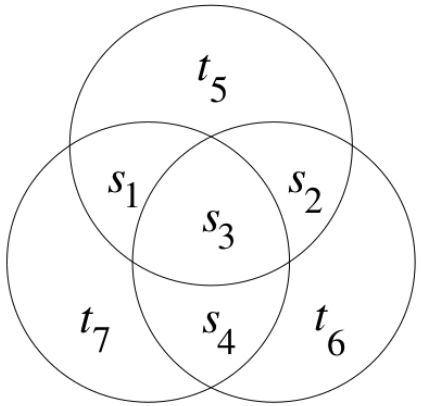


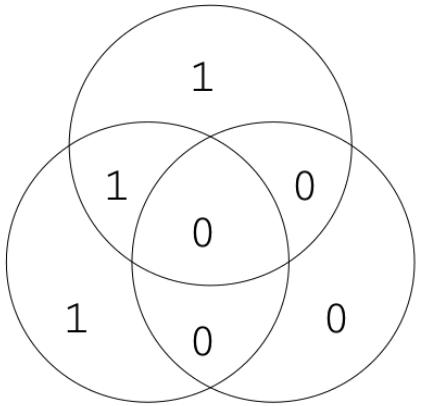
uncorrected errors



The (7,4) Hamming code: encoder

- Introducing the concept of parity-checks:





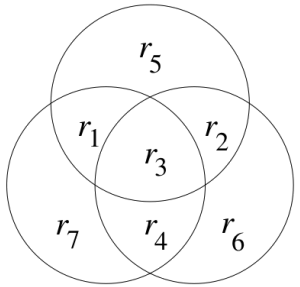
$$\mathbf{t} = \mathbf{G}\mathbf{s} \qquad \mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}$$

Rate of information transfer: $R[H(7,4)] = \frac{4}{7}$

MacKay (2003), chap. 1

The (7,4) Hamming code: decoder

- Introducing the concept of syndrome:



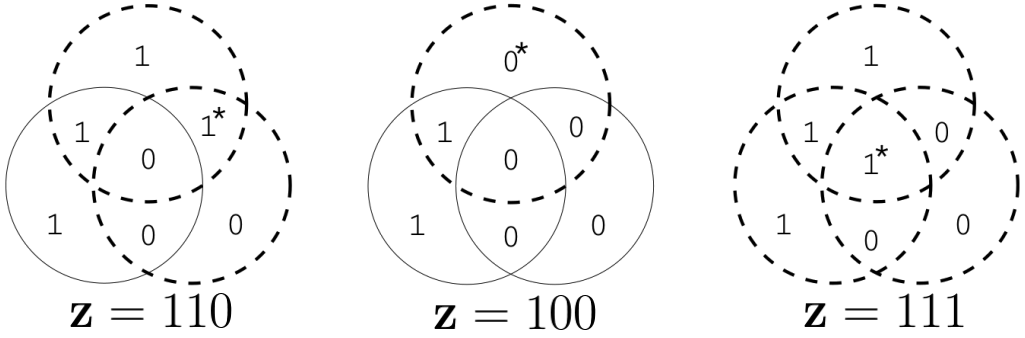
syndrome

$$\mathbf{z} = \mathbf{Hr}$$

$$\mathbf{H} = \left[\begin{array}{cccc|ccc} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right]$$

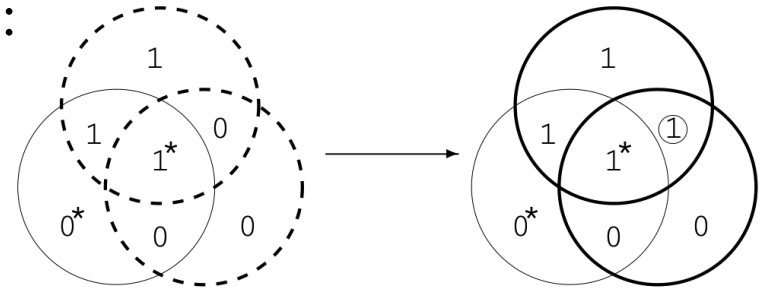
parity-checks identity

- Error correction:



Syndrome $\mathbf{z}$	000	001	010	011	100	101	110	111
Unflip this bit	None	r_7	r_6	r_4	r_5	r_1	r_2	r_3

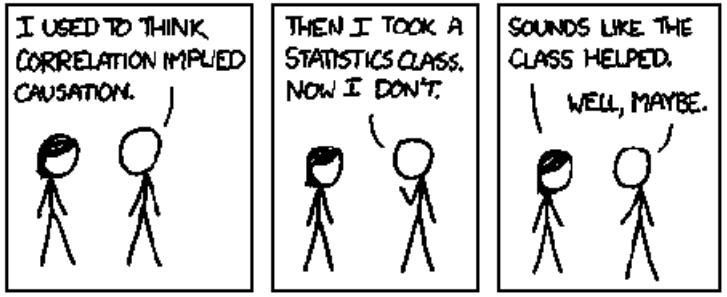
- Unsuccessful error correction:



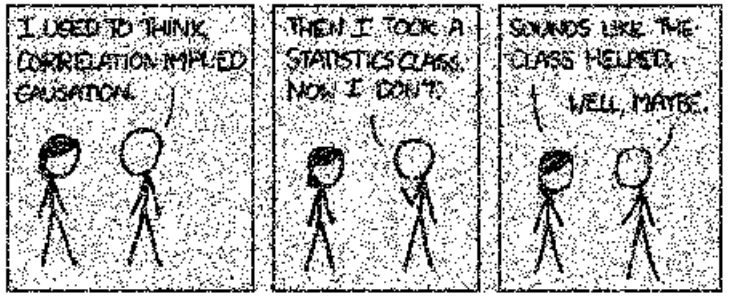
MacKay (2003), chap. 1

The (7,4) Hamming code: example

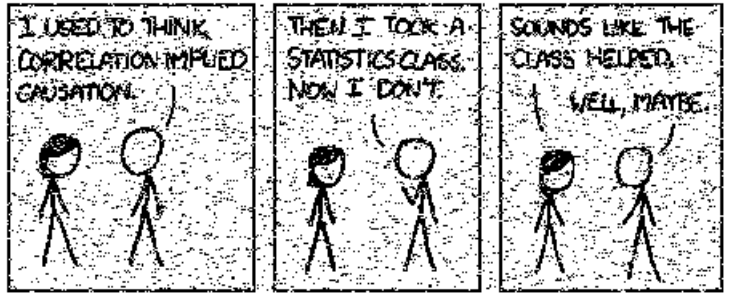
transmitted image



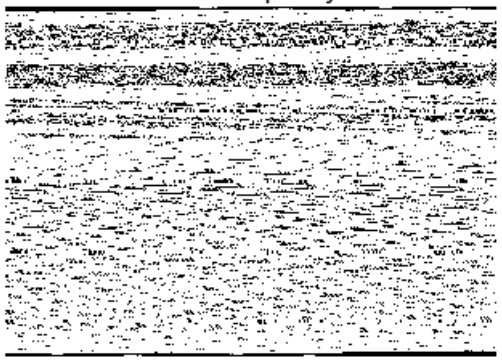
received noisy image



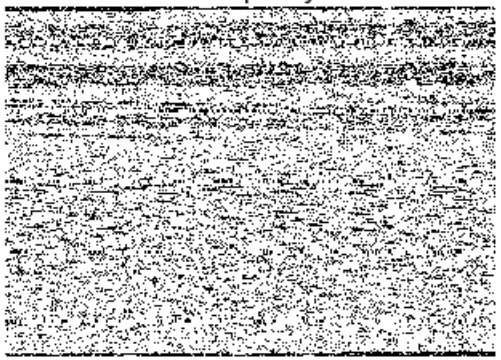
decoded



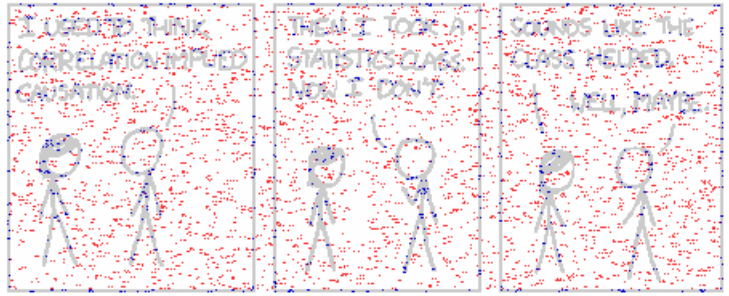
transmitted parity bits



received parity bits

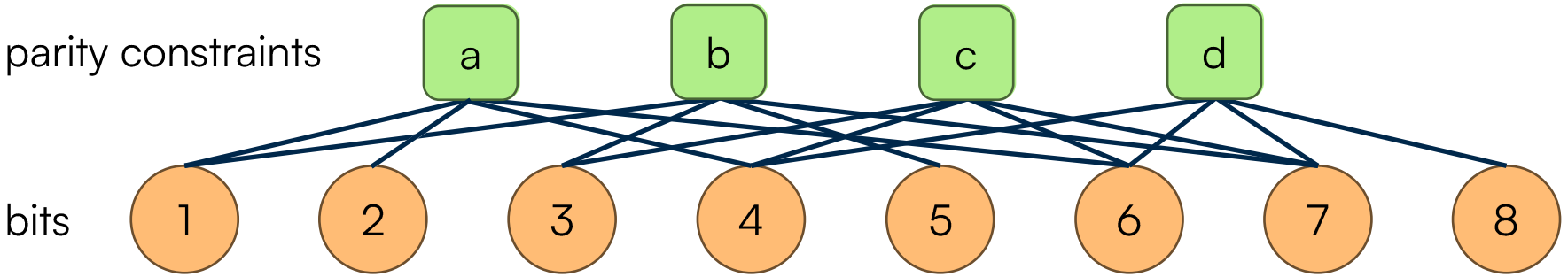


uncorrected errors



Low-density parity check (LDPC) codes

- Tanner graph:



- N bits, $M = (1 - R)N$ parity-check constraints
  - 2^{RN} possible "words": the dictionary
  - A sparse parity-check matrix
- Decoding LDPC codes: the general theory borrows from statistical physics: Ising spins in interaction and the BP mean field approximation (Bethe-Peierls — Belief Propagation)

MacKay (2003), chap. 47

SHANNON'S NOISY-CHANNEL CODING THEOREM

Shannon's noisy-channel coding theorem

- Rate of information transfer $R[R_N]$ and probability of error p_b of repetition codes (for odd N):

$$R[R_N] = \frac{1}{N} \qquad p_\mathrm{b} = \sum_{n=(N+1)/2}^{N} \binom{N}{n} f^n (1-f)^{N-n}$$

- Definitions:
  - Binary entropy function:

$$H_2(x) \equiv x \log_2 \frac{1}{x} + (1-x) \log_2 \frac{1}{1-x}$$

  - Capacity of the noisy channel:

$$C(f) \equiv 1 - H_2(f)$$

  - Shannon's limit:

$$R = \frac{C(f)}{1 - H_2(f)}$$

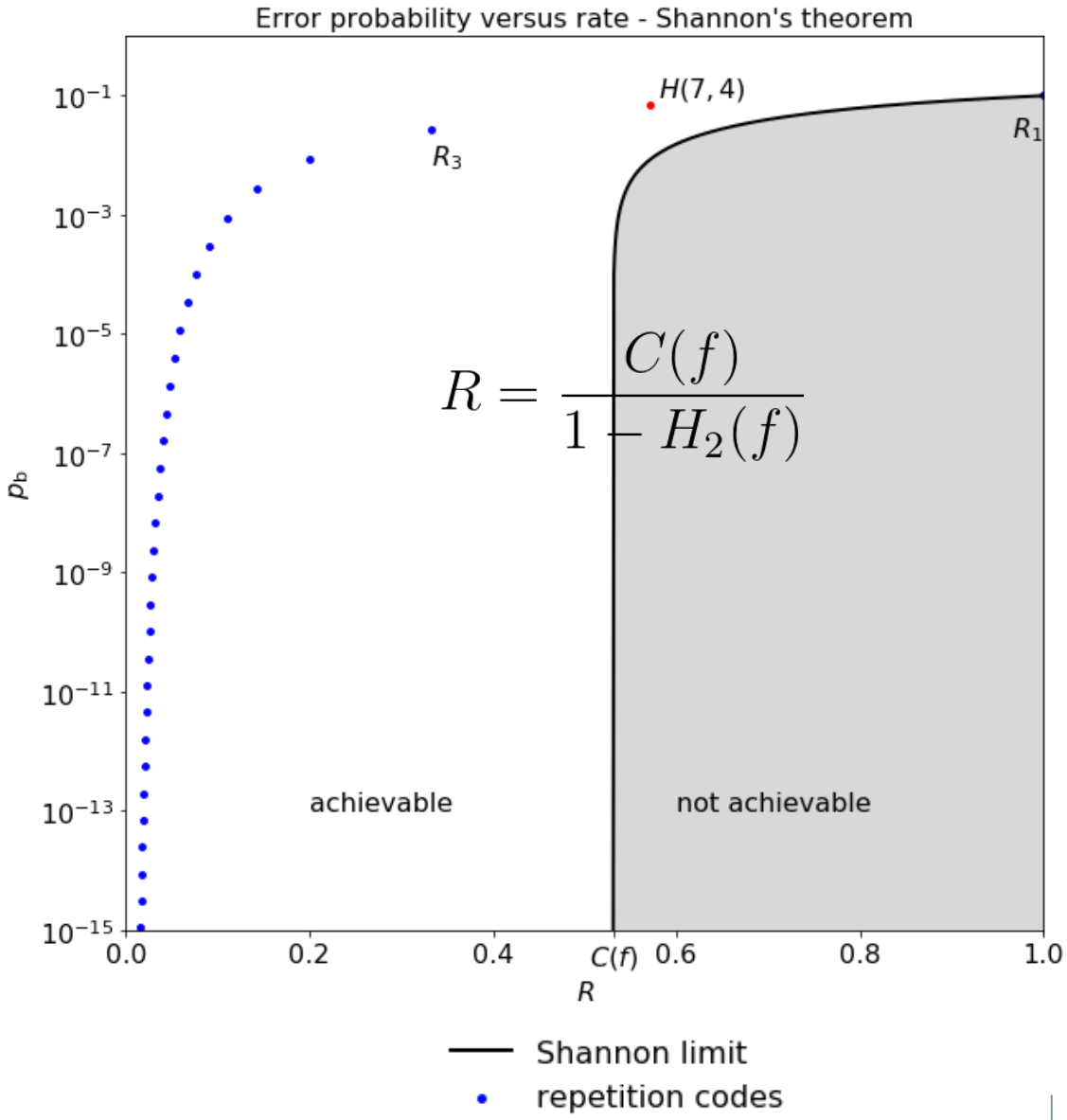
MacKay (2003), chap. 10

Shannon's noisy-channel coding theorem

- Shannon's noisy-channel coding theorem (1948): rates above the Shannon limit are achievable, rates below the Shannon limit are not achievable.

- The boundary between achievable and non-achievable points meets the R-axis at a non-zero value $R = C(f)$. We don't necessarily have $R \to 0$ if we want $p_\mathrm{b} \to 0$.

- Rates up to $R = C(f)$ (the capacity of the channel) are achievable with arbitrarily small probability of error p_b.



MacKay (2003), chap. 10

05 INFORMATION AND ENTROPY

MEASURES OF ENTROPY AND INFORMATION

Information content, entropy, and conditional entropy

- Information content:

$$I[X] \equiv -\sum_{x \in \mathcal{X}} \log_2 p(x)$$

- Entropy:

$$H[X] \equiv \langle I[X] \rangle_{p(X)} = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$

- Conditional entropy:

$$H[X|Y] \equiv \langle H[X|Y=y] \rangle_{p(Y)} = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y) \log_2 \left(\frac{p(y)}{p(x,y)} \right)$$

- Properties:
  - chain rule:

$$H[X|Y] = H[X,Y] - H[Y]$$

  - "Bayes's theorem":

$$H[X|Y] + H[Y] = H[Y|X] + H[X]$$

Crooks (2016); FL *et al.* (2016), 1606.06758 (Appendix A)

Mutual information and relative entropy

- Mutual information:

$$I[X:Y] \equiv \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

- Properties:
  - $I[X:Y] \geq 0$
  - $I[X:X] = H[X]$
  - $I[X:Y] = H[X] - H[X|Y]$
    $= H[Y] - H[Y|X]$
    $= H[X] + H[Y] - H[X,Y]$
    $= H[X,Y] - H[X|Y] - H[Y|X]$
  - Consequence: $H[X|Y] \leq H[X]$
- Relative entropy/Kullback-Leibler divergence/Information gain:

$$D_{\mathrm{KL}}[p||q] \equiv \sum_{x\in\mathcal{X}} p(x)\log_2\left(\frac{p(x)}{q(x)}\right)$$

- Properties:
  - Gibbs's inequality: $D_{\mathrm{KL}}[p||q] \geq 0$
  - Relation to mutual information: $I[X:Y] = D_{\mathrm{KL}}[p(x,y)||p(x)p(y)] = \langle D_{\mathrm{KL}}[p(x|y)||p(x)]\rangle_{p(Y)}$

Crooks (2016); FL *et al.* (2016), 1606.06758 (Appendix A)

Symmetrisation

- Symmetrisation procedures:
  - Jeffreys symmetrisation: $C_s[A:B] = \frac{1}{2}C_a[A||B] + \frac{1}{2}C_a[B||A]$
  - Jensen symmetrisation: $C_s[A:B] = \frac{1}{2}C_a[A||M] + \frac{1}{2}C_a[B||M] \quad \text{with} \quad M = \frac{A+B}{2}$
- Jeffreys divergence:

  $$D_{\mathrm{J}}[p:q] = \frac{1}{2}D_{\mathrm{KL}}[p||q] + \frac{1}{2}D_{\mathrm{KL}}[q||p]$$
- Jensen-Shannon divergence:

  $$D_{\mathrm{JS}}[p:q] = \frac{1}{2}D_{\mathrm{KL}}[p||r] + \frac{1}{2}D_{\mathrm{KL}}[q||r] \quad \text{with} \quad r \equiv \frac{p+q}{2}$$
- Properties:
  - $D_{\mathrm{JS}}[p:q] = H[r] - \frac{1}{2}H[p] - \frac{1}{2}H[q]$
  - $0 \leq D_{\mathrm{JS}}[p:q] \leq 1$
  - $D_{\mathrm{JS}}[p:q] = I[m:z] \quad \text{with} \quad m \equiv z\,p + (1-z)\,q \qquad z = \begin{cases} 0 & (p = 1/2) \\ 1 & (p = 1/2) \end{cases}$

Crooks (2016); FL *et al.* (2016), 1606.06758 (Appendix A)

INFORMATION-THEORETIC EXPERIMENTAL DESIGN

Information-theoretic experimental design

- Back to Bayesian experimental design. The general problem is to maximise

(expected) data experimental design

$$U(\xi) = \langle U(d, \xi) \rangle_{p(d|\xi)} = \int p(d|\xi) U(d, \xi) \, \mathrm{d}d$$

- We can use information theory to write down utility functions for three kinds of problems:

PARAMETER INFERENCE **MODEL COMPARISON** **PREDICTION**

FL *et al.* (2016), 1606.06758 (Appendix A)

Information-theoretic experimental design for parameter inference

1. Parameter inference utility functions:

   - Maximal information gain:

$$U(d,\xi) \equiv D_{\mathrm{KL}}[p(\theta|d,\xi)||p(\theta|\xi)] \qquad\qquad U(\xi) = I[\theta:d|\xi]$$

   - A-optimality:

$$U_{\mathrm{A}}(d,\xi) \equiv \frac{1}{\mathrm{tr}(\mathrm{cov}(\theta|d,\xi)^{-1})}$$

   - D-optimality:

$$U_{\mathrm{D}}(d,\xi) \equiv \det(\mathrm{cov}(\theta|d,\xi))$$

Bassett (2004), astro-ph/0407201; FL *et al*. (2016), 1606.06758 (Appendix A)

Information-theoretic experimental design for model selection

2. Model selection utility functions:

- Maximal Bayes factor:

$$U(\xi) \equiv \mathcal{B}_{12}(\xi) = \frac{p(d|\xi, \mathcal{M}_1)}{p(d|\xi, \mathcal{M}_2)}$$

It is hard to predict Bayes factors…
but see Trotta (2007), astro-ph/0703063

- Maximal mutual information between the model indicator and the data:

$$U(\xi) \equiv I[\mathcal{M} : d|\xi]$$

e.g. Cavagnaro *et al.* (2010)

- Maximal Jensen-Shannon divergence between posterior predictive distributions:

$$U(\xi) \equiv D_{\mathrm{JS}}[p_1 : p_2] = I[\mathcal{M} : r|\xi] \quad \text{with} \quad r \equiv \frac{p_1 + p_2}{2}$$

Vanlier *et al.* (2014), FL *et al.* (2016), 1606.06758 (Appendix A)

FL *et al.* (2016), 1606.06758 (Appendix A)

Information-theoretic experimental design for prediction of future observations

3. Utilities for prediction of future observations:

$$U(d,\xi) \equiv D_{\mathrm{KL}}[p(t|d,\xi)||p(t|\xi)] \qquad U(\xi) = I[t:d|\xi] = H[p(t|\xi)] - H[p(t|d,\xi)]$$

- i.e. the "supervised machine learning" utility: $U(a) = H[T] - H[T|a]$

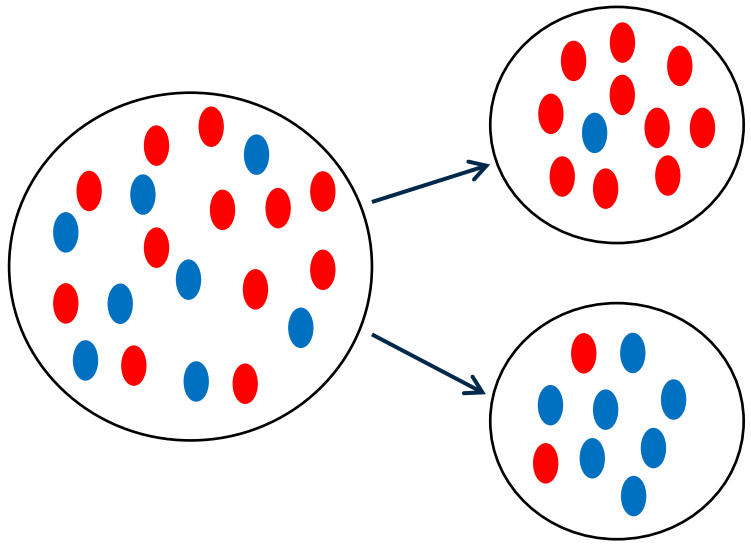
FL *et al.* (2016), 1606.06758 (Appendix A)

SUPERVISED MACHINE LEARNING BASICS

Supervised machine learning basics

Exercise: Titanic example

- How is the information gain computed? $U(a) = H[T] - H[T|a]$



child1 entropy:

$$H = -\frac{10}{11}\log_2\left(\frac{10}{11}\right) - \frac{1}{11}\log_2\left(\frac{1}{11}\right) = 0.4395$$

child2 entropy:

$$H = -\frac{7}{9}\log_2\left(\frac{7}{9}\right) - \frac{2}{9}\log_2\left(\frac{2}{9}\right) = 0.7642$$

parent entropy:

$$H = -\frac{8}{20}\log_2\left(\frac{8}{20}\right) - \frac{12}{20}\log_2\left(\frac{12}{20}\right) = 0.9709$$

weighted average entropy of children:

$$\frac{11}{20} \times 0.4395 + \frac{9}{20} \times 0.7642 = 0.5856$$

information gain for this split: $0.9709 - 0.5856 = 0.3853$ Sh

05 INFORMATION GEOMETRY

Approximating the learning process as a geometrical process

- If entropy is the right information measure, then there is something fundamentally irreversible in statistical learning (going from the prior to the posterior): the information gain is asymmetric:

$$D_{\mathrm{KL}}[p(\theta|d)||p(\theta)] = \int p(\theta|d)\log_2\left(\frac{p(\theta|d)}{p(\theta)}\right)\mathrm{d}\theta \neq D_{\mathrm{KL}}[p(\theta)||p(\theta|d)]$$

- However, for small perturbations in the parameters $\theta' = \theta + \delta\theta$, it is possible to show that close to the peak of any likelihood $p(d|\theta)$,

$$D_{\mathrm{KL}}[p(d|\theta')||p(d|\theta)] = \frac{1}{2}\delta\theta^{\intercal} I(\theta)\delta\theta + o(||\delta\theta^2||) = D_{\mathrm{KL}}[p(d|\theta)||p(d|\theta')]$$

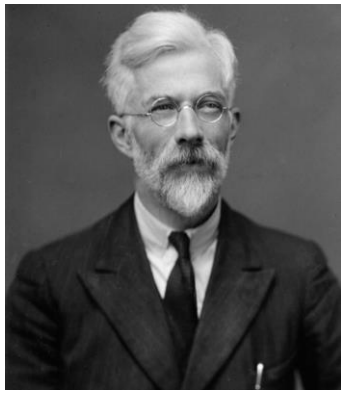
  which is symmetric, where $I(\theta)$ is the Fisher information matrix.

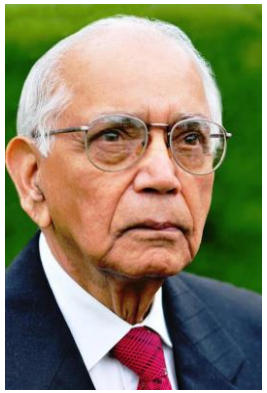
- Can you show this?
  - Hints: Taylor-expand the log-likelihood $\log_2 p(d|\theta)$ to second-order around θ. Substitute this into the KL divergence and use known properties of expectations of the log-likelihood to perform the integrals.
- Therefore, the Fisher information matrix locally defines a metric: the Fisher-Rao metric (1945).

The Fisher-Rao metric in information geometry

- The Fisher information matrix defines a metric, the Fisher-Rao metric.
- $I(\theta)_{\alpha\beta}$ measures the curvature of the log-likelihood surface: steeper curvature implies better parameter constraint.
  - Diagonal elements, $I(\theta)_{\alpha\alpha}$, relate to the sensitivity of the log-likelihood to changes in θ_α .
  - Off-diagonal elements capture correlations between estimates of the parameters.
- In information geometry, all of the notions appearing in differential geometry on curved spaces (or general relativity) have their equivalent: geodesics, parallel transport, Ricci curvature, Christoffel symbols, etc.

(...) I suggested the ***differential geometric approach*** in my 1945 paper by considering the space of probability distributions. I used Fisher information matrix in defining the metric, so it was called Fisher-Rao metric. Differential geometry was not well known at that time, and in order to compute the geodesic distance from the metric, I had to learn the mathematics from papers on ***relativity describing Einstein metric***. It was only 30 years later, my work received attention (...).
Calyampudi Radhakrishna Rao, Scholarpedia


Ronald Aylmer Fisher (1890-1962)


Calyampudi Radhakrishna Rao (1920-2023)

Nielsen, 1808.08271

The Fisher-Rao distance

- The Fisher-Rao distance measures the geodesic distance between probability distributions on a statistical manifold.
- If the Fisher information matrix $I(\theta)$ is constant, it is defined as:

$$d_{\mathrm{FR}}(\theta_0, \theta_1) \equiv \sqrt{(\theta_0 - \theta_1)^{\intercal} I(\theta)(\theta_0 - \theta_1)}$$

- The metric is derived from the local sensitivity (curvature) of the likelihood function, reflecting an intrinsic Riemannian geometry on the space of parameters.

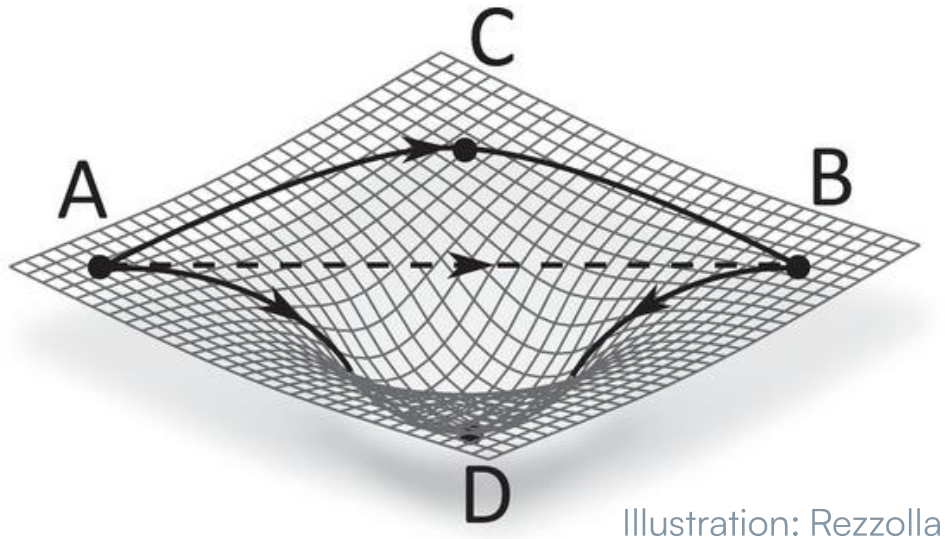


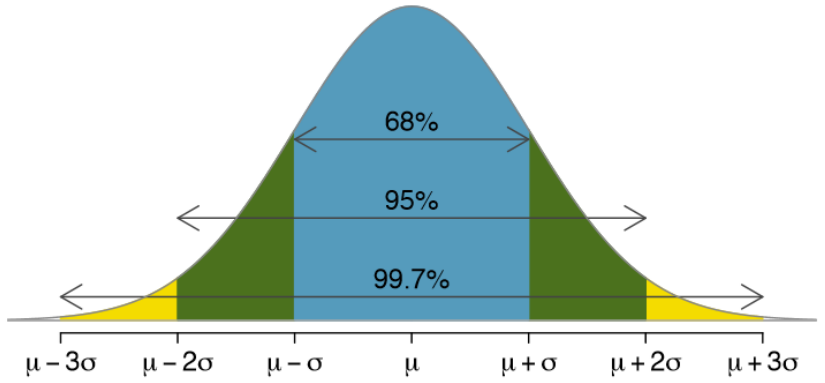
Illustration: Rezzolla (2023)

- There exists some limited analytical results (e.g. the Fisher-Rao distance between two Gaussian distributions).
- The Fisher-Rao distance is widely used in information geometry, statistical inference, and machine learning to compare and analyse the structure of probability distributions.

The Mahalanobis distance

- Given a (Gaussian, or approximated by a Gaussian) probability distribution Q with mean μ and (positive semi-definite) covariance matrix C,
  - the Mahalanobis distance from a point x from Q is:
    $$d_{\mathrm{M}}(x, Q) = \sqrt{(x-\mu)^{\intercal} C^{-1} (x-\mu)}$$
  - the Mahalanobis distance between two points x and y with respect to Q is:
    $$d_{\mathrm{M}}(x, y; Q) = \sqrt{(x-y)^{\intercal} C^{-1} (x-y)}$$
  - which implies:
    $$d_{\mathrm{M}}(x, Q) = d_{\mathrm{M}}(x, \mu; Q)$$

Prasanta Chandra Mahalanobis (1893-1972)

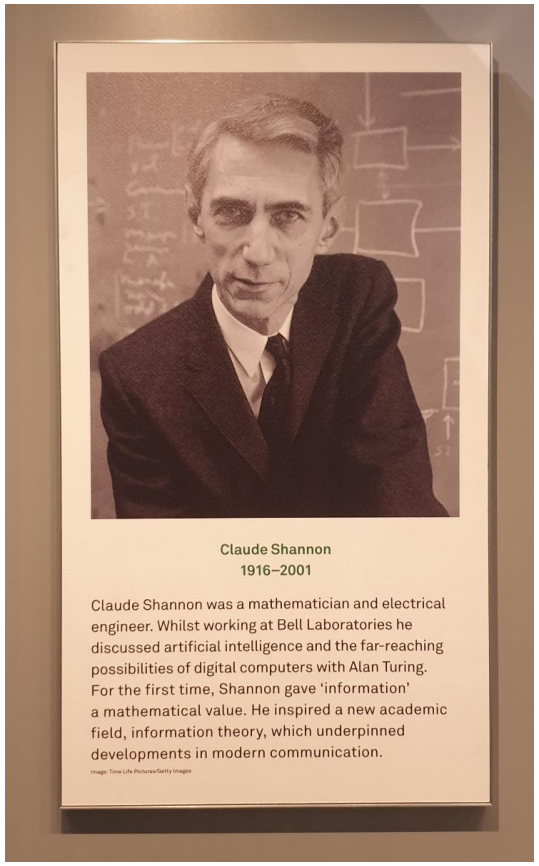


- The Mahalanobis distance is commonly used in classification, outlier detection, and clustering, particularly when the features have different scales or are correlated.
- It accounts for the correlations and variability in the data, effectively standardising the space.
- When C is the identity matrix, the Mahalanobis distance reduces to the standard Euclidean distance.

05

THERMODYNAMICS AND INFERENCE

Physical and information-theoretic entropy



- Physical entropy:

$$S = k_\mathrm{B} \ln W$$ for W equiprobable "microstates"

Boltzmann (1877)

- Information-theoretic entropy:

$$H[p] \propto -\sum_n p_n \log_2 p_n$$ with $p_n = \frac{1}{N}$ for equiprobable events

Shannon (1948)

Why don't you call it entropy? In the first place, a mathematical development very much like yours already exists in Boltzmann's statistical mechanics, and in the second place, no one understands entropy very well, so in any discussion you will be in a position of advantage.
von Neumann to Shannon, about a name for "missing information"

$$1\,\mathrm{nat} = k_\mathrm{B}\,\mathrm{J/K} \iff 1\,\mathrm{Sh} = (k_\mathrm{B} \ln 2)\,\mathrm{J/K} \iff 1\,\mathrm{Sh} = (\ln 2)\,\mathrm{nats}$$

$$k_\mathrm{B} = 1.380649 \times 10^{-23}\mathrm{J/K} \quad \text{(exact value)}$$

$$S_\mathrm{nats} = \frac{S_\mathrm{J/K}}{k_\mathrm{B}} \iff S_\mathrm{Sh} = \frac{S_{J/K}}{k_\mathrm{B} \ln 2}$$

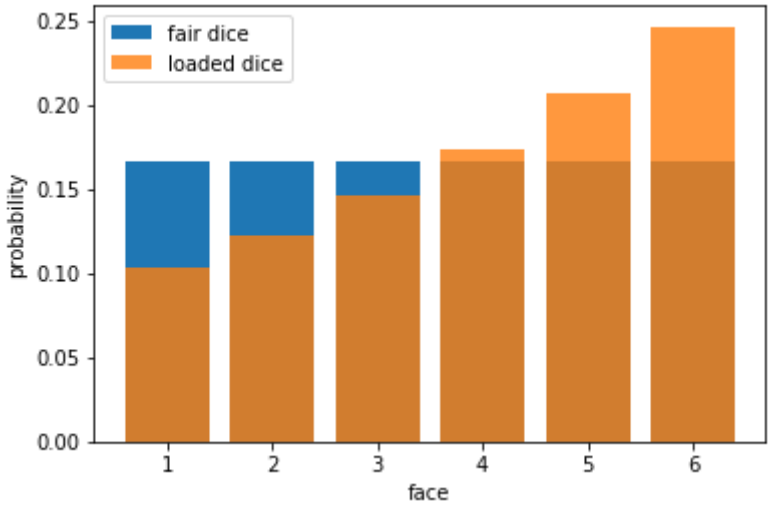
Maximum entropy principle in thermodynamics and in statistics

- Using the maximum entropy principle with the loaded dice example, we have already seen a thermodynamics analogy:
  - Fair dice = microcanonical ensemble: $p_n = \frac{1}{N}$
  - Loaded dice (with fixed mean) = canonical ensemble:

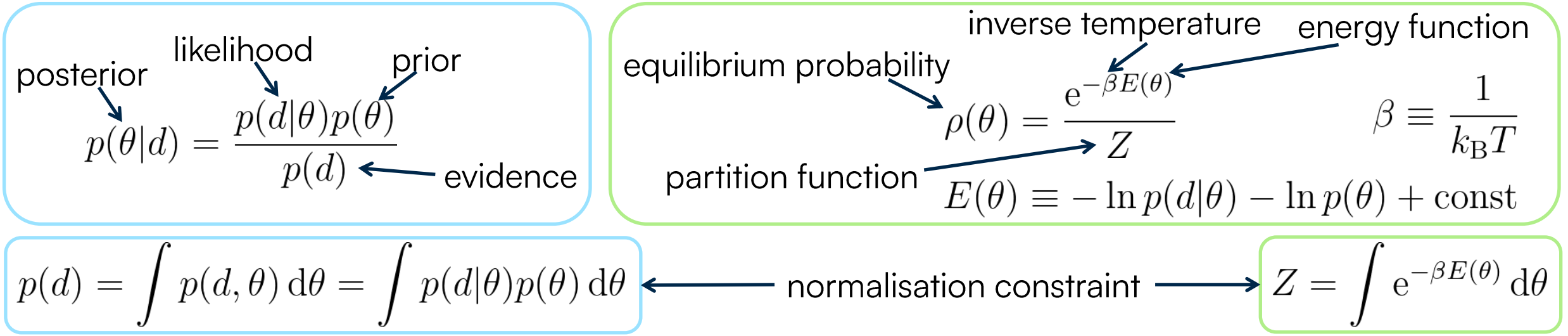
$$p_n = \frac{\mathrm{e}^{-\beta E_n}}{Z} \qquad \beta \equiv \frac{1}{k_\mathrm{B} T}$$

E_n = energy of different states

Z = partition function ≡ evidence in Bayesian statistics



Bayes' theorem in thermodynamic and statistical notations



- Comments on the analogy:
  - Updating beliefs vs. reaching equilibrium:
    Just as Bayesian inference updates beliefs from prior to posterior in light of data, a thermodynamic system relaxes to an equilibrium distribution determined by the energy landscape.
  - Energy and information:
    The mapping $E(\theta) \equiv -\ln p(d|\theta) - \ln p(\theta) + \mathrm{const}$ reveals that lower "energy" configurations correspond to more probable hypotheses, paralleling the principle that systems occupy lower energy states.
  - Normalisation:
    The partition function Z in thermodynamics plays the same role as the evidence $p(d)$ in Bayesian inference: both serve as normalisation constants ensuring the total probability sums to one.

Adjusting the statistical power of a likelihood

- Generalising Bayes' theorem:

$$p(d|\theta)^{\beta} p(\theta) \xrightarrow{\beta \to 0} p(\theta)$$
$$p(d|\theta)^{\beta} p(\theta) \xrightarrow{\beta \to 1} \frac{p(\theta|d)}{p(d)} \propto p(\theta|d)$$
$$\beta \equiv \frac{1}{k_{\mathrm{B}} T}$$

$$\frac{1}{k_{\mathrm{B}} T} \ln p(d|\theta) + \ln p(\theta) + \text{const} \xrightarrow{\beta \to 0} \ln p(\theta) + \text{const}$$
$$\frac{1}{k_{\mathrm{B}} T} \ln p(d|\theta) + \ln p(\theta) + \text{const} \xrightarrow{\beta \to 1} \ln p(\theta|d) + \text{const}$$

- Using this analogy,
  - one can start the statistical analysis (e.g. MCMC run) with the likelihood in a "hot" state, so that the pdf analysed is close to the prior, and progressively "cool down" the likelihood to reach the posterior.
  - or, if one want to intentionally reduce the statistical power of the data, without changing the model (e.g. due to concerns of model misspecification), one can "heat up" the likelihood (see Jasche & Lavaux 2019, 1806.11117 for an example).

Thermodynamic-Bayesian correspondence

Thermodynamics			**Statistics**	
Quantity:	**Interpretation:**		**Quantity:**	**Interpretation:**
$\beta = T^{-1}$	Inverse temperature	$\leftrightarrow$	N	Sample size
$\boldsymbol{\theta}$	State variables/vector	$\leftrightarrow$	$\boldsymbol{\theta}$	Model parameters
X^N	Quenched disorder	$\leftrightarrow$	X^N	Observations
$E_X(\boldsymbol{\theta})$	State energy	$\leftrightarrow$	$\hat{H}_X(\boldsymbol{\theta})$	Cross entropy estiamtor
E_0	Disorder-averaged ground state energy	$\leftrightarrow$	H_0	Shannon entropy
$\rho(\boldsymbol{\theta})$	Density of states	$\leftrightarrow$	$\varpi(\boldsymbol{\theta})$	Prior
Z	Partition function	$\leftrightarrow$	Z	Evidence
$Z^{-1}\rho\ \exp -\beta E_X$	Normalized Boltzmann weight	$\leftrightarrow$	$\varpi(\boldsymbol{\theta}\|X^N)$	Posterior
$F = -\beta^{-1}\log Z$	Free energy	$\leftrightarrow$	$F = -N^{-1}\log Z$	Minus-log-evidence
$U = \partial_\beta \beta F$	Average energy	$\leftrightarrow$	$U = \partial_N N F$	Minus-log-prediction
$C = -\beta^2 \partial_\beta^2 \beta F$	Heat capacity	$\leftrightarrow$	$C = -N^2 \partial_N^2 N F$	*Learning capacity*
$S = \beta^2 \partial_\beta F$	Gibbs entropy	$\leftrightarrow$	$S = N^2 \partial_N F$	*Statistical Gibbs entropy*

Enßlin & Weig, 1004.2868; LaMont & Wiggins, 1706.01428

References and acknowledgements

References:

- D. MacKay (2003), *Information Theory, Inference, and Learning Algorithms*
- G. E. Crooks (2016), *On measures of entropy and information*
- Leclercq *et al.* (2016), 1606.06758, *Comparing cosmic web classifiers using information theory* (Appendix A)
- F. Nielsen (2018), 1808.08271, *An elementary introduction to information geometry*

- For his lectures, thanks to Torsten Enßlin

https://florent-leclercq.eu/teaching.php