# VERS UN PARADIGME SYNTHETIQUE DE LA COGNITION : L'INFERENCE PROBABILISTE

# TOWARD A SYNTHETIC COGNITIVE PARADIGM: PROBABILISTIC INFERENCE

**Pierre Bessière**[1]

**RESUME :**

Les sciences cognitives connaissent actuellement un renouveau certain. Elles sont à la croisée des chemins d'idées venues de domaines scientifiques très divers. On peut espérer que cet intense échange interdisciplinaire entrainera l'émergence de nouveaux paradigmes cognitifs. Le but de cet article est de répondre à la question : existe-t-il des indices indiquant l'émergence de tels nouveaux paradigmes? La principale thèse de ce papier est qu'il existe au moins un candidat intéressant pour ce rôle : l'**inférence probabiliste**.

Afin d'argumenter la thèse ci-dessus la structure de l'article est la suivante :

- dans une première partie, nous proposons cinq critères pour reconnaitre un paradigme cognitif intéressant (la validité, la consistence, la compétence, la faisabilité et la faculté mimétique) ;
- au second paragraphe, les principes de l'inférence probabiliste sont exposés et les critères de validité et de consistence sont discutés ("computational level" suivant Marr) ;
- le critère de compétence est abordé ensuite en montrant la capacité de l'inférence probabiliste à traiter les classiques "énigmes" cognitives et en la comparant avec plusieurs autres théories de la cognition ("algorithmic level") ;
- enfin, la faisabilité (possibilité et conditions d'informatisation) et la faculté mimétique (adéquation avec ce que l'on connait de l'architecture du système nerveux) sont débatues dans la quatrième partie ("implementation level").

En conclusion, il apparaîtra que l'inférence probabiliste est, pour le moins, un cadre très intéressant pour acquérir une vue synthétique de nombreux travaux en cours dans ce domaine et pour identifier et formaliser certaines des questions les plus brûlantes. Certains de ces problèmes seront énumérés. En fait, l'inférence probabiliste apparaîtra finalement, comme pouvant jouer pour la neuro-informatique le même rôle que la logique formelle joue pour l'intelligence artificielle symbolique: le rôle d'une théorie mathématique saine qui sert de fondement et de ligne directrice et est utilisée comme constante référence et comme source d'inspiration.

Mots clefs : science cognitive, réseau de neurones, inférence probabiliste, entropie

**ABSTRACT:**

Cognitive science is a very active field of scientific interest. It turns out to be a "melting pot" of ideas coming from very different areas. One of the principal hopes is that some synthetic cognitive paradigms will emerge from this interdisciplinary "brain storming". The goal of this paper is to answer the question: "Given the state of the art, is there any hints indicating the emergence of such synthetic paradigms?" The main thesis of the paper is that there is a good candidate, namely, the **probabilistic inference** paradigm.

In support of the above thesis the structure of the paper is as follows:

- in a first part, we identify five criteria to qualify as a synthetic cognitive paradigm (validity, self consistency, competence, feasibility and mimetic power);
- in the second paragraph, the principles of probabilistic inference are reviewed and justifications of validity and self consistency of this paradigm are given (Marr's computational level);
- then, the competence criterion is discussed, considering the efficiency of probabilistic inference for dealing with the different classical cognitive riddles and analyzing the relationships of probabilistic inference with several of the usual connexionist formalisms (Marr's algorithmic level);
- the criteria of feasibility (condition of computer implementation) and mimetic power (adequation with what is known of the architecture of the nervous system) are finally considered in the fourth part (Marr's implementation level).

As a conclusion, it will appear that probabilistic inference is at least a very interesting framework to get a synthetic overview of a number of works in the area and to identify and formalize the most puzzling questions. Some of these questions will be listed. In fact, probabilistic inference will appear finally to be able to play the same role for computational cognitive science that formal logic has played for classical symbolic Artificial Intelligence: a sound mathematical foundation serving as a guide line, as a constant reference and as a source of inspiration.

Keywords: cognitive science, neural network, probabilistic inference, entropy,

---

[1]*L.G.I. - I.M.A.G. / LA.S.CO.3 (Laboratoire de Génie Informatique - Institut d'Informatique et de Mathématique Appliquée de Grenoble / LAboratoire de Sciences COgnitive) BP53X, F-38041 Grenoble Cedex, FRANCE.*
*E-mail: bessiere@imag.imag.fr; Phone: 33-76.51.45.72; Fax: 33-76.44.66.75; Telex: UJF 980 134 F*

## I. SEEKING FOR A SYNTHETIC PARADIGM OF COGNITIVE SCIENCES

The purpose of cognitive science is to get a better understanding of advanced information processing systems. Natural information processing systems, given their amazing capabilities and overwhelming complexity, are the privileged subject of study. The human brain is the archetype of such structures and consequently is the center of interest and the constant reference.

Interdisciplinarity of such a scientific project is obvious. Occasional cooperation between specialists of closely related subjects is not sufficient. What is really needed is to have teams of scientists from the humanities, from the life sciences and from engineering sciences, working together toward this goal.

As for most scientific projects of any importance, there is a technological counterpart to cognitives sciences. Its goal is to imagine new kinds of computation and to build new types of machines in order to solve practical problems yet unsolved with classical techniques.

A synthetic cognitive paradigm based on a sound mathematical theory is definitely needed. Such a metaphor would be useful:

- as a common language between the different scientific "cultures" concerned;
- as a reference to compare and contrast the different models and practical implementation;
- as a guide line for forthcoming research by formalizing the most puzzling questions;
- as a source of inspiration for finding answers and solutions to those questions.

An interesting parallel can be made with classical Artificial Intelligence (A.I.), where the concept of formal system is the reference paradigm. This mathematical theory has been the constant reference and the source of inspiration of most of the developments in symbolic computing. Most of the applied A.I. systems can be seen has formal systems where some constraints has been relaxed, some "adhoceries" have been added to face a given practical problem, some programming tricks have been used to improve the performances (see [Bessière90a]).

Looking for a synthetic cognitive paradigm, we must first identify the necessary criteria to qualify as such.

A synthetic cognitive paradigm should propose a formal description of a way to code knowledge, a way to acquire and stock knowledge and a way to do reasoning, using the previously acquired knowledge.

Furthermore, such a synthetic cognitive paradigm should prove, both, its mathematical soundness and its utility. Mathematical soundness, on one hand, may be divided into the two usual criteria: self-consistency and validity. Utility, on the other hand, may be divided into three criteria: competence, feasibility and mimetic power. Competence concerns the efficiency of the proposed paradigm to deal with the different cognitive riddles; feasibility concerns the possibility to implement the described paradigm either by software or by hardware; and mimetic power concerns the adequateness of the proposed theory with the supposed architecture of the nervous system.

In what follows, we will analyse probabilistic inference, using those different criteria: self-consistency and validity in section II, competence in section III and feasibility and mimetic power in section IV.

## II. PROBABILISTIC INFERENCE

In this section we introduce the bases of probabilistic inference. The description given here is formulated in mathematical terms. It corresponds to a formal analysis of a well defined problem. This may be summed up by saying that the level of description of this section corresponds to Marr's computational level (see [Marr82]).

### II.1. Representation of knowledge

In probabilistic inference theory, knowledge is represented using the usual formalism of probability theory. Knowledge is encoded using a set $\Xi = \{X_1, ..., X_n\}$ of variables. Those variables may have discrete or continuous domains. The value space of $\Xi$ is called $\Omega$. The basic assumption is that a knowledge state of a cognitive system is defined as being a probability distribution P over $\Omega$.

Probabilistic inference has to deal with two different problems:

1 - given a knowledge state (a probability distribution P) and some new information (a set of constraints $\Phi$ on $\Xi$), how to infer a new knowledge state (a probability distribution Q) taking into account the new information;

2 - given a knowledge state (a probability distribution P) and values of some of the variables of the set $\Xi$, how to infer the most probable values (according to P) of the other variables.

The first problem may be called the "dynamic inference problem" because it concerns how the knowledge state changes in order to take into account new information, while the second problem, the "static inference problem", concerns the consistency conditions of a knowledge state at a given time (see [Hunter86]).

### II.2. Dynamic inference problem

Given a set of variables $\Xi = \{X_1, ..., X_n\}$, given $\Omega$ its value space, given a prior knowledge state (a probability distribution P) and given some new information (a set of constraints $\Phi$ on $\Xi$); the dynamic inference process has to find a posterior knowledge state (a probability distribution Q).

In the general case, there is an infinity of probability distributions which are potential solutions of this problem. However, all the probability distributions are not equivalent. Some appear to be more "coherent", more "probable", more "interesting" than some others. Those notions will be explained, defined and discussed in § II.4..

For the time being, let us say that we have a function H(Q,P) (called Kulbach entropy, relative entropy or cross entropy) which is a way of measuring the "interest" of a given probability distribution Q relative to P and $\Phi$, the smaller H the better Q. H is defined by: $H(Q,P) = + \int_{\Omega} Q(\omega) \log \frac{Q(\omega)}{P(\omega)} d\omega$ [f.1] in the continuous case; and by: $H(Q,P) = + \sum_{i=1}^{N} Q(\omega_i) \log \frac{Q(\omega_i)}{P(\omega_i)}$ [f.2] in the discrete case, where N is the cardinal of $\Omega$ and where $\omega_i$ is an element of the set $\Omega$.

According to this, the dynamic inference problem may be restated as follow: given P and $\Phi$, find the probability distribution Q which **minimizes** H. It can be shown that if $\Phi$ is a consistent set of constraints, there is one and only one solution Q* to this problem. Finding Q* is not a trivial mathematical problem.

However, for a very important class of problems, where $\Phi$ takes the form of a set of real functions ($\Phi = \{f_1, ..., f_m\}$) such that the mean value $a_i$ ($A = \{a_1, a_2, ..., a_m\}$) of every function $f_i$ is known, then it can be shown that the solution take the following form: $q^*(\omega) = \frac{1}{Z^*} e^{- \sum_{i=1}^{m} \lambda_i f_i(\omega)}$ [f.3] in the continuous case, where q* is the density of Q*, $\lambda_i$ are the Lagrange multipliers and Z* is a normalizing constant; and $Q^*(\omega_j) = \frac{1}{Z^*} e^{- \sum_{i=1}^{m} \lambda_i f_i(\omega_j)}$ [f.4] in the discrete case[1].

*Let us take an example: given two variables A (the Age of the captain) and L (the Length of the boat), the problem we want to solve is find A given L or find L given A.*

*We have: $\Xi = \{A,L\}$; $\Omega = [7,77] \times [4,444]$.*

*Starting from scratch (no prior information), $P_0$, the initial knowledge state, is a uniform distribution over $\Omega$ (figure 1). Let us suppose that we first learn the mean value of A ($E(A) = m_A$) and the variance of A ($E((A - m_A)^2) = \sigma_A^2$). An infinity of probability distributions over $\Omega$ have this mean value and this variance. However one and only one minimizes H: the normal distribution $P_1$ having this mean value and this variance as parameters (figure 2). If we then learn the mean value and variance of L, by the same process, we get the probability distribution $P_2$ (figure 3). Iterating this process for all the data we can get about our problem, and especially for information expressing correlations between A and L we will finally get a probability distribution Q* which will sum up all the previously acquired information (figure 4).*

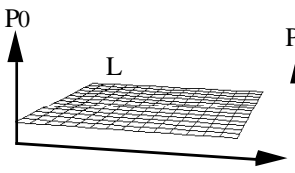

Figure 1

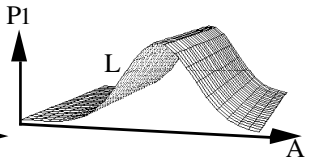

Figure 2

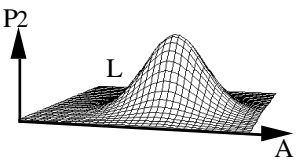

Figure 3

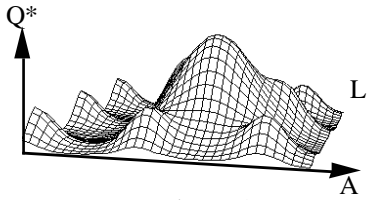

Figure 4

II.3. Static inference problem

Given a set of variables $\Xi = \{X_1, ..., X_n\}$, given $\Omega$ its value space, given P a probability distribution over $\Omega$ and given values of some of the variables of the set $\Xi$; the static inference process has to find the most probable values of the unspecified variables of $\Xi$.

According to the previous paragraph, the interesting cases to consider are those cases where P takes either the form [f.3] or the form [f.4].

Therefore, finding the most probable values of the non specified variables (i.e. maximizing P) corresponds to the **minimization** of the function: $U(\omega) = \sum_{i=1}^{m} \lambda_i f_i(\omega)$ [f.5] over the sub-space of $\Omega$ defined by the given values on $\Xi$.

---

[1] *See, for instance, [Jaynes79] for a more detailed explanation of this in the discrete case and [Robert90b] for the continuous case.*

*Getting back to our example, this means that for a given value a of the variable A we want to find the most probable value l of the variable L. This process may be visualized by looking for the maximum of the curve defined by the intersection of Q* and the vertical plane corresponding to A = a (this curve is in bold on figure 5).*

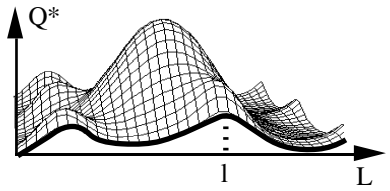

Figure 5

II.4. Validity of probabilistic inference

In the two preceding subsections, we gave a mathematical description of the dynamic and static inference processes. However, the entire approach is based on the "magic" function H accepted in § II.2. as a means for measuring the relative "interest" of different probability distributions. We then postponed the discussion about the validity of H. This discussion is the object of this section.

*II.4.1. Combinatorial argument*

The combinatorial argument (see [Jaynes79]) is directly related to the historical background of probabilistic inference, namely, mechanical statistics. It was, in fact, first proposed by Boltzmann.

Let us suppose that we have a set of p identical particles and that each of these particles may have q different equiprobable states. We are interested in the state of one of these particles denoted by variable X ($\Xi = \{X\}$ & $\Omega = \{1, 2, ..., q\}$).

Let us define $\nu_k$ as a set $\{p_1, p_2, ..., p_q\}$ of q numbers such that $p_i$ is the number of particles in state i.

Let us call $W(\nu_k)$ the number of ways in which $\nu_k$ may be realized. We have: $W(\nu_k) = \frac{p!}{p_1^k!\, p_2^k! \ldots p_q^k!}$ [f.6] with:

$\sum_{i=1}^{q} p_i^k = p$ [f.7][1].

Boltzmann's position is that the most probable distribution $\nu_k$ is the one which may be realized in the greatest number of ways, i.e. the one which maximizes $W(\nu_k)$, or, using the Stirling formula, the one which maximizes:

$\log (W(\nu_k)) \approx - p \sum_{i=1}^{q} \frac{p_i^k}{p} \log \frac{p_i^k}{p}$ [f.8].

The summation in [f.8] is a simplified form[2] of H corresponding to the restricted problem considered in this paragraph.

For p = 1000 and q = 2, if we call $\nu_1$ the distribution such that all the particles are in the same state and if we call $\nu_2$ the distribution such that 500 particles are in one state and 500 others are in the other state, we get: $W(\nu_1) = 1$ and $W(\nu_2) \approx 2^{1000}$ ; $\nu_2$ seems, indeed, to be more probable than $\nu_1$.

We should notice that all this is only an argument, a hint, definitely not a proof. However, the combinatorial argument is very interesting because it gives a good intuitive flavor of the meaning of the entropy concentration theorems of next section.

*II.4.2. Entropy concentration theorems*

Edwin T. Jaynes in [Jaynes82] proves the following entropy concentration theorem:

- let us suppose that we are interested in p trials of a random process described by a variable X and that this variable X may take q different equiprobable values,
- let us define $\nu_k$ as a set $\{p_1, p_2, ..., p_q\}$ of q numbers such that $p_i$ is the number of trials having value i. Let us define $\phi_k$ as the associated set of frequencies $\{f_1, ..., f_q\}$ such that $f_i = p_i / p$;
- for a given distribution $\nu_k$, we have the associated entropy: $H(\nu_k) = H(\phi_k) = - \sum_{i=1}^{q} f_i \log f_i$ [f.9];

---

[1] *In physics, there is usually an aditional constraint in terms of conservation of total energy of the form :* $\sum_{i=1}^{q} p_i^k e_i^k = E$

[2] *Expressed in terms of numbers of possibilities rather than in terms of probabilities and assuming that there is no prior knowledge.*

- let us say that our problem is constrained by a set of r linear constraints of the form: $\sum_{i=1}^{q} b_{ij} f_i = a_j \; 1 \le j \le r$ [f.10];

- let us call F the ratio of the number of distributions $\phi_k$ having their entropy in the interval $H_{max} - \Delta H \le H(\phi_k) \le H_{max}$ [f11] to the number of all possible distributions;

The concentration entropy theorem state that $\Delta H$ is related to F by the following relation: $\Delta H = \frac{\chi^2_s (1 - F)}{2p}$ [f.12] where $\chi^2$ is Pearson's chi-squared and s = q - r - 1 is the number of degrees of freedom taking into account the r constraints.

Claudine Robert in [Robert90b], demonstrates, using large deviation techniques, a very important generalization of this theorem. Unfortunately, it's too long to be restated here.

These theorems may be seen as mathematical justifications of the use of the Maximum Entropy Principle to deal with the dynamic inference problem. They prove that the probability distributions are concentrated around the maximum entropy distributions and moreover quantify the number of such distributions in a given entropy interval. In fact, they demonstrate that the maximum entropy distribution is the **most probable one** given the available informations, and the **only one that doesn't assume any information that hasn't been given**.

*II.4.3. Argument of coherence with conditional probability and axiomatic justification*

Two others very interesting justifications of the use of probabilistic inference should be mentioned. Unfortunately, they can not be developed in the context of this paper:

- Edwin T. Jaynes in [Jaynes79] and Van Campenhout in [Van Campenhout81] develop arguments to justify the use of H, founded on proofs of coherence between probabilistic inference and conditional probability as basically expressed by Baye's theorem;
- John E. Shore (see [Shore80], [Shore81] & [Shore86]) goes further in the same direction when he proposes a justification of the use of H on an axiomatic basis. He postulates four axioms, namely, uniqueness, invariance, system independence and subset independence and derives from these axioms the form of function H.

## III. PROBABILISTIC INFERENCE AND THE COGNITIVE RIDDLES

In this section probabilistic inference is reviewed at Marr's algorithmic level [Marr82]. Our concerns will therefore be:

- how should data be represented and what are the consequences of the choices made at this level?
- how may probabilistic inference be practically used?
- how does probabilistic inference deal with the different cognitive riddles?
- how may it be compared with the other cognitive models?

### III.1. Representation of knowledge

Given that the purpose of probabilistic inference is to do reasoning on incomplete and uncertain knowledge, the choice of probability formalism for encoding knowledge may seem obvious. However, it is worth noticing that some other formalisms may be considered (fuzzy logic [Zadeh65], [Goguen67] & [Dubois80]; possibility theory [Zadeh78])[1].

More unusual is the use of a probability distribution to model a knowledge state of a cognitive system. Such a conception of probability is not, even at present, shared by everyone. The "frequentist" approach of probability has still a lot of supporters and was, only a few years ago, the only way around to think about probability[2]. Nevertheless, this conception is not recent and may be traced back to Bayes.

It should be noticed that the choice of probability formalism, at the earliest stage of cognitive modelling ("representation of knowledge"), is of prime importance. Actually, it is a definite breakthrough compared to formal symbolic artificial intelligence, in order to deal with incomplete and uncertain knowledge.

Some efforts have been made to merge formal logic and probabilistic reasoning. A flavor of those works may be given by the simple observation that "A => B" is equivalent to "P(B/A) = 1"; logical rules may be seen as special cases of conditional probabilities where probability values are either equal to 1 or 0 (see [Cox79]). Practical applications of this have been made to include production rules as expressed by medical experts in a daily used probabilistic expert system devoted to diagnosis of children meningitis [Robert90a].

### III.2. Dynamic inference problem

At this point in the paper, it should be clear that dynamic inference process is a model of "learning by experience", in the sense that it is a way to evolve from one knowledge state to a new one, taking into account new information. In some sense, it is also a model of "memory", because it proposes a means of stocking information, namely, probability

---

[1] *There is no room in this paper to get into the passionate argument about the relative merits of probability theory and fuzzy theories. For arguments in favor of probability see, for instance, [Cox61], [deFinetti72] , [Cheeseman85] and [Robert90c].*

[2] *See [Jaynes79] for a fascinating epistemological study of the evolution of these two paradigms of probability.*

distributions. Finally, physical systems that support this dynamic inference process will see their internal structures, encoding the probability distributions (see § IV.), "self-organize" as new information is provided.

*III.2.1. Using H as the objective function*

The Dynamic Inference problem has been shown to be an optimization problem (c.f. § II.2.). The search space is the space of all possible probability distributions over $\Omega$.

H, the Kullbach entropy, is the best possible objective function for this optimization problem in the sense of the entropy concentration theorems of § II.3..

However, it has been shown that this optimization problem may be dealt with only when $\Phi$ takes the specific form of a set of real functions such that the mean values of every function are known (c.f. § II.2.). In that case we know that the solution sought has the following form: $q^*(\omega) = \frac{1}{Z^*} e^{-\sum_{i=1}^{m} \lambda_i f_i(\omega)}$

When the sets $\Phi$ and A are explicitly stated, the maximum entropy principle appears to justify the use of the most popular distributions:

- the normal distribution with mean 0 and variance $\sigma^2$ is the maximum entropy distribution under the constraint $E(X^2) = \sigma^2$;
- the exponential distribution with parameter 1/m is the maximum entropy distribution under the constraint of a non negative variable with mean m;
- the distribution of the air density as a function of height in the atmosphere, is the maximum entropy distribution under the proper energy constraint;
- etc... (see [Jaynes79] & [Van Campenhout81]).

**- a - "Discovering" the observables ($f_i$)**

As we just said, the $f_i$ may be explicitly stated by the problem. However, this is generally not the case. Most often, the $f_i$ are implicitly given by a set of instances, of prototypes. In that case, "discovering" the $f_i$ is part of the Dynamic Inference problem. The $f_i$ are called "observables" ([Robert90a] & [Robert90b]) by analogy with physics where the $f_i$ are the measurable macroscopic quantities which are accessible to the observer.

In fact, it is the most difficult part. No general satisfactory solution has yet been proposed to solve this problem.

A first solution to this problem, the most generally adopted, is to restrict the problem to cases where the $X_i$ are binary variables and where the $f_i$ depend on no more than two variables: $f_i(X_1, ..., X_n) = \alpha_i X_k X_l + \beta_i X_k + \gamma_i X_l + \delta_i$ [f.13].

In that case we have: $Q^*(\omega) = \frac{1}{Z^*} e^{-\sum_{i<j} W_{ij} X_i X_j - \sum_i W_i X_i}$ [f.14] where Q* is a 1-Gibbs distribution. Such distributions have well known dynamics with nice simple properties. This is the choice made, for instance, for Hopfield nets [Hopfield82] and for Boltzmann machines [Hinton87]. They have been proved to be equivalent to Markoff Random Field (M.R.F.) [Geman84]. The dynamic inference problem has new parameters, the "weights" $W_{ij}$, in place of the old ones, the Lagrange multipliers $\lambda_i$.

A second solution as been proposed in Harmony theory ([Smolensky86]). The relevant observables ($f_i$) are explicitly given by the programmer. Actually, the programmer has to specify a "knowledge vector" which is a list of parameters (-1, 0 or 1) stating how the "knowledge atoms" depend on the "representational features". The drawback of this approach is that, often, for a number of practical problems, the designer of the system is unable to specify the relevant observables.

In observable networks (see [Robert90a], [Robert90b] & [Robert90c]) some of the observables are explicitly stated by the experts as in harmony theory, some are stated using inference rules, but, furthermore, some other observables are "discovered" using standard data analysis methods (mean square estimation, factorial discriminant analysis, etc...) on sets of prototypes. This is a very elegant solution to the drawback noted in the previous paragraph.

**- b - "Learning" the parameters ($a_i$, $\lambda_i$ or $W_{ij}$)**

As for the $f_i$, most of the times the $a_i$ are not explicitly stated by the problem, but are rather implicitly given by a set of prototypes. The second part of the dynamic inference problem is to learn either the $a_i$ or the $\lambda_i$ from this set of instances.

In the case of Gibbsian neural nets the parameters $\lambda_i$ are replaced by the parameters $W_{ij}$. The search space, which was the space of all probability distributions over $\Omega$, is approximated by the space of weights' values. The dynamic inference process is replaced by a gradient descent dynamic in the space of the weights' values with H as objective function. One of the remarkable properties of the Gibbsian neural nets is that this gradient descent may be done, without any explicit computation of H, using simply classical local adaptation rules (for instance, Widrow-Hoff's rule for Boltzmann machines) on the system at "thermodynamic" equilibrium (see [Hinton87]). Reaching thermodynamic equilibrium itself is strictly equivalent to treating the static inference process and is done using simulated annealing (see § III.3.).

For harmony theory the parameters considered are either the knowledge atoms' strength $\sigma$ of "Harmonium" or the coefficients $\lambda$ of the "Simulation Machine". The learning algorithm is very similar to that of the Boltzmann machine.

For observable networks, acquisition of parameters $\lambda_i$ occur at the same time as observables are discovered, as the common result of the same calculus.

*III.2.2. Using other objective functions*

In a certain sense (see § II.4.) H has been proved to be the best possible objective function for the Dynamic Inference problem. However, other objective functions may be considered. In fact, in an other sense, H is one of the worst possible objective function, because, in the general case, it is a very difficult function to minimize.

One may rather try to find a specific objective function (or its associated learning rule and algorithm) adapted to a particular practical concern. This concern might be, for instance, to maximize the information storage capacity (see [Gardner89b]) or to optimize convergence of the dynamic (see, for instance, the "Minover algorithm" in [Krauth87] or some works by E. Gardner in [Gardner89a]). This is the guide line of the mechanical statistic approach of neural networks[1].

For supervised learning algorithms, like the family of back-propagation algorithms (see, for instance, [Rumelhart87]), the learning mechanism is also a dynamic process on the space of weights' values. Gradient descent techniques are used in this space to minimize the quadratic difference between the actual and the desired output. This difference is the objective function.

III.3. Static inference problem

The static inference process is a way to retrieve information previously stocked during learning. Considering the nature and the number of variables X of $\Xi$ which are missing, it can be either seen as a model of recalling information from an associative memory, or as a model of pattern recognition (classification), or, finally, as a model of the very general concept of pattern association which may cover, for instance, decision-making.

The static inference problem has been shown to be an optimization problem (c.f. § II.3.). The search space is $\Omega$. The objective function is the probability distribution learned in the previous phase by the dynamic inference process.

If H was the objective function of the dynamic inference process and if $\Phi$ had the adequate form, then we may considered that the objective function is $U(\omega)$ (c.f. [f5] in § III.3.). The search space is not $\Omega$, but $\Theta$ sub-space of $\Omega$ defined by the known and imposed values of the variables $X_i$. The static inference problem is to find the optimum of U on $\Theta$.

For Gibbsian neural nets, U has the exact form of an energy function. The minimum of U, is searched like an energy minimum using, for instance, the well known, simulated annealing algorithm. Simulated annealing is also used by Harmony theory. However, any other optimization algorithms like, for instance, genetic algorithms, could be used as well, leading in some cases to much better results (see [Ackley87]).

A large class of neural networks, namely, feedforward networks have a very simplified dynamic for the static inference process. This dynamic is reduced to a one path propagation of activity as there is no feed back.

## IV. PROBABILISTIC INFERENCE: OPENING THE "BLACK BOX"

In this section probabilistic inference is reviewed at Marr's implementation level [Marr82]. Our concern will be on one hand the conditions of computer implementation of the two previously described processes and, on the other hand, the eventual adequateness of these mathematical models with what is known of the nervous system's architecture.

IV.1. Representation of knowledge

For practical computer implementation of probabilistic inference, two main options for representing knowledge may be considered.

The first one is to stick close to theory, using probability formalism as it is. In that case, many out of the shell statistical and optimization softwares are available. Furthermore, a lot of know-how has been acquired by statisticians. Hence, this option deserves very serious consideration.

However, at present time, the second option, namely, artificial neural network is more in fashion. Variables X are represented as activity of cells and probability distributions on $\Omega$ are approximated by "weights" on links between those cells.

Two main reasons may explain why the second solution is more in favor than the first one:

- artificial neural networks are intrinsically parallel, even if their parallelism is often difficult to map onto the available parallelism of existing machines;
- artificial neural networks may easily be considered as a model of biological nervous systems, even if they appear to be drastic simplifications.

One of the most disturbing simplifications is that in most artificial neural network models, the activity of a neuron is represented only by a level of activation, corresponding to a mean firing rate. This seems to be biologically very unrealistic, because there is great evidences that synchronicity and temporal distribution of the neural pulses play a very important role in information coding in the brain (see, for instance [Abeles88a]). This simplification is made by computer scientists for obvious reasons of software simplicity, but is also often made by neurophysiologists because the experimental study of temporal distribution of pulses in assembly of neurons is extremely difficult [Abeles88b]. This question of information coding in the brain is of the main importance, and solving it could be a very important

[1] *see the special issue of Journal of Physics A, Volume 22, 1989 in memory of the late E. Gardner for an overview of this subject.*

breakthrough for cognitive sciences. Works are under way to propose mathematical formalisms allowing to model, in detail, spatio-temporal distributions of pulses (see, for instance, [Hervé90]).

IV.2. Dynamic and static inference problem

Dynamic and static inference algorithms are made of three embedded loops:

- loop 1, corresponding to the learning dynamic, iterates on the different prototypes (or on the different mean values $a_i$ of observables $f_i$);
  - loop 2, corresponding to the static inference dynamic, iterates for different parameters (for instance, decreasing temperature in simulated algorithm) searching for the most probable states of the variables;
    - loop 3, is an iteration on each cell to propagate activity from cells to cells.

Loop 1 may be sometimes suppressed, given that some algorithms, like back-propagation, use to update the weights, a linear combination of each individual training prototypes' contributions. In these cases, the different training patterns may be processed in parallel. This is called "training set parallelism" (see [Singer90]).

However, loop 3 is where can be found the so called "inherent" parallelism of artificial neural networks. Each cell is considered as an independent process, running in parallel with all the others and exchanging information through communication links.

The main difficulty encountered to implement such algorithms is due to the number of processes, and overall, to the overwhelming number of communication links that are needed. That is why, most parallel implementations take advantage of the fact that for numerous neural algorithms, communication between cells may be represented by operations on vectors and matrix. Hence, most parallel implementation of neural networks are in fact parallel implementation of matrix operations. More biologically realistic implementations of neural networks using one process (or better one processor) for each cell and having the wanted number of communication links, seem to be, for the time being, beyond the scope of available parallel machines; even if works are going in this direction (see, for instance [Bessière90b]).

Reversing the problem, let us wonder if in any way, the nervous system may be considered as "a dynamical complex system solving optimization problems like the dynamic and static inference problems"? We obviously can not get in that question in a few word at the end of this paper, let us just say that the "Darwinian" approach of cognition seems to be on a not so different track (see [Changeux76], [Edelman78], [Changeux83], [Changeux84] & [Edelman87]).

## V. CONCLUSION

At the end of this paper, we hope that you are convinced that probabilistic inference is at least a very interesting framework for a synthetic overview of a number of works in the area of neuro-computing and that it is worth considering it as a good candidate for a synthetic paradigm for cognitive science. However, a lot of work has still to be done on this base and especially, a number of comparisons with classical techniques have been just sketched and need to be much further examined.

On the basis of probabilistic inference, some probing questions may be asked and formalized. Let us just list few of them:

- What knowledge representations will be best for computer implementation of probabilistic inference like models?
- Do these representations have any biological plausibility? Do they suggest any new ideas about information coding in the brain?
- How may we discover the relevant observables? How can we take into account, in probabilistic inference systems, knowledge and inference rules as expressed by human experts?
- What are the "good" objective functions for the dynamic inference problem? May we associate types of objective functions with types of problems?
- Among all optimization technics, how can we choose? Dynamic and static inference process correspond to very different search spaces, does this make a difference to pick up the right algorithm?
- May we substitute the debate "formal logic versus probabilistic inference" for the debate "symbolic versus sub-symbolic systems"? does it bring any new light?
- Can we formalize some ideas of the "Darwinian" approach of cognition?
- Finally, given that neural networks are dynamical systems, do not we have much to learn from works done about non equilibrium dynamical systems (see, for instance, [Nicolis77])?

# BIBLIOGRAPHY


[Abeles88a] Moishe Abeles
*Neural codes for higher brain functions*
in *Information processing by the brain* edited by H.J. Markowitsh, Hans Huber Publishers, 1988

[Abeles 88b] Moshe Abeles and George L. Gerstein
*Detecting spatiotemporal firing patterns among simultaneously recorded single neurons*
Journal of Neurophysiology, Vol 60, N°3, 1988

[Ackley87] David H. Ackley
*A connectionist machine for genetic hillclimbing*
Kluewer Academic Publishers, 1987

[Bessière90a] Pierre Bessière
*Un possible paradigme synthetique pour le connexionisme: l'Inférence Probabiliste*
Proceedings of N.S.I.90, 1990

[Bessière90b] P. Bessière, A. Chams, A. Guérin, J. Hérault, C. Jutten & J-C. Lawson
*From hardware to sftware: designing a "Neurostation"*
in *Introduction to V.L.S.I. design of artificial neural networks*, Kluever Academic Publishers, 1990

[Changeux76] J-P. Changeux & A. Danchin
*Selective stabilisation of developping synapses as a mechanism for the specification of neural networks*
Nature, N° 264, 1976

[Changeux83] J-P. Changeux
*L'homme neuronal*
Fayard, Paris, 1983

[Changeux84] J-P. Changeux, T. Heidman & P. Patte
*Learning by selection*
in *The biology of learning* edited by P. Marler & H.S. Terrace; Springer Verlag, 1984

[Cheeseman85] P. Cheeseman
*In defense of Probability*
Proceedings of AAAI85, 1985

[Cox61] R. T. Cox
*The algebra of probable inference*
The John Hopkins Press, Baltimore, 1961

[Cox79] R. T. Cox
*Of inference and inquiry, an essay in inductive logic*
in *The maximum entropy formalism*, edited by Raphael D. Levine & Myron Tribus; M.I.T. Press, 1979

[deFinetti72] B. de Finetti
*Probability, induction and statistics*
John Willey & sons, 1972

[Dubois80] D. Dubois & H. Prade
*Fuzzy sets and systems: theory and applications*
Academic Press, New York, 1980

[Edelman78] G. Edelman
*Group selection and phasic reentrant signalling: a theory of higher brain function*
M.I.T. Press, 1978

[Edelman87] G. Edelman
*Neural Darwinism*
Basic Books, New York, 1987

[Gardner89a] E. Gardner
*Optimal basins of attraction in randmly sparse neural network models*
Journal of physics A, volume 22, number 12, 1989

[Gardner89b] E. Gardner
*Three unfinished works on the optimal storage capacity of networks*
Journal of physics A, volume 22, number 12, 1989

[Geman84] S. Geman & D. Geman
*Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images*
IEEE transactions on pattern analysis and machine intelligence, 6, 1984

[Goguen67] J.A. Goguen
*L-fuzzy sets*
Journal of mathematical analysis and applications, N°18, 1967

[hervé90] T. Hervé, J.M. Dolmazon & J. Demongeot
*Random field and neural information*
Proc. Natl. Acad. of Sciences, U.S.A., Vol 87, 1990

[Hinton87] G. E. Hinton & T. J. Sejnowski
*Learning and relearning in Boltzmann machines*
in *Parallel Distributed Processing*
by D.E. Rumelhart, J.L. McClelland & the P.D.P. Research Group; The M.I.T. Press, 1986

[Hopfield82] J. J. Hopfield
*Neural networks and physical systems with emergent collective computational abilities*
Proceedings of the National Academy of Sciences (U.S.A.), 1989

[Hunter86] Daniel Hunter
*Uncertain reasoning using Maximum entropy Inference*
in *Uncertainty in Artificial Intelligence*; edited by L. N. Kanal & J. F. Lemmer ; Elsevier Science Publishers, 1986

[Jaynes79] Edwin T. Jaynes
*Where do we stand on maximum entropy?*
in *The maximum entropy formalism*; edited by Raphael D. Levine & Myron Tribus; M.I.T. Press, 1979

[Jaynes82] E. T. Jaynes
*On the rationale of maximum-entropy methods*
Proceedings of the IEEE, 1982

[Krauth87] W. Krauth & M. Mezard
*MIinoverl algorithm*
Jouranal of physics A, Volume 20, 1987

[Marr82] D. Marr
*Vision*
W.H. Freeman & Co, 1982

[Nicolis77] G. Nicolis & I. Prigogine
*Self-organisation in nonequilibrium systems*
John Wiley & sons, 1977

[Robert90a] Claudine Robert, B. Cremilleux, P. François & J. Demongeot
*Markof Random fields for medical decision making: Observable Networks*
Proceedings of the 11th Prague conference on information theory, statistical decision functions and random processes

[Robert90b] Claudine Robert
*An entropy concentration theorem: applications in artificial intelligence and descriptive statistics*
Journal of Applied Probabilities, Vol. 27, 1990

[Robert90c] Claudine Robert
*Méthodes statistiques pour l'intelligence artificielle; application en médecine.*
to appear in éditions Masson, Paris, 1991

[Rumelhart87] D.E. Rumelhart, G.E. Hinton & R.J. Williams
*Learning internal representation by error propagation*
in *Parallel Distributed Processing,* by D.E. Rumelhart, J.L. McClelland & the P.D.P. Research Group; The M.I.T. Press, 1986

[Shore80] J. E. Shore & R. W. Johnson
*Axiomatic derivation of the principle of Maximum Entropy and the Principle of Minimum Cross-Entropy*
IEEE Transactions on Information Theory, 1980

[Shore81] J. E. Shore & R. W. Johnson
*Properties of cross-entropy minimization*
IEEE Transactions on Information Theory, 1981

[Shore86] John E. Shore
*Relative entropy, probabilistic inference and A.I.*
in *Uncertainty in Artificial Intelligence*; edited by L. N. Kanal & J. F. Lemmer; Elsevier Science Publishers, North-Holland, 1986

[Singer90] Alexander Singer
*Exploiting the inherent parallelism of Artificial Neural Networks to achieve 1300 million interconnect per second*
in Procceedings of I.N.N.C.90, Kluwer Academic Publishers, 1990

[Smolensky86] P. Smolensky
*Information processing in dynamical systems: foundations of Harmony theory*
in *Parallel Distributed Processing* by D.E. Rumelhart, J.L. McClelland & the P.D.P. Research Group; The M.I.T. Press, 1986

[Van Campenhout81] J. M. Van Campenhout & T.M. Cover
*Maximum entropy and conditional probability*
IEEE Transactions on Information Theory, 1981

[Zadeh65] L. A. Zadeh
*Fuzzy sets*
Information and Control, N°8, 1965

[Zadeh78] L. A. Zadeh
*Fuzzy sets as a basis for a theory of possibility*
Fuzzy Sets and Systems, N°1, 1978