Supelec

Theoretical Foundations of Flexible Radios

February, 2008

Program

Course 1: **Overview and historical development.**
Course 2: **Plausible reasoning and quantitative rules.**
Course 3: **The entropy principle**
Course 4: **Ignorance priors and transformation groups**
Course 5: **Cognitive radios..results soon..**

Presentation

Overview and Historical development

Flexible Radio: Intelligence as a third resource

- The design of wireless networks requires an increase for new capacity and higher performance.
- The development of these capabilities is limited severely by the scarcity of two of the principal resources in wireless networks, namely
  - Energy
  - Bandwidth.
- Recently, the community has turned to a third principal resource,

  the deployment of intelligence at all layers of the network

  in order to exploit increases in processing power afforded by Moore's Law type improvements in microelectronics.

Flexible Radio: a Shannon historic perspective

THREE LANDMARK PAPERS FROM ALCATEL-LUCENT

1948 ”A Mathematical Theory of Communication”, C. Shannon, Bell System Technical Journal, Vol. 27 (July and October 1948), pp. 379-423 and 623-656.

1949 ”Communication Theory of Secrecy Systems”, C. Shannon, Bell System Technical Journal, Vol. 28 (1949), pp. 656-715.

1950 ”Programming a Computer for Playing Chess”, C. Shannon, Philosophical Magazine, Series 7, Vol. 41 (No. 314, March 1950), pp. 256-275

Flexible Radio: a Shannon historic perspective

Philosophical Magazine, Ser.7, Vol. 41, No. 314 - March 1950.

XXII. Programming a Computer for Playing Chess[1]

By CLAUDE E. SHANNON

Bell Telephone Laboratories, Inc., Murray Hill, N.J.[2]

[Received November 8, 1949]

1. INTRODUCTION

This paper is concerned with the problem of constructing a computing routine or "program" for a modern general purpose computer which will enable it to play chess. Although perhaps of no practical importance, the question is of theoretical interest, and it is hoped that a satisfactory solution of this problem will act as a wedge in attacking other problems of a similar nature and of greater significance. Some possibilities in this direction are: -

Flexible Radio: a Shannon historic perspective

(1)Machines for designing filters, equalizers, etc.
(2)Machines for designing relay and switching circuits.
(3)Machines which will handle routing of telephone calls based on the individual circumstances rather than by fixed patterns.
(4)Machines for performing symbolic (non-numerical) mathematical operations.
(5)Machines capable of translating from one language to another.
(6)Machines for making strategic decisions in simplified military operations.
(7)Machines capable of orchestrating a melody.
(8)Machines capable of logical deduction.

Flexible Radio: a Shannon historic perspective

Computers and Automata*

CLAUDE E. SHANNON†, FELLOW, IRE

C. E. Shannon first became known for a paper in which he applied Boolean Algebra to relay switching circuits; this laid the foundation for the present extensive application of Boolean Algebra to computer design. Dr. Shannon, who is engaged in mathematical research at Bell Telephone Laboratories, is an authority on information theory. More recently he received wide notice for his ingenious maze-solving mechanical mouse, and he is well-known as one of the leading explorers into the exciting, but uncharted world of new ideas in the computer field.

The Editors asked Dr. Shannon to write a paper describing current experiments, and speculations concerning future developments in computer logic. Here is a real challenge for those in search of a field where creative ability, imagination, and curiosity will undoubtedly lead to major advances in human knowledge.—*The Editor*

***Summary*—This paper reviews briefly some of the recent developments in the field of automata and nonnumerical computation. A number of typical machines are described, including logic machines, game-playing machines and learning machines. Some theoretical questions and developments are discussed, such as a comparison of computers and the brain, Turing's formulation of computing machines and von Neumann's models of self-reproducing machines.**

* Decimal classification: 621.385.2. Original manuscript received by the Institute, July 17, 1953.

† Bell Telephone Laboratories, Murray Hill, N. J.

Introduction

SAMUEL BUTLER, in 1871, completed the manuscript of a most engaging social satire, *Erewhon*. Three chapters of *Erewhon*, originally appearing under the title "Darwin Among the Machines," are a witty parody of *The Origin of Species*. In the topsy-turvy logic of satirical writing, Butler sees machines as gradually evolving into higher forms. He considers the classification of machines into genera, species and vari-

What is a Flexible Radio 50 years later?

Joseph Mitola III, ”Cognitive Radio: An Integrated Agent Architecture for Software Defined Radio”, Royal Institute of Technology (KTH) Stockholm, Sweden, 8 May, 2000.

It is the ultimate point where devices are sufficiently computationally intelligent about radio resources to detect user communication needs as a function of use context, and to provide (without human intervention) radio resources and wireless services most appropriate to those needs.

Definition of Mitola: A radio that employs model based reasoning to achieve a specified level of competence in radio-related domains.

Flexible Radio: the road ahead

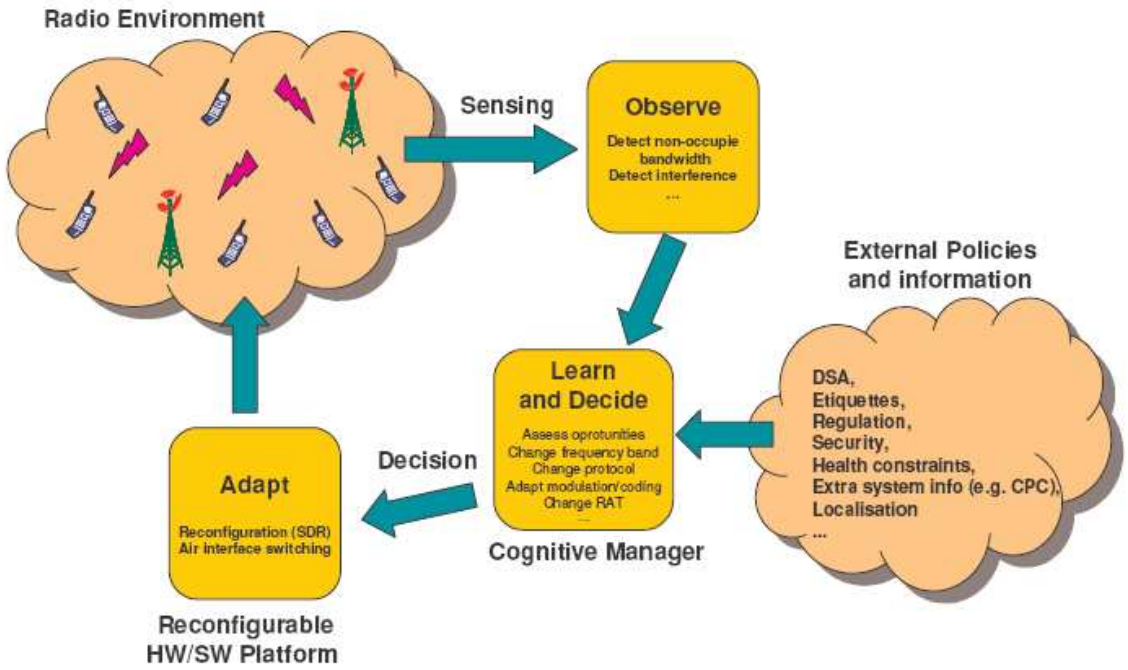


Fig. 1 : *Cognition Cycle*

It is also called a Cognitive radio (”Cogito, ergo sum” [I think, therefore I am]- René Descartes).

Flexible terminals or flexible networks?

Usual flexible radio research problems are user oriented and not infrastructure oriented.

However, bringing the intelligence at the network is even more important.

Example of infrastructure based problems:

- Collaborative multi-cell MIMO (also known as Network MIMO)
- Feedback and protocol overhead optimization.
- Cross-system and cross-layer resource allocation
- Network self-management
- Software reconfiguration of base stations.
- Spectrum management

The big dilemma

Three important aspects of the problem:

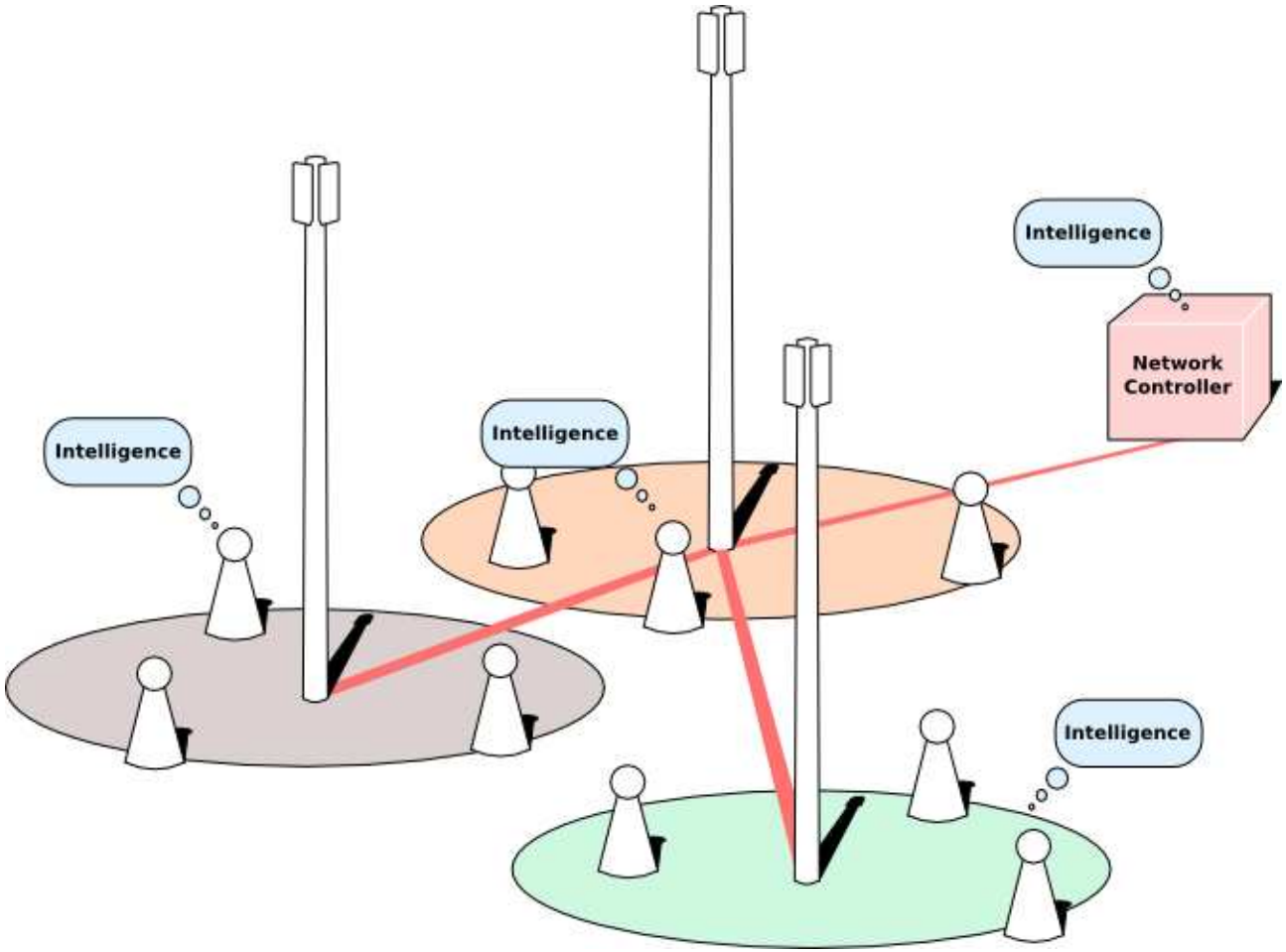
- Heterogeneity: systems are heterogeneous in transmit power, frequencies, range, QoS requirements, spectral efficiency and systems.
- Limited information: there may be limited or no communication between different systems and decisions have to be made based on such distributed information.
- Temporal requirements: systems change rapidly and the flexible radio needs to adapt fast.

One of the most challenging problems in the development of this theory is to manage complexity. The key is to develop:

- The right abstractions to reason about the spatial and temporal dynamics of complex systems
- Understand how information can be processed, stored, and transferred in the system with bounded delay.

We refer to this as the ”theoretical foundations of flexible radio”.

Example of theoretical problems we are concerned with



Interdisciplinary Research

The research is highly interdisciplinary and is a blend of

- Statistical inference methods to build radios which would carry plausible reasoning (Maximum Entropy methods,..).
- Game theoretic techniques (based on rational players) to promote decentralized/adaptive resource allocation schemes.
- Control theory with the use of feedback mechanisms in the realm of cybernetics.
- Random matrix theory and free probability theory to reduce the dimensionality of the problem i.e find the parameters of interest in a network rather than optimizing through simulations with 1 billion parameters.
- Physics to study how information can be processed, stored, and transferred in the network.
- Non-equilibrium Information theory to understand the fundamental limits achievable with intelligent radio.

References

E. T. Jaynes, ”Probability Theory: The Logic of Science”, Cambridge, 2003.

J. von Neumann and O. Morgenstern, ”Theory of Games and Economic Behavior”, Princeton, NJ: Princeton University Press, 1944.

N. Wiener, ”Cybernetics”, John Wiley, 1948.

V.A. Marchenko and L.A. Pastur, ”Distribution of Eigenvalues for Some Sets of Random Matrices”, Math USSR Sb, 1967, vol.1, pp. 457–483.

D.V. Voiculescu and K.J. Dykema and A. Nica, ”Free Random Variables”, American Mathematical Society, CRM Monograph Series, Volume 1, Providence, Rhode Island, USA, 1992.

C. H. Bennett and R. Landauer, ”Fundamental Physical Limits of Computation”, Scientific American 253:1 48-56, July 1985.

C. Shannon, ”A Mathematical Theory of Communication”, Bell System Technical Journal, Vol. 27 (July and October 1948), pp. 379-423 and 623-656.

A tribute to Jaynes

*********MORE TO COME**********

Jaynes unfinished masterpiece

Scientific inference, extended logic and probability theory

- The task is difficult and requires to essentially capture a mathematical model of human common sense.
- Our goal is to build an intelligent radio which would carry out useful plausible reasoning, following clearly defined principles expressing an idealized common sense.
- We have to design a radio which reasons according to certain rules and these rules will be deduced from simple desiderata.

Interpreting information

J. Bernoulli, ”The Art of Conjecture”, 1713

Jakob Bernoulli, 1654-1705

Bernoulli was among the first to realize the difference between deductive logic and inductive logic.

”The chance depends mainly upon our knowledge”

He was however unable to use the outcomes to makes inferences about the way in which an observed situation could have occurred.

Interpreting information

Reverend Thomas Bayes, ”An Essay towards solving a Problem in the Doctrine of Chances”, 1763

Reverend Thomas Bayes, 1701-1761

Bayes turned the situation and provided a formula to make inferences about the causes using the outcomes.

The Birth of Plausible reasoning

Théorie Analytique des Probabilités, Tomes VII, P.S. Laplace, Edition Jacques Gabay, Third edition, 1820, volumes I and II, 1847, Reprint 1995,ISBN 2-87647-161-2

Pierre Simon Laplace, 1749-1827

Laplace set out a mathematical system of inductive reasoning based on probability, which we would today recognize as plausible reasoning and rediscovered Bayes theorem.

Probability theory is nothing but common sense reduced to calculation

Plausible reasonning

R. T. Cox, ”Probability, Frequency, and Reasonable Expectation,” Am. Jour. Phys., 14, 1-13, (1946).

R.T Cox in 1946 derived three requirements known as Cox's Theorem:

- Divisibility and comparability: the plausibility of of a statement is a real number between 0 (for false) and 1 (for true) and is dependent on the information we have related to the statement.
- Common sense: Plausibilities should vary with the assessment of plausibilities in the model.
- Consistency: If the plausibility of a statement can be derived in two ways, the two results should be equal.

which leads to Bayes rule and Bayesian probability Theory.

Deduction versus plausible reasoning

Deduction from fact: $A \rightarrow B$

- A is true, therefore B is true.
- B is false, therefore A is false

Plausible reasoning from fact: $A \rightarrow B$

- B is true. What does it say about A?
- A is false. What does it say about B?

Priors

Prior distributions

- If correct, increase the accuracy of the conclusions.
- If wrong, do the opposite.

One of the big debate has been the construction of priors (in other words how to translate information into ”plausibilities”!)

Indeed, probability theory here is regarded as the calculus of inductive reasonning.

The principle of indifference

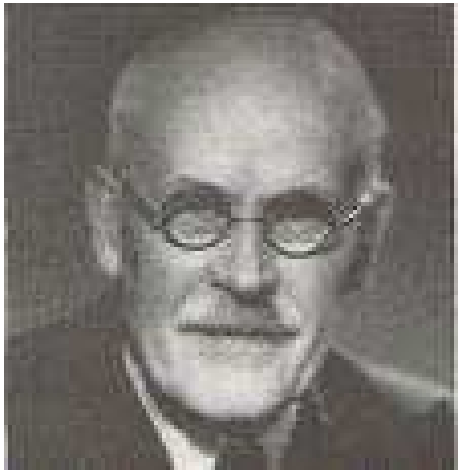
”A Treatise on Probability”, J.M Keynes, MacMillan and Co, London, 1921

J. M. Keynes, 1883-1946

Although he revolutionized economics with his classic book, The General Theory of Employment, Interest and Money (published in 1936), Keynes is mostly known in mathematics for introducing already in 1921 the principle of indifference (after formalized as the principle of maximum entropy) to express our indifference in attributing prior values when no information is available.

Uninformative priors

H. Jeffreys, ”Theory of Probability”, Oxford University Press, London, 1939



Sir Harold Jeffreys, 1891-1989

Jeffrey shows that the theory of probability is nothing else than a theory of inductive inference founded on the principle of inverse probability. His whole objective was to use probability to represent human information (extending the work of Laplace where prior probabilities represent a state of knowledge).

The controversy with Fisher

R. A. Fisher, ”Statistical Methods for Research Workers ”, London, 1925

Ronald Fisher, 1890-1962

A probability statement can only be made about random variables and not about unknown fixed parameters. Probability is defined as a limiting frequency of an experiment.

Note that reactions against Laplace began already in the mid-19th century (Cournot, Boole and Venn) and were continued by Von. Mises, Neyman, Feller and Savage.

The orthodox statistics school

- For a century, probability was identified as a limiting frequency of events in a series of identically repeated and independent experiences.
- This point of view is known as the classical, frequentist or orthodox school.
- In the bayesian framework, a probability ia a measure of the plausibility of a proposition with incomplete information. The measure is a real number between 0 and 1 and is somehow a numerical coding of information.
- The frequentist approach has to infer on hypothetical data as the data (and not the hypothesis) are considered as random variables.

Frequentist versus bayesian approach

For some reason, data in the frequentist school are always considered random, almost everything else is nonrandom.

However, in practice, the data (information) are usually the only things that are definite and known and almost everything else in the problem is unknown and conjectured.

For our cognitive problems, the opposite choice seems far more natural.

Frequentist versus bayesian approach: an example

To specify the geographical layout of a country, there are two possible methods:

- Specify the points that are in it
- Specify its boundary

The first method is the Venn-Kolmogorov viewpoint while the second is more related to the information (scientific inference) we have in real problems.

Frequentist versus bayesian approach

The bayesian question: What is the probability conditional on the data that the hypothesis is true?

The frequentist question: If the hypothesis being tested is in fact true, what is the probability that we shall get data indicating that it is true?

The Bayesian revival: Jaynes approach and the maximum entropy principle

Information Theory and Statistical Mechanics, Part 1&2, E. T. Jaynes, Phys. Rev, 1957, vol. 106 and vol. 108.

E. T. Jaynes, 1922-1998

The principle of Maximum entropy:
The principle states that, of the distribution P that satisfy the constraints, one should choose the one which maximizes: $-\int P(x)\log(P(x))dx$
Why?
We need a measure of uncertainty which expresses the constraints of our knowledge and the desire to leave the unknown parameters to lie in an unconstrained space.

The principle of maximum entropy

J. Shore and R. Johnson, ”Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy”, Information Theory, IEEE Transactions on Volume 26, Issue 1, Jan 1980 Page(s):26 - 37

Shore proved that the maximum entropy principle is the correct method of inference when given new information in terms of expected values using four consistency axioms:

- Uniqueness: If one solves the same problem twice the same way then the same answer should result both times.
- Invariance: If one solves the same problem in two different coordinate systems then the same answer should result both times.
- System independence: It should not matter whether one accounts for independent information about independent systems separately in terms of different densities or together in terms of joint density.
- subset independence: It should not matter whether one treats an independent subset of system states in terms of a separate conditional density or in terms of the full system density.

An example in Spectrum estimation

J. P. Burg, "Maximum Entropy Spectral Analysis", PhD dissertation, Stanford University, 1975

Back to 1975: what is the consistent estimator subject the following autocorrelation constraints $\mathbb{E}(x_i x_{i+k}) = \tau_k, k = 0, \ldots, p$ for all i?

Answer (Burg's Theorem): Use the principle of Maximum entropy! The process is shown to be a p^{th} AR order model of the form:

$$x_i = -\sum_{k=1}^{p} a_k x_{i-k} + b_i$$

for which the spectrum is given by:

$$P(f) = \frac{\sigma^2}{\mid 1 + \sum_{k=1}^{p} a_k e^{-j2\pi kf} \mid^2}$$

Where do we stand now: information theory versus science

L. Brillouin, ”Science and Information Theory”, Academic Press, New York , 1956..I strongly recommend...

H. P. Yockey, ”Information theory from molecular biology”, Cambridge University Press, 1992

F. B. Roy, ”Science from Fisher Information: A unification”, Cambridge University Press, 2004

- Information theory is a principle used to derive most laws of physics, biology, chemistry and economics.

Presentation

Plausible reasoning and quantitative rules

Some notions on Boolean Algebra

G. Boole, ”An investigation of the laws of thought”, Macmillan, London, 1854

- The symbol AB called the logical product or the conjunction, denotes the proposition ”both A and B” are true.
- The expression A+B called the logical sum or disjunction, stands for ”at least one of the propositions A,B is true”.
- Of course, in each case, the order in which things are stated does not matter.

Some notions on Boolean Algebra

- In Boolean algebra, when ”A=B”, the equal sign is used to denote not equal numerical value but equal truth values.
- The denial of a proposition is indicated by a bar

$$A = \bar{A}$$

  is false.
- Moreover,
  - $$A\bar{B} = AB$$

    is false.
  - $$\bar{A}\bar{B} = \text{both} A \text{and} B$$

    are false.

Basic identities

- Idempotence:

$$AA = A$$
$$A + A = A$$

- Commutativity

$$AB = BA$$
$$A + B = B + A$$

- Associativity

$$A(BC) = (AB)C = ABC$$
$$A + (B + C) = (A + B) + C = A + B + C$$

Basic identities

- Distributivity:

$$A(B+C) = AB + AC$$
$$A + (BC) = (A+B)(A+C)$$

- Duality

$$\text{If} C = AB, \text{then} \bar{C} = \bar{A} + \bar{B}$$
$$\text{If} D = A + B, \text{then} \bar{D} = \bar{A}\bar{B}$$

Implication

The proposition

$$A \Rightarrow B$$

does not assert that either A or B is true.

It means $A\bar{B}$ is false or $(\bar{A} + B)$ is true.

Extended logic, axioms and desiderata

- The flexible radio is designed by us, so it reasons according to certain define rules.
- We are free to adopt any rules we please.
- We call these rules ”desiderata” and not axioms because they do not assert that anything is ”true” but only state what appear to be desirable goals.

Desideratum (I)

- To each proposition about which it reasons, the rradio must assign some degree of plausibility, based on the information given.
- Whenever it receives new evidence, it must revise these assignments to take that new evidence into account.

(I) Degrees of plausibility are represented by real numbers

We adopt a natural but essential convention: that a greater plausibility shall correspond to a greater number and that an infinitesimally greater plausibility ought to correspond only to an infinitesimally greater number.

Remarks

- The plausibility of a proposition A will depend in general whether we are told that some other proposition is true (the background information) which is indicated by

$$A \mid B$$

  i.e ”The conditional plausibility that A is true given that B is true” or just ”A given B”.
- $A \mid BC$ represents the plausibility that A is true, given that both B and C are true.
- $A + B \mid CD$ represents the plausibility that at least one of the propositions A and B is true, given that C and D are true.

Remarks

- The plausibility of a proposition A will depend in general whether we are told that some other proposition is true (the background information) which is indicated by

$$A \mid B$$

  i.e ”The conditional plausibility that A is true given that B is true” or just ”A given B”.
- $A \mid BC$ represents the plausibility that A is true, given that both B and C are true.
- $A + B \mid CD$ represents the plausibility that at least one of the propositions A and B is true, given that C and D are true.
- $A \mid B > C \mid B$ means that given B, A is more plausible than C.

Desideratum (II)

- We want the radio to reason in a way that is at least qualitatively the humans try to reason.
- If it has old information C which gets updated to C' in such a way that the plausibility of A is increased but the plausibility of B given A is not changed then :
  - this can produce only an increase, never a decrease, in the plausibility that A and B are true.
  - It must produce a decrease in the plausibility that A is false.

(II) Qualitative correspondence with common sense

Desideratum (III)

- We want the radio to reason in a consistent way

(III) If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same results.

interface conditions

- (I) The radio always takes into account all of the evidence it has relevant to a question. It does not arbitrarily ignore some of the information, basing its conclusions only on what remains. In other words, the radio is completely nonideological
- (II) The radio always represents equivalent states of knowledge by equivalent plausibility assignments. That is, if in two problems the radio's state of knowledge is the same (except perhaps for the labeling of the propositions), then it must assign the same plausibilities.

Comments on the human brain

- In general, we form a judgement about a proposition not only as to whether is plausible but also whether it is desirable, important, interesting, moral..
- Hence, a fully adequate description of the human brain should be represented by a vector in a space of a rather large number of dimensions.
- Actually, such a theory is still under development but we will ignore the other coordinates for the moment.

The product rule

- We seek to relate the plausibility of AB to the plausibility of A and B separately. In particular, we would like to derive $AB \mid C$.
- The process of deciding that AB is true can be broken down to elementary pieces:
  - Decide that B is true and having accepted that B is true, decide that A is true.
  - Decide that A is true and having accepted that A is true, decide that B is true.

The product rule

- In order for AB to be a true proposition, it is necessary that B is true. Thus the plausibility of $B \mid C$ should be involved. Moreover, if B is true, it is further necessary that A should be true, so the plausibility of $A \mid BC$ is needed.
- But if B is false, then of course AB is false independently of whatever one knows about A. If the radio reasons fist about B, then the plausibility of A will be relevant only if B is true.
- Therefore, if the radio has $B \mid C$ and $A \mid BC$, it will not need $A \mid C$.
- Similarly, $A \mid B$ and $B \mid A$ are not needed. Whatever plausibility A and B might have in the absence on information C could not be relevant to judgements of a case in which the radio knows that C is true.

The product rule

- Since the product is commutative, we can interchange A and B in the above statements. Knowledge of $A \mid C$ and $B \mid AC$ would serve equally well to determine $AB \mid C$.
- Hence, in more definite form, we can write:

$$(AB \mid C) = F[(B \mid C), (A \mid BC)]$$

The product rule

Let us denote $u = (AB \mid C)$, $v = (A \mid C)$, $x = (B \mid C)$ and $y = (A \mid BC)$.

Desideratum (II) imposes that $F(x, y)$ must be a continous monotonic increasing function of both x and y. For simplicity reasons, we will assume it differentiable i.e

$$F_1(x, y) = \frac{\delta F}{\delta x} \geq 0$$

$$F_2(x, y) = \frac{\delta F}{\delta y} \geq 0$$

The product rule

Let us know make use of the consistency axiom.

$$(ABC \mid D) = F[(BC \mid D), (A \mid BCD)]$$

$$(ABC \mid D) = F\left[F[(C \mid D), (B \mid CD)], (A \mid BCD)\right]$$

But we have also:

$$(ABC \mid D) = F\left[(C \mid D), F[(B \mid CD), (A \mid BCD)]\right]$$

The product rule

A necessary and sufficient condition that the radio reasons in a consistent say is that:

$$F\left[F(x,y),z\right] = F\left[x,F(y,z)\right]$$

The function that satisfies this relation has been solved in:

R. T Cox, ”The algebra of probable inference”, John Hopkins University Press, Baltimore, 1961

The product rule

Let us denote $u = F(x, y)$ and $v = F(y, z)$.

We need to solve: $F(x, v) = F(u, z)$

Differentiating with respect to x and y, we obtain:

$$\begin{aligned} F_1(x, v) &= F_1(u, z)F_1(x, y) \\ F_2(x, v)F_1(y, z) &= F_1(u, z)F_2(x, y) \end{aligned}$$

Elimination of $F_1(u, z)$ provides $G(x, v)F_1(y, z) = G(x, y)$ (the left hand-side is independent of z) with $G(x, y) = \frac{F_2(x,y)}{F_1(x,y)}$. This can also be rewritten as:

$$G(x, v)F_2(y, z) = G(x, y)G(y, z)$$

The product rule

Prove that $G(x, y)G(y, z)$ is independent of y.

The most general function $G(x, y)$ with this property is

$$G(x, y) = r\frac{H(x)}{H(y)}$$

(r is a constant and the function H is arbitrary.

One can show that:

$$F_1(y, z) = \frac{H(v)}{H(y)}$$
$$F_2(y, z) = r\frac{H(v)}{H(z)}$$

The product rule

The relation $dv = F_1 dy + F_2 dz$ takes the form

$$\frac{dv}{H(v)} = \frac{dy}{H(y)} + r\frac{dz}{H(z)}$$

which gives:

$$Q(v) = Q(y)Q^r(z)$$

where $Q(x) = e^{\int^x \frac{du}{H(u)}}$.

By taking $Q(.)$ of $F(x, v) = F(u, z)$ and applying the previous formula, we obtain:

$$Q(x)Q^r(v) = Q(u)Q^r(z)$$

which again by applying the same formula yields:

$$Q(x)Q^r(y)[Q(z)]^{r^2} = Q(x)Q^r(y)Q^r(z)$$

The product rule

$$Q(x)Q^r(y)[Q(z)]^{r^2} = Q(x)Q^r(y)Q^r(z)$$

We obtain a non-trivial solution only if $r = 1$ which yields:

$$Q(u) = Q(x)Q(y)$$

In other words:

$$Q(AB \mid C) = Q(A \mid BC)Q(B \mid C) = Q(B \mid AC)Q(A \mid C)$$

Qualitative correspondence with common sense requires that $Q(x)$ be a positive continous monotonic function.

We will without loss of generality adopt the choice

$$0 \leq Q(x) \leq 1$$

as a convention.

The sum rule

- Note that:
  - The logical product $A\bar{A}$ is always false.
  - The logical sum $A + \bar{A}$ is always true.
- The plausibility that A is false must depend in some way on the plausibility that it is true.
- If we define $u = Q(A \mid B)$ and $v = Q(\bar{A} \mid B)$, there must exist some functional relation $v = S(u)$.

The sum rule

The function S must be:

- a continuous monotonic function ("qualitative correspondence with common sense") with extreme values $S(0) = 1$ and $S(1) = 0$.
- By consistent with the product rule
  - $Q(AB \mid C) = Q(A \mid C)Q(B \mid AC)$.
  - $Q(A\bar{B} \mid C) = Q(A \mid C)Q(\bar{B} \mid AC)$.
- which yields

$$Q(AB \mid C) = Q(A \mid C)S[Q(\bar{B} \mid AC)] = Q(A \mid C)S\left[\frac{Q(A\bar{B} \mid C)}{Q(A \mid C)}\right]$$

- The symmetry argument yields:

$$Q(A \mid C)S\left[\frac{Q(A\bar{B} \mid C)}{Q(A \mid C)}\right] = Q(B \mid C)S\left[\frac{Q(B\bar{A} \mid C)}{Q(B \mid C)}\right]$$

The sum rule

The previous relation must hold for all propositions A, B, C and in particular for $\bar{B} = AD$

In this case, we have:

$$A\bar{B} = \bar{B}$$
$$B\bar{A} = \bar{A}$$

and

$$Q(A\bar{B} \mid C) = Q(\bar{B} \mid C) = S\left[Q(B \mid C)\right]$$
$$Q(B\bar{A} \mid C) = Q(\bar{A} \mid C) = S\left[Q(A \mid C)\right]$$

By noting $x = Q(A \mid C)$ and $y = Q(B \mid C)$, S has to solution of the following functional equation:

$$xS\left[\frac{S(y)}{x}\right] = yS\left[\frac{S(x)}{y}\right]$$

The sum rule

J. Aczél, ”Lectures on Functional Equations and their applications”, Academic Press, New York 1966

$$xS\left[\frac{S(y)}{x}\right] = yS\left[\frac{S(x)}{y}\right]$$

The only solution of the equation satisfying $S(0) = 1$ is

$$S(x) = (1 - x^m)^{\frac{1}{m}}$$

Which shows that (m is a positive integer):

$$Q^m(A \mid B) + Q^m(\bar{A} \mid B) = 1$$

The sum rule

$$Q^m(A \mid B) + Q^m(\bar{A} \mid B) = 1$$

We also have:

$$Q^m(AB \mid C) = Q^m(B \mid AC)Q^m(A \mid C) = Q^m(A \mid BC)Q^m(B \mid C)$$

If we define $P(x) = Q^m(x)$, we have:

$$P(A \mid B) + P(\bar{A} \mid B) = 1$$

$$P(AB \mid C) = P(B \mid AC)P(A \mid C) = P(A \mid BC)P(B \mid C)$$

We got exactly the basic rules of probability theory.

Probability theory is nothing but common sense reduced to calculations!

Q or P?

From now on, P will be called a probability.

The measure we have found can be changed and it we could as well use Q (which will cause only to rescale our measure).

The situation is analogous to thermodynamics, where out of all empirical temperature scales, which are monotonic functions of each other, we finally decide to use the Kelvin state, NOT BECAUSE IT IS MORE CORRECT THAT OTHERS BUT BECAUSE IT IS MORE CONVENIENT.

The logical sum $A + B$

$$
\begin{aligned}
P(A + B \mid C) &= 1 - P(\bar{A}\bar{B} \mid C) \\
&= 1 - P(\bar{B} \mid \bar{A}C)P(\bar{A} \mid C) \\
&= 1 - P(\bar{A} \mid C)\left[1 - P(B \mid \bar{A}C)\right] \\
&= P(A \mid C) + P(\bar{A} \mid C) - P(\bar{A} \mid C)\left[1 - P(B \mid \bar{A}C)\right] \\
&= P(A \mid C) + P(\bar{A} \mid C)P(B \mid \bar{A}C) \\
&= P(A \mid C) + P(\bar{A}B \mid C) \\
&= P(A \mid C) + P(B \mid C)P(\bar{A} \mid BC) \\
&= P(A \mid C) + P(B \mid C)\left[1 - P(A \mid BC)\right] \\
&= P(A \mid C) + P(B \mid C) - P(AB \mid C)
\end{aligned}
$$

Qualitative properties

- The statement "$A \rightarrow B$, B is true, therefore A becomes more plausible" corresponds to the product rule

$$P(A \mid BC) = P(A \mid C)\frac{P(B \mid AC)}{P(B \mid C)}$$

- The statement "$A \rightarrow B$, A is false, therefore B becomes less plausible" corresponds to the product rule

$$P(B \mid \bar{A}C) = P(B \mid C)\frac{P(\bar{A} \mid BC)}{P(\bar{A} \mid C)}$$

Mutually exclusive propositions

One can show that:

$$\begin{aligned} P(A_1 + A_2 + A_3 \mid B) &= P(A_1 \mid B) + P(A_2 \mid B) + P(A_3 \mid B) - P(A_1 A_2 \mid B) \\ &- P(A_2 A_3 \mid B) - P(A_3 A_1 \mid B) + P(A_1 A_2 A_3 \mid B) \end{aligned}$$

If we suppose the propositions exclusive and exhaustive (the background information B supposes stipulated that one and only one of them must be true), then:

$$P(A_1 + A_2 + A_3 \mid B) = P(A_1 \mid B) + P(A_2 \mid B) + P(A_3 \mid B) = 1$$

which generalizes to:

$$P(A_1 + A_2 + \ldots + A_m \mid B) = \sum_{i=1}^{m} P(A_i \mid B) = 1$$

Marginalization procedure

Note that we have:

$$A = A.(B + \bar{B}) = AB + A\bar{B}$$

Therefore,

$$P(A \mid I) = P(A.(B + \bar{B}) \mid I) = P(AB + A\bar{B} \mid I)$$

Since AB and $A\bar{B}$ are mutually exclusive propositions, we have:

$$P(A \mid I) = P(AB + A\bar{B} \mid I) = P(AB \mid I) + P(A\bar{B} \mid I)$$

The results can be generalized to any $B_1, B_2, \ldots, B_m$ which are exclusive and exhaustive:

$$P(A \mid I) = \sum_{i=1}^{m} P(AB_i \mid I)$$

The entropy principle

The principle of indifference

- The machine is still not flexible as it can not translate information into numerical values of probabilities.
- In general, the principle of indifference is not sufficient as there are some reasons for preferring one possibility to another.
- We will introduce two principles giving this ability.
- Note that more principles are yet to be found.

Shannon's theorem for discrete probabilities: the desiderata claims

- We assume that some numerical measure $H(p_1, \ldots, p_n)$ associating "amount of uncertainty" and real number exists.
- We assume a continuity property: H is a continuous function of p_i.
- We require that this measure should correspond qualitatively to common sense in that, when there are many possibilities, we are more uncertain than when there are few.
- We require that the measure be consistent in the same sense as before: if there are more than one way of working out its value, we must get the same answer for every possible way.

Let us start

- Suppose the radio perceives two alternatives, to which it assigns probabilities p_1 and $q = 1 - p_1$. Then the amount of uncertainty represented by this distribution is $H(p_1, q)$.
- But now the radio learns that the second alternative consists of two prossibilities and it assigns probabilities p_2, p_3 to them satisfying $p_2 + p_3 = q$.
- What is the radio's full uncertainty $H(p_1, p_2, p_3)$ as to all three possibilities?

Let us start

- The process of choosing one of the three can be broken into two steps:

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

- This equation is consistent but adds an additional assumption which we did not include in out list i.e that the measure should be additive.

Generalization

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

Since $H(p_1, \ldots, p_n)$ is to be continuous, it will suffice to solve the equation for all rational values $p_i = \frac{n_i}{\sum_i n_i}$ with n_i integer.

Solution

Let us denote $h(n) = H(\frac{1}{n}, \ldots, \frac{1}{n})$, then for a general choice of the n_i, we have:

$$h(\sum n_i) = H(p_1, \ldots, p_m) + \sum_i p_i h(n_i)$$

With the choice $n_i = m$, we have that

$$h(mn) = h(m) + h(n)$$

which is solved with

$$h(n) = K \log(n)$$

with $K > 0$.

Solution

- The solution can be shown to be unique.
- Different choices of K amount to the same thing as taking logarithms to different bases.
- Finally, we have found that the only function $H(p_1, \ldots, p_n)$ satisfying the condition we have imposed on a reasonable measure of "amount of uncertainty" is:

$$H(p_1, \ldots, p_n) = -K \sum_{i=1}^{n} p_i \log(p_i)$$

- Shore and Johnson extended the proof and showed that any other choice of information measure will lead to inconsistencies.

MAXENT in action

Suppose that the mean $\hat{m}$ is given in a problem where m can only take the values $1, 2, 3$. We can use the Lagrangian multiplier argument to get:

$$p_m = \frac{e^{-\lambda m}}{Z(\lambda)}$$

with $Z(\lambda) = \sum_{m=1}^{3} e^{-\lambda m}$ and $p_1 + 2p_2 + 3p_3 = \hat{m}$.

The distribution which has maximum entropy, subject to a given average value, is of exponential form.

MAXENT in action

Suppose a variable x can take on n different discrete values $(x_1, \ldots, x_n)$ which correspond to the n different propositions $(A_1, \ldots, A_n)$ and that there are different function of x,

$$f_k(x)$$

with the expectation

$$F_k = \sum_{i=1}^{n} p_i f_k(x_i)$$

The solution is given by:

$$p_i = \frac{e^{-\sum_{j=1}^{n} \lambda_j f_j(x_i)}}{Z(\lambda_1, \ldots, \lambda_n)}$$

with $Z(\lambda_1, \ldots, \lambda_n) = \sum_{i=1}^{n} e^{-\sum_{j=1}^{n} \lambda_j f_j(x_i)}$.

Remarks and objections

- The information put as constraints may be so meager that no reliable predictions can be made.
- if we emerge with a very broad distribution for some quantity α, then the maximum entropy procedure is telling us that the prior information and data are too meager to permit any inference about α.
- If one has additional information but fails to incorporate it into his calculus, the result is not a failure but a misuse of the MAXENT procedure.

Remarks and objections

- The prior probabilities represent prior information and are to be determined by logical analysis.
- In formulating a problem, the final conclusions depend necessarily on both the prior information and the data.
- Our goal is that inferences are to be completely objective in the sense that two persons with the same prior information must assign the same prior probabilities.

Recalling wave-corpuscule experiment

We do not describe reality-only our information about reality!

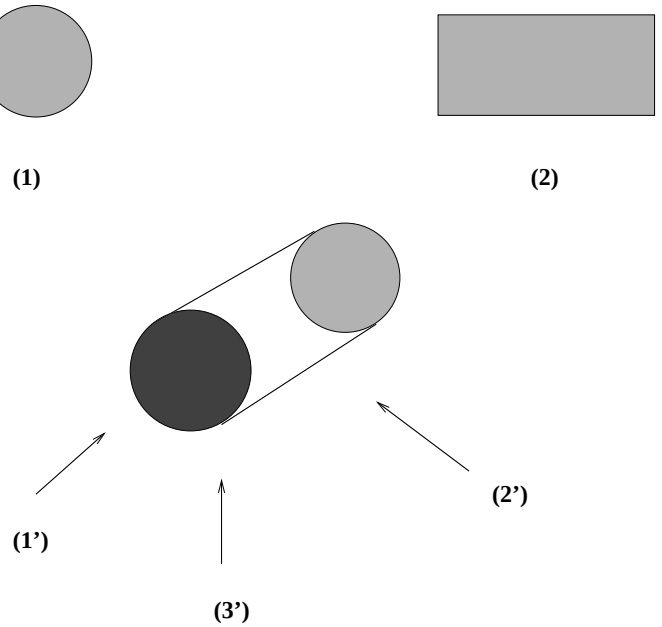


Figure 1: Duality wave-corpuscule?

The channel modelling methodology does not pretend to seek reality but only to represent view (1') or view (2') in the most accurate way.

Continuous distributions

- Unfortunately, the well-know quantity

$$H(P) = -\int dx p(x) \log[p(x)]$$

  lacks invariance under a change of variable $x \rightarrow y(x)$ and can not be used to be a correct information measure.
- In fact, if we change into a new coordinate,

$$H(P) = -\int p(x) J(\frac{x}{y}) \log[p(x) J(\frac{x}{y})] dy$$

  shows that the quantity (J is the Jacobian) is a measure of information relative to an assumed coordinate system.
- Why did Shannon not care about this?

Continuous distributions

- In the case of a model $y = x + n$, the capacity is given by:

$$C = H(x) - H(x \mid y)$$

- Channel capacity depends on the difference of two entropies and this difference does not depend on the coordinate frame, each of the two terms being changed by the same amount!
- Note that Shannon did warn us and called it differential entropy.
- How can we define a notion of uncertainty irrespective of the coordinates?

How are we going to solve our problem?

Well are you sure that:

$$H_d = -\sum_{i=1}^{n} p_i \log(p_i) \rightarrow -\int dx p(x) \log(p(x))?$$

First of all,

$$\lim_{n\rightarrow\infty} \frac{1}{n}\{\text{Number of points in} a < x < b\} = \int_a^b dx g(x)$$

Now,

$$p_i = p(x_i)(x_{i+1} - x_i) \rightarrow p(x_i)\frac{1}{ng(x_i)}$$

which yields

$$H_d = -\int dx p(x) \log(\frac{p(x)}{ng(x)})$$

How are we going to solve our problem?

The continuous information measure is defined as:

$$H_c = -\int dx p(x) \log(\frac{p(x)}{g(x)})$$

- $g(x)$ is called the invariant measure function (we will see why afterwards).
- The continuous information measure is also known as the Kullback distance.
- By taking the relative distance with an invariant measure, information is calculated irrespective of the tool with which we measure.

Continous Maxent in Action

- We seek a probability density which is normalized $\int dx p(x \mid I) = 1$
- Constrained by fixing the mean values of m different functions $f_k(x)$:

$$F_k = \int dx p(x \mid I) f_k(x)$$

- The solution is given by:

$$p(x \mid I) = Z^{-1} g(x) e^{\lambda_1 f_1(x) + \ldots + \lambda_m f_m(x)}$$

with the partition function

$$Z = \int dx g(x) e^{\lambda_1 f_1(x) + \ldots + \lambda_m f_m(x)}$$

Intuitive meaning of the invariant measure

Consider the one-dimensional case and suppose it is known that $a < x < b$ but we have no other prior information. In this case, we have:

$$p(x \mid I) = \frac{g(x)}{\int_a^b g(x)dx}$$

Except for a constant factor, the measure $g(x)$ is also the prior describing ”complete ignorance” and is proportional to the limiting density of discrete points.

Maximum continuous entropy

Note that in many cases, people take $g(x) = 1$ as previously but the result is then only true in the provided coordinate system.

This is of course not consistent. In particular, a change of scale and shift of location should not change the state of knowledge.

How to find therefore $g(x)$, which expresses in some sense the prior distribution describing ”complete ignorance’ of x ?

Remarks

- In the case of discrete probabilities, complete ignorance is represented by a uniform probability assignment.
- For continuous probabilities, the problem is much more difficult as it is not possible to define the finite set of possibilities.
- How can we in this case express complete ignorance?
- A formal theory of how to find express complete ignorance is needed.
- In the discrete case, the principle of maximum entropy is THE solution to the problem of assigning discrete priors whereas in the continuous case, it will rely on transformation groups, marginalization theory and coding theory.

Some remarks

- In our equations, $P(A \mid I)$ is not defined numerically unless alternatives A', A'', etc..are specified (by I). This enumeration of A and all the alternatives to be considered is called the sample space or the hypothesis space. Then it is essentially the principle of indifference or its generalization to Maxent on that space that assigns our initial probabilities.
- The problem of applying maximum entropy is not in deciding what constraints should be applied but in deciding what is the underlying measure (in other words, what is the hypothesis space).
- In problems where we do not have a hypothesis space, we have at present no officially approved way of applying probability theory.
- As a consequence, only relative measures are possible for the moment.
- In spectrum analysis, the Burg solution implied independent uniform weighting to all possible values (relatively therefore to this measure).
- However, the future may bring some surprises here. Any persistent failures would point out to a new hypothesis space and therefore to the possibility of still better predictions.

Invariance measure

- Therefore, the right question in the continuous case is to ask: ”What probability distribution has maximum entropy subject to the basic measure chosen and the constraints imposed?”
- We cannot answer the question ”What prior distribution represents this specific information?” unless we first learn how to answer ”What prior distribution represents complete ignorance?”
- Having answered this, the ”invariance measure” is known and application of the principle of maximum entropy to incorporate specific prior information then becomes independent of our choice of parameters.

How can we still go forward with complete ignorance

- Complete ignorance of a location and scale parameter is a state of knowledge such that a scale and shift of location does not change that state of knowledge.
- If we merely specify ”complete initial ignorance”, we cannot hope to obtain any definite prior distribution, because such a statement is too vague to define any mathematically well-posed problem.
- We are defining this state of knowledge far more precisely if we can specify a set of operations which we recognize as transforming the problem into an equivalent one.
- Having found such a set of operations, the basic desideratum of consistency then places non-trivial restrictions no the form of the prior.

Complete ignorance

- Jeffrey suggested that we assign a prior $\frac{d\sigma}{\sigma}$ to a continuous parameter σ known to be positive, on the grounds that we are the saying the same thing whether we use the parameter σ or σ^m.

Transformation groups: Location and scale parameter

- Complete ignorance needs to be define in a specific way.
- Suppose that we express that a change of scale and shift of location does not change our state of knowledge.

$$\begin{aligned} \mu' &= \mu + b \\ \sigma' &= a\sigma \end{aligned}$$

We would like to know what would be the prior distribution to give to these two parameters.

Transformation groups: Location and scale parameter

We introduce the prior distribution:

$$f(\sigma, \mu) d\sigma d\mu$$

The prior is changed to $g(\mu', \sigma')$, where from the Jacobian we get:

$$g(\mu', \sigma') = \frac{1}{a} f(\mu, \sigma)$$

Know, our desideratum demand (within the consistency framework) that we assign the same probability distribution in them:

$$f(\mu, \sigma) = g(\mu, \sigma)$$

in other words:

$$f(\mu, \sigma) = a f(\mu + b, a\sigma)$$

whose general solution is:

$$f(\mu, \sigma) = \frac{\text{Constant}}{\sigma}$$

known as Jeffrey's rule!

How to still apply maxent?

In any new problem, the strategy is:

- Think hard about the appropriate hypothesis space by looking for some symmetry/invariance properties.
- If the desired useful result appear, then there is no need to point to a different hypothesis space and one is done.
- If the results are unsatisfactory and all the relevant constraints have been taken into account, then this is evidence that Nature is using a different hypothesis space. One must therefore go back to the first step.

Remarks

- One estimates the value of the parameters had when the data were taken. Inference is drawn about what actually did happen and not about what might have happened and did not.
- Before, failure of the predictions was seen as a calamity to be avoided; now we look eagerly for such failures because they tell us new things about the dynamics of the problem.

Information and entropy

Do not confuse information and entropy

- In the bayesian point of view, probability is not taking else than a measure of our state of knowledge (information).
- More information does not necessarily decrease the entropy (measure).
- Example: learning the results of a new poll represents ”more information” but it can make us either more certain or less certain about the result of the election.

Some important (final) remarks

- In the frequentist point of view, ”good statistics” (in the sense that the values are logically independent) are important in hope to average out errors.
- However, it is idle to repeat a physical measurement an enormous number of times. Only (reliable) information should be taken into account.
- http://youtube.com/watch?v=dMFsuBlkoIQ
- Ten million uninformed opinions are not as good as one expert opinion (a fact by the way many politicians and pollsters have forgotten).

Application Example

Cognitive radios..coming soon..