The Fifteenth International Workshop on
MAXIMUM ENTROPY AND BAYESIAN METHODS
July 31 - August 4, 1995

Rational neural models based on information theory

Robert L. Fry
The Johns Hopkins University/Applied Physics Laboratory Laurel, MD 20723

I. INTRODUCTION AND OVERVIEW

Biological organisms which possess a neurological system exhibit varying degrees of what can be termed rational behavior. One can hypothesize that rational behavior and thought processes in general arise as a consequence of the intrinsic rational nature of the neurological system and its constituent neurons. A similar statement may be made of the immunological system [1]. The concept of rational behavior can be made quantitative. In particular, one possible characterization of rational behavior is as follows

(1) A physical entity (observer) must exist which has the capacity for both measurement and the generation of outputs (participation). Outputs represent decisions on the part of the observer which will be seen to be rational.

(2) The establishment of the quantities measurable by the observer is achieved through learning. Learning characterizes the change in knowledge state of an observer in response to new information and is driven by the directed divergence information measure of Kullback [2].

(3) Output decisions must be made optimally on the basis of noisy and/or missing input data. Optimally here implies that the decision-making process must abide by the standard logical consistency axioms which give rise to probability as the only logically consistent measure of degree of plausible belief. An observer using decision rules based on such is said to be rational.

Information theory can be used to quantify the above leading to computational paradigms with architectures that closely resemble both the single cortical neuron and interconnected planar field of multiple cortical neurons all of which are functionally identical to one another. A working definition of information in a neural context must be agreed upon prior to this development, however. Such a definition can be obtained through the Laws of Form - a mathematics of observation originating with the British mathematician George Spencer-Brown [3].

It is commonly accepted that neurons process information although any precise definition of information within a neural context remains outstanding. Such a definition is fundamental to the analysis and understanding of biological neurons and their ability to process information. The following definition is proposed: *Information is that which is measurable by an observer*. The assumption of this definition means that what is information to one measurement device (observer) is not necessarily information to another physically independent system (observer) as

related to a pervading state of nature. This definition implies that a quantity distinguishable by an observer is information hence information can be seen to be a unique property of the observer. The concept of distinguishability on the part of an observer can be seen to lie at the basis of the Maximum Entropy formulation and can be seen to lie at the foundation of information theory itself.

II. DISTINGUISHABILITY AS INFORMATION

In George Spencer-Brown derived his *Calculus of Indications* as a means of quantifying the concept of "distinguishability" in his *Laws of Form (LoF)* [3]. Indicational expressions within the *Calculus of Indications* provide a quantitative means of describing an observer in terms of the quantities measurable by the observer. In fact, any system can be described using indicational expressions which explicitly delineate system observables and outputs. The most primitive act within the *LoF* is that of distinction which requires both a measurement device and corresponding measurable quantities and is enacted in physics through the Hermitian operator. The lack of any generated indication or observable on the part of an observer means that the observer is a closed physical system unobservable to any other physically independent system/observer and hence is not information as based on the previous definition and therefore is unimportant in our analysis.

The act of distinction on the part of an observer arises from symmetry-breaking on the part of the observer and is the dual-state system as created through learning. The defining distinction and the observer are synonymous. For instance, biological neurons observe spatio-temporal patterns of activity and must decide on which patterns it will fire - the outward manifestation of which is the action potential. Certain pattern classes will more likely give rise to an action potential which represents the *decision* of the neuron. As described here, the observing neuron can probabilistically decide in favor of one distinction over the its contradiction and hence either generate an action potential (indication) or not as based on measured spatio-temporal information. The phenomenology leading to the formation of the defining distinction is deemed to correspond to the process of learning in a neural context. Learning can then be interpreted as the search for the appropriate *question* as opposed to the appropriate *answer*. The use of probability by the neuron to perform its decision-making is what intrinsically makes the neuron rational.

Simple *Calculus of Indications* expressions can be used to describe information flow within a functional neuron q or neural field Q which extract information from the environment through a set of measurement functions $\mathfrak{F}$ which depend logically on the input training ensemble $\mathbf{x}=(x_1\ x_2\ x_3 \ldots x_N)$ and the outputs $\mathbf{y}$ where $\mathbf{x}\varepsilon B^N$, $\mathbf{y}\varepsilon B^M$, $B=\{0,1\}$and in general $M\leq N$. The adoption of $B=\{0,1\}$ is in part arbitrary since the information measures used here are scale-invariant. Therefore, 0 and any numerical value distinguishable from 0 are acceptable although the choice of B lends itself to more concise derivations upon the application of the discussed information-theoretic measures. Boolean expressions arise as a consequence of one limited interpretation of the *LoF*. Below are the basic *LoF* forms for the noted Boolean expressions. The most important is the logical “implies” which mathematically corresponds to physical observation. For instance, the notation $\bar{a}|b$ denotes the distinction of a measured quantity a by the observer b. The situation

$\overline{b}|b = \overline{\ }|$ means that the observer has observed a response matched to its defining distinction b leading to the indication $\overline{\ }|$ which is an action potential in a neural context.

LoF Interpreted as Logic

Literal	Boolean Expression	*LoF*	Delineation
not a	~a	$\overline{a}\|$	~0 = 1 ~1 = 0
a or b	a∨b	ab	0∨0 = 0 0∨1 = 1 1∨0 = 1 1∨1 = 1
a and b	a∧b	$\overline{\overline{a}\|\overline{b}\|}\|$	0∧0 = 0 0∧1 = 0 1∧0 = 0 1∧1 = 1
a implies b (b observes a)	a⊃b	$\overline{a}\|b$	0⊃0 = 1 0⊃1 = 1 1⊃0 = 0 1⊃1 = 1

Neural indicational expressions include feedback of the indicational output **y** to the set of measurement functions internal to Q thereby providing a means of adapting the measurement functions in response to supplied elements $\mathbf{x}\varepsilon B^N$. Equation 1 shows the correspondence between q and the *calculus of indications* expressions for the sampling functions dictated by the set $\mathcal{F}_1=\{x_1y,\ x_2y,...,x_N\ y,y\}$. Literally, this particular set can be described as the case where the observer q observes "the conjunctions x_1y *with*[1] x_2y *with ... with* x_Ny *with* the indication y." Note that these are not just conjunctions of signals in space, but are also conjunctions of the respective signals in time. The *LoF* expression for the observer making measurements through the set $\mathcal{F}_1$ is given in Equation 1.

$$\overline{\overline{\overline{x_1}|\overline{y}|}\ \overline{\overline{x_2}|\overline{y}|}|\cdots\overline{\overline{x_N}|\overline{y}|}\ \overline{y}|}\ |q = y \qquad \textbf{(1)}$$

This form can be mathematically simplified at the expense of loosing insight into the flow of information into and within the model neuron.

[1] Here, *with* means Boolean disjunction "or" and is more accurately used here to convey literal meaning.

III. PROBABILITY AS THE ONLY RATIONAL BASIS FOR DECISION-MAKING

Probability is the only logically consistent measure of degree of plausible belief [4] on the part of an observer. An observer which uses such must be deemed rational. As such, let the joint probability $P(\mathbf{x},y)$ describe the internal statistical model of the environment for a rational neuron q which is developed through learning. Suppose that q uses the measurement set $\mathscr{F}_1$ to extract information from environmental input in order to compute $P(\mathbf{x},y)$. The observer q is assumed to not be capable of maintaining a complete record of input, but rather can only estimate moments estimated through arithmetic averaging of the outputs of the respective measurement functions possessed by the observer.

The statistical model $P(\mathbf{x},y)$ is really a state of knowledge possessed by q. $P(\mathbf{x},y)$ provides q with the capacity of performing rational decision-making regarding the validity of q given the observable information $\mathbf{x}$. It should be noted that $P(\mathbf{x},y)$ reflects knowledge of possible environmental states as possessed (stored) by the subject neuron. The evaluation of $P(\mathbf{x},y)$ can only practically take place in response to a supplied in $\mathbf{x}$ and the resulting self-generated output y at which time an action potential either is or is not generated.

IV. MAXIMUM ENTROPY (ME) FORMULATION

It is assumed that the neural field Q is capable of extracting information from external inputs $\mathbf{x}\varepsilon B^N$ and its own outputs $\mathbf{y}$ through the use of a set the union of the individual neural measurement sets $\mathscr{F}_n$ and is given by $\mathscr{F}=\bigcup_{n=1}^{N}\mathscr{F}_n$ The field Q uses the sampled data to estimate moments on the defined measurement functions $\mathscr{F}$ which it in turn uses to compute the joint distribution $P(\mathbf{x},\mathbf{y})$. The computed moments serve to realize $P(\mathbf{x},\mathbf{y})$ as a unique network ME distribution or equivalently a Gibbs distribution parameterized by synaptic connection weights which are in fact the Lagrange multipliers for the ME distribution as are all the decision thresholds μ_n. For instance, suppose that the measurement functions for a single neuron q consist of $\mathscr{F}_1$ as defined earlier. In this case $P(\mathbf{x},y)=\exp[-\sum\lambda_n x_n y+\mu y]/Z$ where Z is the standard partition function and $\Lambda_1=\{\lambda_1,\lambda_2,...,\lambda_N\ ,\ \mu\}$is the Lagrange set for the corresponding moments on the elements of $\mathscr{F}_1$. In this case, the classical Hopfield neuron arises [5],[6] and the maximum likelihood (ML) decision rule can be used to generate an action potential when $\sum\lambda_n x_n-\mu>0$, otherwise none is generated. If a stochastic decision rule is adopted, then an action potential can be generated according to the sigmoidal transition rule given by $P(y=1|\mathbf{x})=1/1+\exp[-\zeta(\mathbf{x})]$ where $\zeta(\mathbf{x})=\sum\lambda_n x_n-\mu$. A simple calculation will show that ζ is in fact the statistical *evidence* function which is defined to be the log of the odds function given by the likelihood ratio $P(y=1|\mathbf{x})/P(y=0|\mathbf{x})$. Computed statistical evidence is therefore used by the neural observer to perform rational decision-making.

There is a problem with this formulation, however. It is unrealistic to expect that a neuron or collection of neurons has to computationally solve the ME optimization problem to obtain the prerequisite Lagrange set. For one thing, this would imply that the moments of the measurement

set either be specified by a supervisory source or that the moments be extracted and stored as computed from observed data. Secondly, ME optimization does not have a known simple biological plausible implementation which uses locally available information. Consider the following alternative.

It is a well known fact that the Lagrange multipliers and the distributional moments are completely conjugate to one another for an ME formulation with linear constraints. Knowing one set fully constrains the other. Therefore, the specification of the values for a Lagrange set Λ will simultaneously define the complete ME distribution and a corresponding set of moments. The question here then becomes that of how to specify values for Λ or what methodology to use for computing Λ directly from observable data. One possible strategy is to select values for Λ which serve to maximize the mutual information between the input $\mathbf{x}$ to the Q-field and output $\mathbf{y}$ (or in the case of an individual neuron - just y).

V. MAXIMIZED MUTUAL INFORMATION (MMI)

Suppose that the neural field Q rationalizes according to $P(\mathbf{x},\mathbf{y})=\exp[-\boldsymbol{\lambda}\cdot\mathbf{f}]/Z$ where $\mathbf{f}$ is a vectorized representation of the set of measurement functions $\mathscr{F}$ defined architecturally within Q. In this case, one can dynamically compute the Lagrange values $\boldsymbol{\lambda}$ which maximize the mutual information between $\mathbf{x}$ and $\mathbf{y}$ through the use of the Gibbs Mutual Information Theorem (GMIT) as derived in [6] which states that the optimal $\boldsymbol{\lambda}$ is given by the extrema of $J=I(\mathbf{x};\mathbf{y})+\sum\alpha_m g_m(\boldsymbol{\lambda})$ where $I(\mathbf{x},\mathbf{y})$ is the mutual information between $\mathbf{x}$ and $\mathbf{y}$ and each $\alpha_m g_m(\boldsymbol{\lambda})$ represents architectural constraints placed on one or more members of the Lagrange set - realistic biological constraints might include limited neurotransmitter resources, non-negativity of specific synaptic connections, etc. In this case, GMIT states that for all $n=1,2,...,N$ and each $f_n\in\mathscr{F}$.

$$\frac{\partial J}{\partial \lambda_n}=\left\{\mathrm{E}\left[< f_n > - f_n(\mathbf{x},\mathbf{y})\right]\ln\left(\frac{\mathrm{P}(\mathbf{x}|\mathbf{y})}{\mathrm{P}(\mathbf{x})}\right)\right\}+\frac{\partial}{\partial \lambda_n}\sum \alpha_m g_m(\boldsymbol{\lambda})=0$$

As an example, the set $\mathscr{F}_1$ with an ℓ_2-norm constraint on $\boldsymbol{\lambda}$ such that $|\boldsymbol{\lambda}|^2=c$, leads to a Hopfield neuron model [5] which maximizes its output entropy over the ensemble of training patterns. A related results holds for the measurement set $\mathscr{F}$ as defined above. In this case the network can, with some further realizable architectural restrictions, be logically factored into functionally equivalent neural components q_n which correspond to the individual sets $\mathscr{F}_n$, respectively for $n=1,2,...,N$ and $|Q|=N$.

VI. SUMMARY

Together ME and MMI lead to a neural field which is factorable in both form and function into component computational entities which correspond to the Hopfield neuron model including decision threshold, action potential realization, Hebbian learning, sigmoidal transfer characteristic, and conditionalized principal component analysis using a simple modification of an equation originally described by Oja [7]. The resulting physical neuron paradigm can in many regards be considered rational and furthermore has biological plausibility [8],[9].

REFERENCES

[1] Francisco J. Varela, *Principles of Biological Autonomy*, North Holland, 1979.

[2] Solomon Kullback, *Information theory and statistics*, Wiley, 1959 and Dover, 1968.

[3] George Spencer-Brown, *Laws of Form*, E. P. Dutton, New York 1979

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

[5] Hopfield, J. J., "Neural networks and physical systems with emergent collective computational abilities," *Proc. Nat. Acad. Sci. USA*, 79, 2554-2558, 1982.

[6] R. L. Fry, "Observer-participant models of neural processing," *IEEE Trans. Neural Networks*, July, 1995.

[7] E. Oja, "A simplified neuron model as a principal component analyzer," *J. of Math. Biol.* **15**, 267-273., 1982

[8] William B. Levy, Costa M. Colbert, and Nancy L. Desmond, “Elemental adaptive processes of neurons and synapses: A statistical/computational perspective,” in *Neuroscience and connectionists theory*, pp 187-235, 1990.

[9] William B. Levy, “Maximum entropy prediction in neural networks,” in proc. *International Joint Conference on Neural Networks*, 1990.