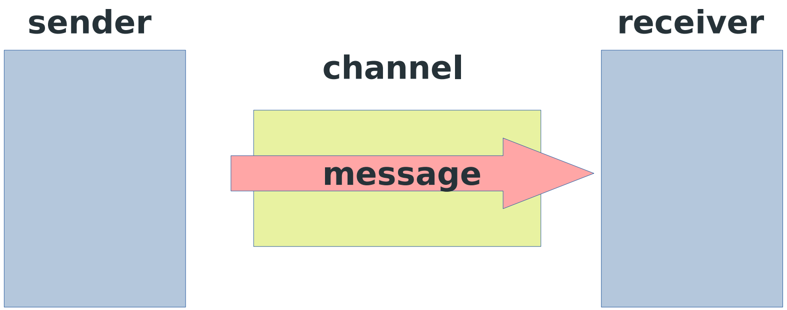
Michele Allegra

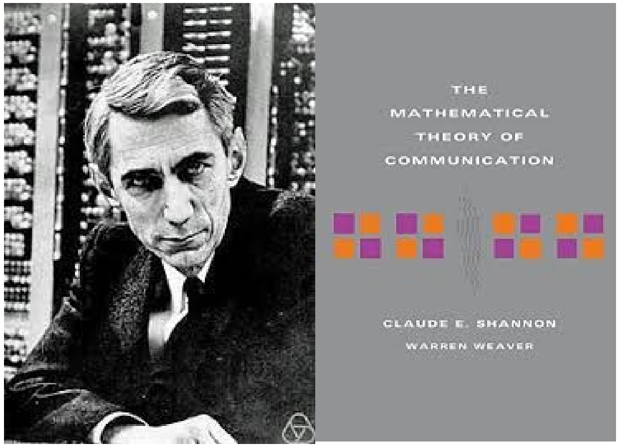
Information theory and Inference

Information theory

“the fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point”

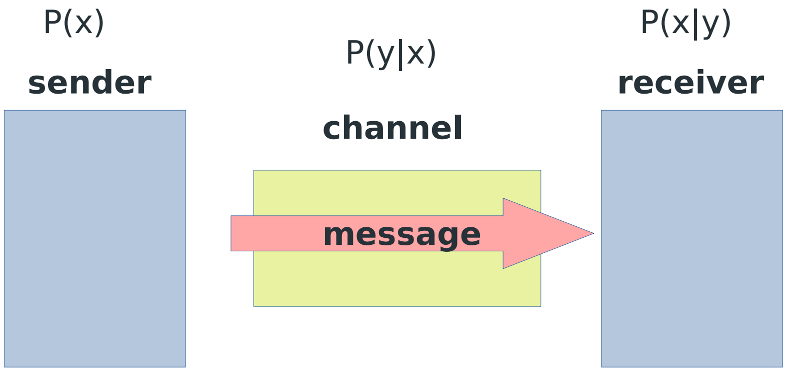
“the significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection”





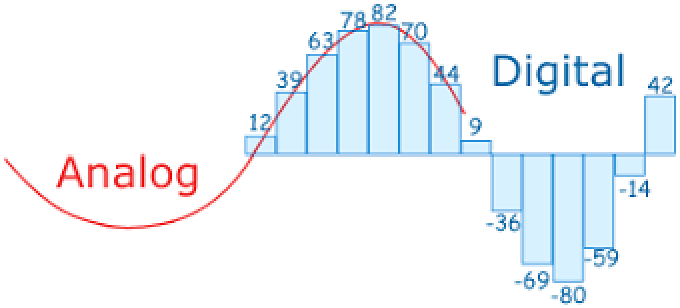
Information theory

- sender selects (input) message x in S with probability P(x)
- channel yields (output) y in R with probability P(y|x)
- Receiver reconstructs input message $\quad P(x|y)=P(y|x)P(x)/P(y)$
  (signal processing)

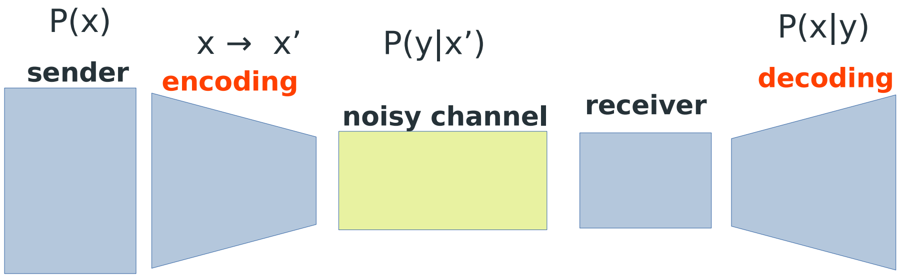


Information theory

encoding x → x'



Any type of information converted into digitized information (string of 0's and 1's)



Information theory

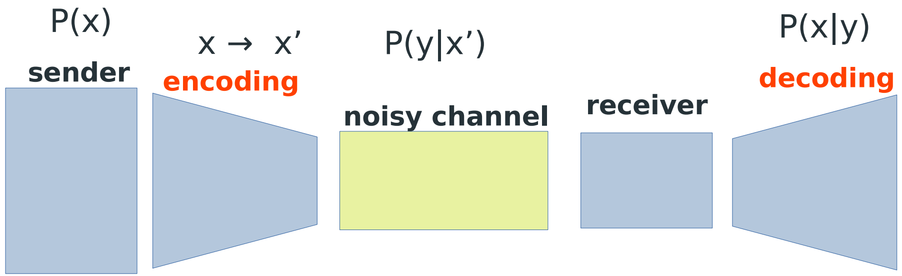
Reliable information transmission limited by mutual information: *I(X:Y)*

I(X:Y) = Mutual information = H(X)+H(Y)-H(X:Y)

H(X)= Shannon entropy =$-\sum p(x)\log[p(x)]$

H(X) information needed to specify *X*

I(X:Y) information Y provides about X



Information theory

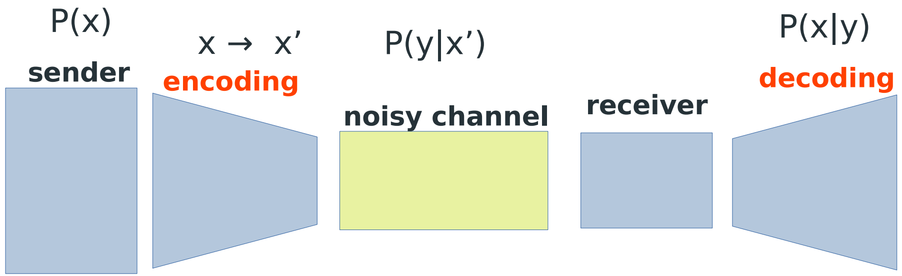
Reliable information transmission limited by mutual information: $I(X{:}Y)$

$I(X{:}Y) = H(X) - H(X|Y)$

$H(X)$ information needed to specify X

$H(X|Y)$ information needed to specify X is Y is known

$H(X) = I(X{:}Y) + H(X|Y)$



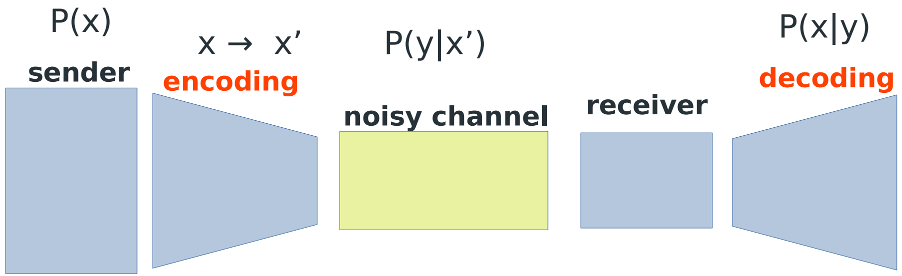
Information theory

Reliable information transmission limited by mutual information: $I(X{:}Y)$

$I(X{:}Y) = D(P(X,Y)||P(X)P(Y))$

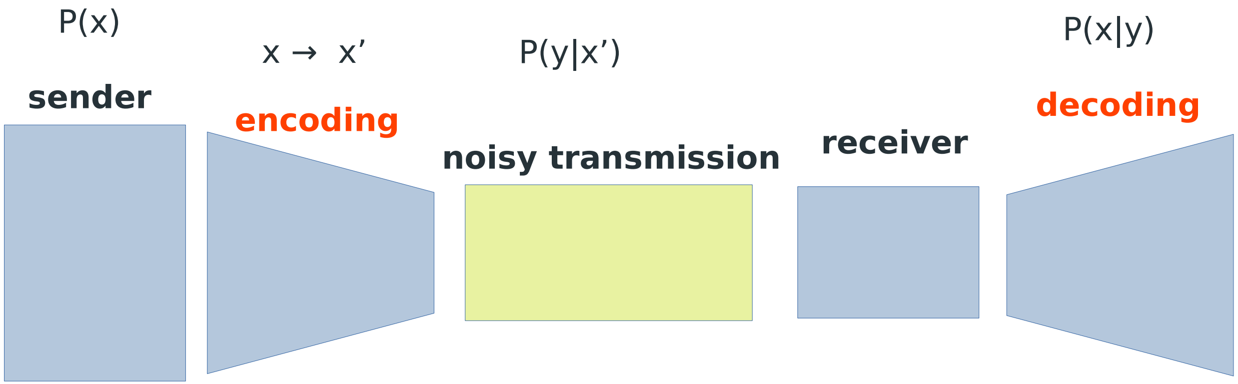
$D(P||Q)$ = *relative entropy/K-L divergence* = $-\sum p(x)\log[p(x)/q(x)]$

$D(P||Q)$ = *"distance between P and Q"*



Information theory

- ***Quantify*** the information that can be reliably transmitted
- ***Design*** channel to maximize information transmission rate



Information theory

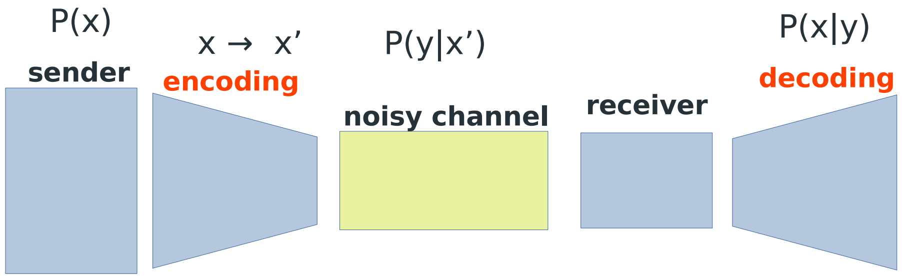
Achieving reliable information transmission with *minimal code length*

design smart encoding/decoding

e.g. send bit (0,1) though noisy channel with bit flip probability of ε

repetition code: $0 \rightarrow 0000$ $1 \rightarrow 1111$

error probability ε^k



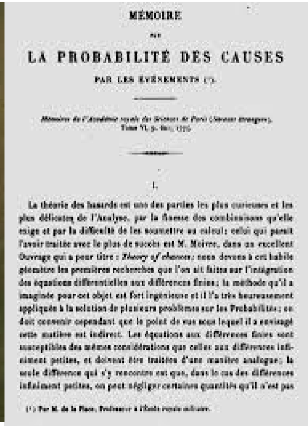
Inference

“the uncertainty of knowledge rests on events or their causes.”

“If we are sure an urn contains black and white papers in a given ratio, and we ask the chance that extracting random paper it will be white, the event is uncertain but the cause determining its probability (the white/black ratio), is known.”

“Hence the following problem: if an urn contains black and white papers in an unknown ratio, and we extract a white paper, determine the chance that the ratio is *p/q*”

“the probability of each cause is equal to the probability of the event given the cause, divided by the sum of all the probabilities of the event given any cause”



MÉMOIRE

SUR

LA PROBABILITÉ DES CAUSES

PAR LES ÉVÉNEMENTS (¹).

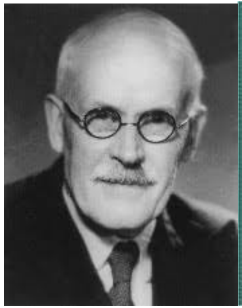
I.

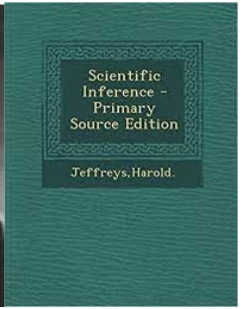
La théorie des hasards est une des parties les plus curieuses et les plus délicates de l'Analyse, par la finesse des combinaisons qu'elle exige et par la difficulté de les soumettre au calcul; celui qui paraît l'avoir traitée avec le plus de succès est M. Moivre, dans un excellent Ouvrage qui a pour titre : *Theory of chances*; nous devons à cet habile géomètre les premières recherches que l'on ait faites sur l'intégration des équations différentielles aux différences finies; la méthode qu'il a imaginée pour cet objet est fort ingénieuse et il l'a très heureusement appliquée à la solution de plusieurs problèmes sur les Probabilités; on doit convenir cependant que le point de vue sous lequel il a envisagé cette matière est indirect. Les équations aux différences finies sont susceptibles des mêmes considérations que celles aux différences infiniment petites, et doivent être traitées d'une manière analogue; la seule différence qui s'y rencontre est que, dans le cas des différences infiniment petites, on peut négliger certaines quantités qu'il n'est pas

(¹) Par M. de la Place, Professeur à l'École royale militaire.

Inference

Harold Jeffreys 1939





Again, suppose that $p_1, p_2, \dots p_n$ are a number of mutually exclusive hypotheses such that one of them must be true.

Therefore

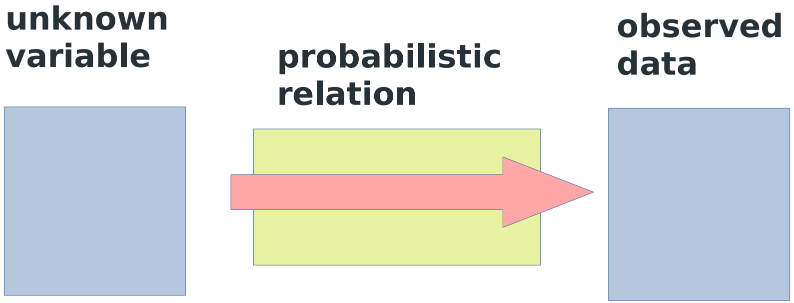
$$P(p_r \mid q \,.\, h) = \frac{P(q \mid p_r \,.\, h)\, P(p_r \mid h)}{\sum_{r=1}^{n} P(q \mid p_r \,.\, h)\, P(p_r \mid h)}. \qquad (4)$$

This theorem* is to the theory of probability what Pythagoras's theorem is to geometry.

Then our result is that the posterior probability of p is the prior probability of p divided by the prior probability of the consequence.

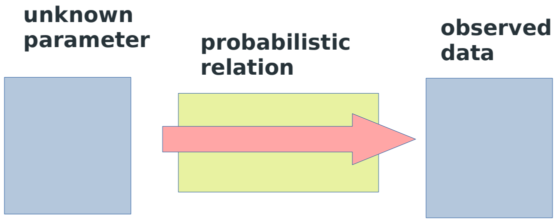
Inference

- unknown variable x in S with probability P(x)
- data y in R with probability P(y|x)
- reconstruct variable $P(x|y)=P(y|x)P(x)/P(y)$
  (data processing)

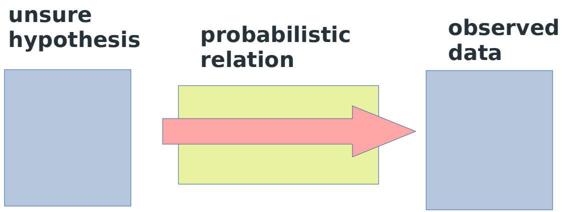


Inference

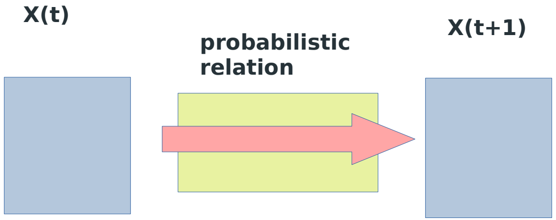
Parameter inference



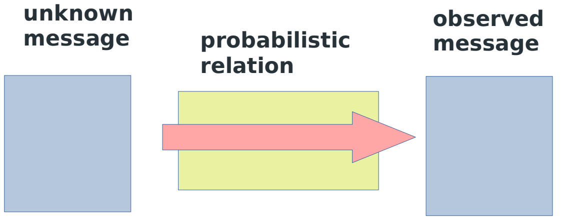
Hypothesis testing



Dynamical system analysis



Communication



Inference and sampling

- Inference requires to compute a posterior $P(x|y)=P(y|x)P(x)/P(y)$

- This allows to obtain *best estimates* and assess *uncertainty* about estimates

$$x^{est} = E[x|y] \qquad\qquad \delta x^{est} = E[(x-x^{est})^2|y]$$

- This involves *complicated high-dimensional sums/integrals*

- *e.g.* $E[x|y] = \int dx\, x\, P(y|x)P(x)/P(y) = \int dx\, x\, P(y|x)P(x)/\int dx\, P(y|x)P(x)$

sampling is needed!!

Inference and sampling

- Ising spins $\mathbf{s} = \{s_1, \ldots, s_n\} \in \{-1, 1\}^N$
- observe spins:

$$P(\mathbf{s}|\mathbf{h}, J) = \frac{e^{-\mathbf{h}\cdot\mathbf{s} - \mathbf{s}^T J \mathbf{s}}}{Z(\mathbf{h}, J)}$$

- reconstruct $\mathbf{h}, J$:

$$P(\mathbf{h}, J|\mathbf{s}^{obs}) = \frac{P(\mathbf{s}^{obs}|\mathbf{h}, J)}{P(\mathbf{s}^{obs})} = \frac{P(\mathbf{s}^{obs}|\mathbf{h}, J)}{\int d\mathbf{h} dJ \;\; P(\mathbf{s}^{obs}|\mathbf{h}, J)}$$

$$\langle \mathbf{h} \rangle = \frac{\int d\mathbf{h} dJ \;\; \mathbf{h} \;\; P(\mathbf{s}^{obs}|\mathbf{h}, J)}{\int d\mathbf{h} dJ \;\; P(\mathbf{s}^{obs}|\mathbf{h}, J)}, \qquad \langle J \rangle = \frac{\int d\mathbf{h} dJ \;\; J \;\; P(\mathbf{s}^{obs}|\mathbf{h}, J)}{\int d\mathbf{h} dJ \;\; P(\mathbf{s}^{obs}|\mathbf{h}, J)}$$

Inference and sampling

Computers were invented for
numerical integration!

Metropolis

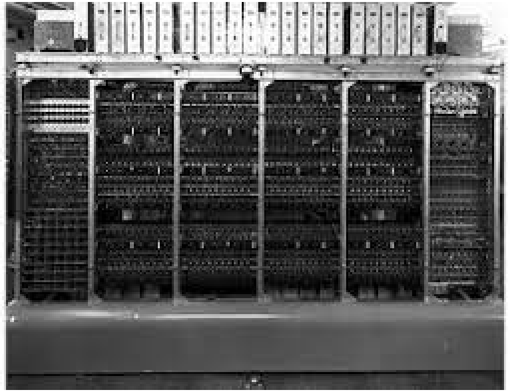
THE JOURNAL OF CHEMICAL PHYSICS VOLUME 21, NUMBER 6 JUNE, 1953

Equation of State Calculations by Fast Computing Machines

NICHOLAS METROPOLIS, ARIANNA W. ROSENBLUTH, MARSHALL N. ROSENBLUTH, AND AUGUSTA H. TELLER,
Los Alamos Scientific Laboratory, Los Alamos, New Mexico
AND
EDWARD TELLER, * *Department of Physics, University of Chicago, Chicago, Illinois*

(Received March 6, 1953)

A general method, suitable for fast computing machines, for investigating such properties as equations of state for substances consisting of interacting individual molecules is described. The method consists of a modified Monte Carlo integration over configuration space. Results for the two-dimensional rigid-sphere system have been obtained on the Los Alamos MANIAC and are presented here. These results are compared to the free volume equation of state and to a four-term virial coefficient expansion.



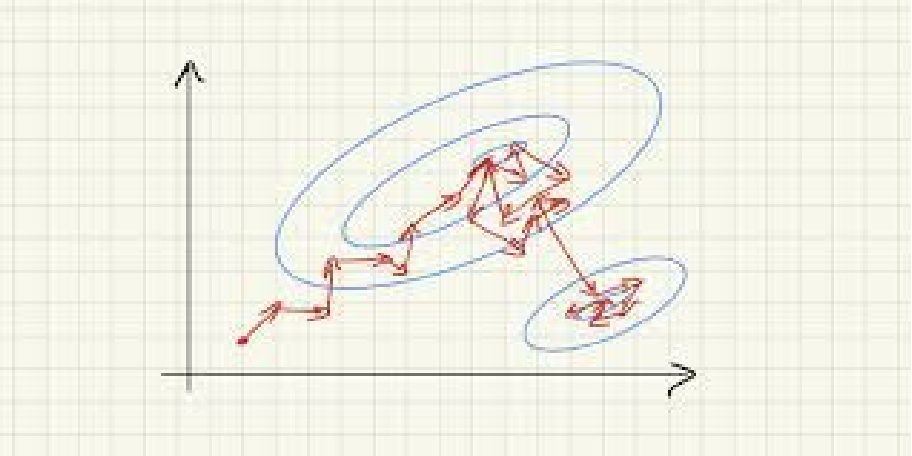
Maniac

Inference and sampling

Large computing power

Smart algorithms (Markov chain Monte Carlo)

Design stochastic dynamical system that automatically sample to required distribution



Information theory and inference

- a general inference process can be thought of as “noisy channel”
- Idea applicable to prediction, sensing, parameter estimation, hypothesis testing
- “message” to be reconstructed can be signal, parameter, hypothesis
- Information theory quantifies *how much information is contained in the data*
- Information theory establishes *fundamental limits* of inference
- Information theory suggests *recipes for data processing*

Example: Fano's inequality

- message $X \in \{1, \ldots, M\}$
- "noisy" output Y
- reconstructed message $\hat{X} = f(Y)$
- probability of error: $P_e = Prob(\hat{X} \neq X)$
- Fano's inequality:

$$P_e \geq 1 - \frac{I(X:Y)+1}{\log(M)}$$

- With n indipendent repetitions

$$P_e \geq 1 - \frac{nI(X:Y)+1}{\log(M)}$$

- to achieve $P_e \leq \delta$,

$$n \geq \frac{(1-\delta)\log(M)-1}{I(X:Y)}$$

Example: Fano's inequality

- **unknown variable** $X \in \{1, \ldots, M\}$
- **known data** Y
- **estimator** $\hat{X} = f(Y)$
- probability of **inference error**: $P_e = Prob(\hat{X} \neq X)$
- **minimum data size** to achieve $P_e \leq \delta$,

$$n \geq \frac{(1-\delta)\log(M) - 1}{I(X:Y)}$$