

# **Discrete prior probabilities: the entropy principle**

Jaynes, Chap. 11

Presented by Frederik

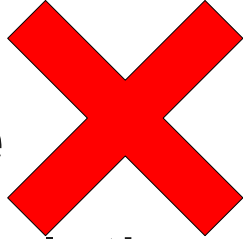
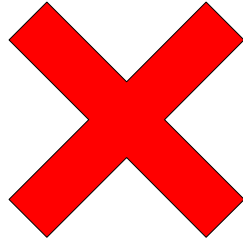
# Outline

- A new kind of prior
- Information entropy: a measure of amount of uncertainty
  - Shannon's derivation
  - Wallis derivation
- Maximum entropy distributions
- Objections against maximum entropy

# A new kind of prior

- **Ex.:** Translating the English “*in*” into French
  - $p(\textit{dans}) + p(\textit{en}) + p(\textit{\grave{a}}) + p(\textit{au cours de}) + p(\textit{pendant}) = 1$
  - From analyzing texts, we know
    - $p(\textit{dans}) + p(\textit{en}) = 3/10$
    - $p(\textit{dans}) + p(\textit{\grave{a}}) = 1/2$
  - Cannot use principle of indifference
- **Goal:** Assign a probability distribution as uniform as possible while agreeing with constraints.

# What *doesn't* work

- Maximizing **variance** 
  - Leads to unjustified solutions
- Minimizing **sum of squares** 
  - May end up with negative  $p_i$
  - “Fixing” them is not an option
    - Different principles of reasoning for different constraint values
    - Assigns zero probability to situations that are not ruled out by prior information

# A measure of uncertainty

Requirements for a measure of uncertainty of a probability distribution:

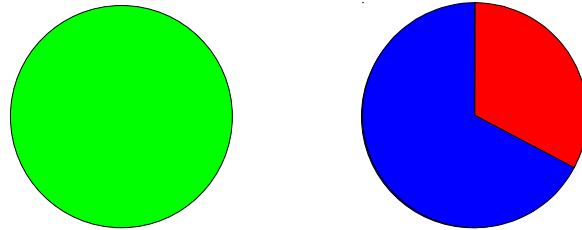
- (1) Measure is a *real-valued* function  $H(p_1, \dots, p_n)$
- (2) *Continuity*: A small change in  $p_i$  may cause only a small change in uncertainty
- (3) *Common sense*: More possibilities  $\rightarrow$  more uncertainty.

Formally:  $h(n) \leq h(n+1)$ , where  $h(n) = H\left(\underbrace{\frac{1}{n}, \dots, \frac{1}{n}}_{n \text{ times}}\right)$

- (4) *Consistency*: All ways of working out  $H$  need to yield the same value

# Functional equations for H

- Given two alternatives with probabilities  $p_1, q$ 
  - Uncertainty:  $H(p_1, q)$



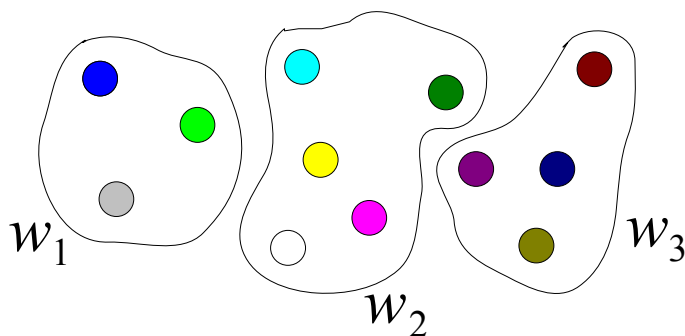
- Second alternative really consists of two different alternatives with probabilities  $p_2, p_3$
- What's  $H(p_1, p_2, p_3)$ ?

$$H(p_1, p_2, p_3) = H(p_1, q) + qH\left(\frac{p_2}{q}, \frac{p_3}{q}\right)$$

# Functional equations for H (cont.)

- Generalization

- n alternatives with probabilities  $p_i$



$$w_1 = p_1 + p_2 + p_3$$

$$w_2 = \dots$$

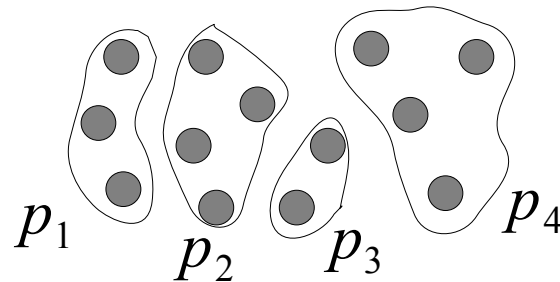
- Group them into composite propositions

$$H(p_1, \dots, p_n) =$$
$$H(w_1, \dots, w_r) + w_1 H\left(\frac{p_1}{w_1}, \dots, \frac{p_k}{w_1}\right) + w_2 H\left(\frac{p_{k+1}}{w_2}, \dots, \frac{p_{k+m}}{w_2}\right) + \dots$$

# Deriving h

- Consider rational  $p_i = \frac{n_i}{N}$ ,  $N = \sum n_j$ 
  - Imagine  $p_i$  stands for a composition of  $n_i$  propositions with equal probabilities.

$$p_1 = \frac{3}{13}, p_2 = \dots$$



- Then  $h(N) = h(\sum n_j) = H(p_1, \dots, p_n) + \sum p_i h(n_i)$
- If all  $n_i = m$ , we get  $h(mn) = h(m) + h(n)$
- This is solved by  $h(n) = K \log(n)$



# Finally, a measure of uncertainty

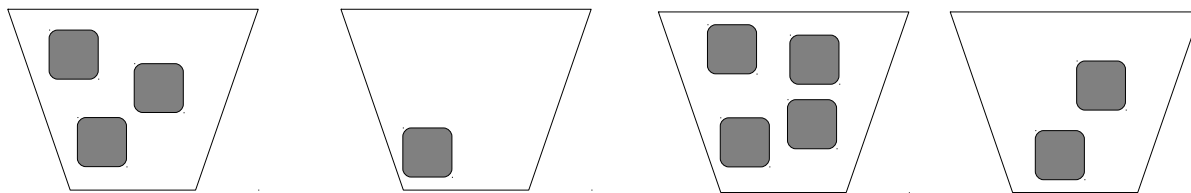
- Using the functional equations and  $h = \log(n)$

$$H(p_1, \dots, p_n) = -\sum p_i \log(p_i)$$

- H is called information entropy
  - Not to be confused with experimental entropy
  - Showed only necessity
  - Proof of uniqueness is in the book ;)

# Wallis derivation

- Goal: Assign probabilities  $p_i$  to  $m$  different propositions subject to constraints
- Game:
  - Distribute the  $n \gg m$  quanta of probability randomly among the  $m$  propositions



$$p_i = \frac{n_i}{n}$$

- Check if the resulting assignment satisfies the constraints
  - If yes: done, else: repeat game.

# Wallis derivation

- What's the probability of getting a specific assignment?
  - Multinomial distribution  $m^{-n} \cdot W$ ,  $W = \frac{n!}{n_1! \cdots n_m!}$
  - Larger  $W \Rightarrow$  more likely result
- As  $n \rightarrow \infty$ :  $\frac{1}{n} \log(W) \rightarrow H(p_1, \dots, p_m)$ 
  - Thus, the most likely assignment is the one that maximizes entropy

# Maximum entropy distributions

- Let's put the measure to work
  - Given propositions  $A_1, \dots, A_n$ , variable  $x$  can take corresponding values  $x_1, \dots, x_n$
  - Of course, we want  $\sum p_i = 1$
  - Our prior tells us that  $F_k = \langle f_k(x) \rangle = \sum p_i f_k(x_i)$ 
    - I.e., the expected values for the functions  $f_k$  are given

# Maximum entropy distributions

- Using the Lagrange method, it is shown that

$$p_i = \exp\left(-\lambda_0 - \sum_{j=1}^m \lambda_j f_j(x_i)\right)$$

- $\lambda_i$  are Lagrange multipliers
  - $\lambda_i$  are chosen so they satisfy the constraints
- Alternative derivation of  $p_i$  shows that it indeed maximizes  $H$ 
  - Necessary because Lagrange method doesn't work if maximum at a cusp

# Objections



# Round 1

- *“Maximum uncertainty’ is a negative thing which can’t possibly lead to any useful predictions.”*
  - This is a “play on words”
  - The principle doesn't create “new” uncertainty, it merely tries to avoid unwarranted assumptions

# Round 2

- *“Probabilities obtained by MAXENT are irrelevant to physical predictions because they have nothing to do with frequencies.”*
  - “The probability distribution which maximizes the entropy is numerically identical with the frequency distribution which can be realized in the greatest number of ways.”
  - “If the information incorporated into the maximum entropy analysis includes all the constraints actually operating in the random experiment, then the distribution predicted by maximum entropy is overwhelmingly the most likely to be observed experimentally.”



# Round 3

- *“The principle only works when the constraints are averages; in practice, they are real measurements, and not averages over anything.” [?]*
  - The principle also works for other constraints
  - If there are constraints on the width of the distribution, we can incorporate them

# Round 4

- *“Different people have different information, so the results are basically arbitrary.”*
- Consider Mr A and Mr B; Mr B has some additional information that Mr A hasn't
  - If Mr B's additional information is implied by Mr A's information, they will find at the same distribution
  - If Mr B's additional information is contradictory to his previous information, no distribution can be found
  - If Mr B's additional information was neither redundant or contradictory, his distribution will indeed have a lower entropy

“The principle of maximum entropy is not an oracle telling which predictions ***must*** be right; it is a rule for inductive reasoning that tells us which predictions ***are most strongly indicated by our present information.***”

The end.

