A Theory of Neural Computation

Robert L. Fry

The Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Road
Laurel, MD 20723-6099
robert_fry@jhuapl.edu

Abstract. A theory of systems is proposed that is complementary to information theory and which provides a means of understanding biological neurons. This theory, its cybernetic nature, and an engineering formalism are summarized and used to derive the architectural elements of a single-neuron system, including its operational characteristics, spatiotemporal adaptation rules, and most importantly, its computational objective. Design assumptions are tailored to obtain an architecture closely resembling that of biological neurons. The underlying theory represents a "meta-theory" in the sense that it captures information theory, systems theory, and hybrid systems such as those described by systems of single-neuron systems that internally both process and transmit information between constituent elemental "systems."

Keywords: Systems theory, information theory, cybernetics, Boolean Algebra, logical questions

INTRODUCTION AND OVERVIEW

Boolean Algebra is almost universally accepted as providing the unique and internally consistent mathematical framework for the manipulation of logical expressions involving Boolean variables and literals as combined using any minimal covering set of logical operators such as {"~", "∨", "•"} or logical complementation, disjunction, and conjunction, respectively. However, one can ask the question "Why do the identities of Boolean Algebra hold and what do they mean?" Boolean Algebra cannot explain itself. At least there is no evidence of this since George Boole's original introduction of his algebraic theory [1]. The work of Cox [2] – [4] provides one explanation and further supports a common theory of systems that communicate information and systems that acquire and exploit it to guide action in regards to some essential purpose.

The essential new idea introduced by Cox [4] is that one can postulate the existence of a logical domain of questions that lies in mutual relation to conventional Boolean Algebra and wherein the rules of Boolean Algebra also apply. Each logical domain is defined by and exists only in mutual relation to the other. Boolean logic has traditionally been viewed as static in nature. The proposition "It is raining outside!" merely represents a logical statement that by the laws of Aristotelian logic can only be true or false. However, viewed dynamically, this is an active *assertion* or proclamation by a physical *assertor* which answers some question. The physical flow of information and the conducting of actions require that every assertion have a corresponding and unique defining question, and vice versa. In this case, the assertion $r \equiv$ "It is raining outside!" implies the existence of a complementary question R that defines r. The usage of the term *complementary* in this context is different from that of simple logical complementation that for example ensures identities like $\sim\sim r = r$. We use the symbol "¬" to denote this new kind of complementation operator and call it *reflection*. In this case, the defining question for the assertion r is given by the reflected question $R = \neg r$. Similarly, $r = \neg R$ defines and makes existent the assertion r. In this case, r and R are an *elementary* assertion and question, respectively. Note that r and R are not variables, but rather subjectively defined literal constants.

In general, questions and assertions are not elementary [5]. A question is more generally defined by any *system of assertions* that answer it. Thus $D \equiv$ "What side of the die will show?" is defined by the system $D \equiv \{d_1, d_2, \ldots, d_6\}$ where the assertions $d_i \equiv$ "The ith face is showing!" are elementary and comprise a system of assertions that define D. Now define the elementary questions $D_i \equiv$ "Is the

*i*th face showing?" Then $D_i = \neg d_i$ follows. Now define the assertion d by $d = \neg D$. Then, if a system has as its *issue* the resolution of the issue D, then the same system holds as its *premise* the assertion d regarding what are possible resolutions to the issue D where $\neg D = d \equiv \{D_1, D_2, \ldots, D_6\}$. The premise d is the disjunction of all d_i while D is the conjunction of all D_i. The use of logical questions enables the manipulation of quantities corresponding to uncertainties and this is of profound theoretic and practical importance.

In a communication system, uncertainties are associated with the questions $S \equiv$ "What code will be selected and sent by the source?" and the question $R \equiv$ "What code will be decoded by the receiver given that the various source codes may have been sent?" The latter question is inclusive of the consideration of the intervening physical communication channel. Alternatively, within a systems context, the questions become $X \equiv$ "What information has been acquired?" and $Y \equiv$ "What action should be selected and conducted given the acquired information X?" The duality between channels and systems implies that the many theorems of information theory become relevant to this systems theory and vice versa [6]. In particular, it holds by way of the analogy between a communication channel that maximizes its source rate and matches it to the channel capacity and neuron system that maximizes its output decision rate and maximizes and matches its acquisition rate to its decision rate.

SINGLE-NEURON SYSTEMS

The present theory provides a simple information-theoretic derivation and system-theoretic explanation of neural computation thereby extending previous work ([7]–[15]). Before providing a mathematical overview, we first provide a qualitative and geometric perspective of the proposed computational paradigm.

One can consider the totality of binary codes received by a single-neuron "system" observed over its lifetime and which drive its operation and adaptation. The totality of presented codes or any conceivable subset can be viewed from the subjective perspective of the system as clouds of points tenuously populating an N-dimensional hypercube where N is the number of binary system inputs. The curse of dimensionality guarantees a sparsely populated space for values of N typical of biological neurons – often $>10^3$ in many cases, over any conceivable operational lifetime even including that of the universe. The adaptation process of the system [6] can be described by the construction the parameters defining an N-dimensional hyperplane having orientation $\boldsymbol{\lambda}$ and offset μ within the high-dimensional input space, where $\boldsymbol{\lambda}$ correspond to the system synaptic efficacies and μ the system decision threshold.

During adaptation, the system navigates along an information manifold defined by its N+1 dimensional parameter space. It does so in such a way as to arrive at local or global adaptation equilibria. At these equilibria, the system anticipates that that in the future that one-half of the codes it will observe will fall on one or another of the constructed hyperplane. Those points observed by the system as falling on some specific side of the prevailing hyperplane will *probably* induce an action potential while those falling on the other side *probably* will not. We will see that it is important to consider the presence of a modulating influence that can regulate these probabilities and the rate at which the system makes errors in its output decisions – firing when it shouldn't and not firing when it should. Once the adaptation process reaches an equilibrium, the system then realizes a condition whereby it has a maximal expected information throughput. This condition is in turn brought about through the maximization of its output decision rate and then matching this rate to its information acquisition rate.

The adaptation process serves to maximize the neuron's ability to acquire and match information derived from its input N+1 dimensional code space and to exploit this information to make decisions via its output code space as accurately and at as high a rate as possible. The adaptation of the parameter μ serves to maximize the decision rate while the adaptation of the parameter vector $\boldsymbol{\lambda}$ modifies what information is extracted by the system from the environment. The vector $\boldsymbol{\lambda}$ is ostensibly the measuring rod of the neuron against which it measures everything it believes can exist "outside." Alternatively, the scalar μ modulates the articulation of decisions made in regards to observed information. The system effectively learns and decides where to place a "stick in the sand"

thereby deciding how it can distinguish and how it can act upon the world it subjectively postulates to exist about it [6].

We now provide a brief mathematical overview of a system-theoretic "design process" for obtaining the architecture of single-neuron system having the design objective A of maximizing its information throughput. One can graphically capture the computational objective of such a system through the use of an information diagram or I-diagram [20] such as shown in Figure I.

Suppose that the objective of the system to reconstruct the input code X given the asserted code Y. Having done so, the information content of Y will have maximal bearing on the answering of the question X and vice versa. That is, answering the question $Y \equiv$ "What code should be asserted?" should ideally answer the questions $X \equiv$ "What code was observed?" Logically, we desire that $Y \rightarrow X$ or that the information content of the output Y contain the uncertainty of what code X was observed. That is, one would like Y to contain the uncertainty of X just as in a communication system one would like to have the question of what code was received R include the uncertainty of the question S regarding what source code was sent, or $R \rightarrow S$. Since this is architecturally impossible for the subject many-to-one single-neuron system, we then desire that the logical condition $Y \rightarrow X$ hold to the *degree* possible where *degree* means that there must exist a quantitative measure of the degree to which the logical condition holds.

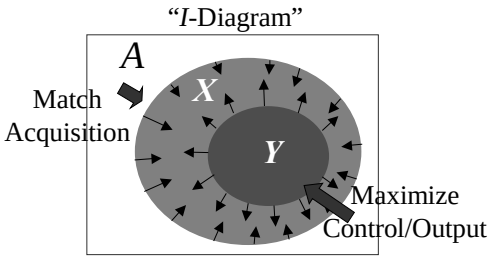

Figure I: This is an I-diagram for a system attempting to match acquired information X to its decision rate Y while maximizing the rate at which it generates accurate decisions Y. Here, the logical condition $X \rightarrow Y$ holds since the region Y is contained within the region X.

Regarding Figure I, the objective A is achievable through two synergistic adaptation processes. The first process attempts to make the area of Y as large as possible in regards to covering A. This in effect corresponds to the maximizing of the decision rate of the system. The second process attempts to make the area X maximally conformal to Y. The arrows shown in the I-diagram in Figure I depict these two adaptation processes that together serve to maximize the *degree* to which $Y \rightarrow X$. In maximizing the area Y, the system is maximizing its decision rate Y relative to the issue A which is the system decision entropy $H(Y)$. Through a conformal mapping of X to Y, the system maximizes the relevancy of the information extracted from X to Y relative to the computational issue A. This is the common information $X \vee Y$ between input X and output Y and corresponds to the mutual information $I(X;Y)$ ([20]). Maximizing $I(X;Y)$ serves to guarantee that the acquired information X is matched to the need for information induced by Y.

Suppose that the single-neuron system possesses N Boolean inputs $X_1, X_2, \ldots, X_N$ with each input X_i representing a question posed by the system to its environment and answered through the existence or nonexistence of externally furnished pre-synaptic potentials. Each X_i is a "binary" question which is a question defined by a system of two complementary assertions. In this case let $X_i \equiv \{x_i, \sim x_i\}$. The defining assertions literally correspond to $x_i \equiv$ "A potential is asserted on the ith synaptic input!" and $\sim x_i \equiv$ "No potential is asserted on the ith synaptic input!" while $X_i \equiv$ "Is there or is there not a potential present?" We have excluded the additional qualifier "at a specific time" on each literal definition to keep the example simple for now. "Time" is explicitly required and is discussed briefly later where this qualifier is needed. The question of what code was observed X is determined by all questions X_i. Specifically, it is the information derived by posing in parallel to the environment all questions X_i, i=1,2,…, N. This mathematically ([3], [4]) means that the system poses the question X corresponding to the conjunction of all X_i or $X = X_1 \bullet X_2 \bullet \ldots \bullet X_N$. The question X is then defined by

the system of 2^N assertions describing every possible code that can in principle be observed by the system. That is, $X \equiv \{\ \sim x_1 \bullet\cdot \sim x_2 \bullet\cdot \ldots \cdot\bullet \sim x_N\ ,\ \ldots\ldots\ ,\ x_1 \bullet\cdot x_2 \bullet\cdot \ldots \cdot\bullet \sim x_N\ \}$. Thus there are 2^N possible defining assertions for the question X. Since neurons can have in excess of 5000 inputs, there can be a unimaginable number of assertions ($>2^{5000}$) in the system that defines X. Any neuron can expect to observe only an infinitesimal fraction of these possible codes over its lifetime. The question X defines the *inquiry space* of the system [6] and represents what it can ask. The inquiry space is the analog of the receiver element in a communication system.

As one might suspect, the system also has an *action space* that is analogous to the source element of a communication system. The action space corresponds to those questions that the system can answer through the selection of specific actions from a suite of possible actions. For the subject single-neuron system, the action space is extremely simple and defined by $Y \equiv \{y, \sim y\}$ where $y \equiv$ "Generate an action potential!" and $\sim y \equiv$ "Do not generate an action potential!" The question Y represents the system's query regarding what to do. The *cybernetic space* of the system is defined by the conjunction of its inquiry and action spaces as elaborated upon in [6] and so $A = X \bullet Y$ which for this system is defined by a system of assertions having exactly twice as many assertions as does the defining system for X since each assertion defining X is conjoined with both y and $\sim y$ in accordance with the definition of the conjunction of questions ([3], [4]).

The final "design step" consists of the specification of the computational resources possessed by the system relative to its ability to extract and record information useful in guiding its two adaptation processes. It is assumed [14] that the single-neuron system cannot record or directly use previously observed input or output codes and that is a discrete memoryless system analogous to a discrete memoryless channel in information theory. However, it is assumed that the system can measure and extract certain things from its input and output codes and that it can retain certain memories of what it has seen in the past and what it did based on what it saw. In particular, we postulate that the system can measure the numerical averages $\langle x_i \bullet y \rangle$ of observed coincidences (conjunctions) of input potentials x_i with its self-generated potential decision events y. We further postulate that the system can measure the fraction of the times that it generated action potentials $\langle y \rangle$ in the past taken over the opportunities that it had to do so. Thus it is postulated that the system is capable of measuring and recording $N+1$ quantities that are somehow related to its $N+1$ adjustable parameters. It is important to note that as the single-neuron system designer, one is free to postulate that the system can measure any function(s) of its inputs and its outputs. Assumptions are tailored here to yield a biologically plausible architecture.

The step of mathematically deriving the single-neuron system architecture requires the application of inductive principles that are both universal and which abide by logical consistency [3]. By universal, it is meant that the principles must apply to all systems. By logical consistency, it is meant that the system must attain a common final state of knowledge regardless of the orders of introduction of information to it as long as the information is provided in its totality. The resulting information processing rules are unique and include Bayes' Theorem and the principal of Maximized Entropy ([18], [19]). Given that the system measures $\langle x_i \bullet y \rangle$ and $\langle y \rangle$, the application of the principal of maximized entropy leads to a precise parametric form for the joint distribution $p(\boldsymbol{x},y|a)$ on input codes $\boldsymbol{x}$ and output codes y where a is the premise capturing all system assumptions made thus far. In particular, the coefficients λ_i are parameters of $p(\boldsymbol{x},y|a)$ and correspond to the synaptic efficacies $\boldsymbol{\lambda}$ and the Lagrange multipliers for the moment constraints $\langle x_i \bullet y \rangle$. Likewise, μ is another parameter of $p(\boldsymbol{x},y|a)$ which also serves as the system decision threshold and the Lagrange multiplier for the moment constraint $\langle y \rangle$. The distribution $p(\boldsymbol{x},y|a) = 1/Z \exp[-\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{x}y + \mu y]$ represents the model constructed by the system of its external world regarding what it will see and what it should do given what it has seen.

Again. the computational issue A of the system is to maximize its information throughput as realized through the dual adaptation processes prescribed by the *I*-diagram in Figure I. We thus desire that the decision entropy rate $H(Y)$ be maximized and that the acquisition rate $I(X;Y)$ (mutual information or the information shared between X and Y) be matched to this decision rate. Since the maximal amount of information that X can have in common with Y is the decision entropy rate $H(Y)$ (since $H(Y) = I(X;Y) + H(Y|X)$, it follows that $H(Y) \geq I(X;Y)$ since all measures are positive), then

mutual information between the acquired information and possible controls can be independently maximized subject to any additionally imposed system constraints. The mutual information $I(X;Y)$ specifies the system information acquisition rate of the system viz. the transmission rate of a channel.

Maximizing the output entropy [14] leads to the equilibrium condition for the decision threshold μ that $\mu = \left\langle \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{x} \,\middle|\, y = 1 \right\rangle$. The quantity $\zeta(\boldsymbol{x}) = \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{x} - \mu$ is important to system operation and adaptation and represents the effective induced somatic potential useful to the system used as a basis for the decision of whether or not to fire given observed codes $\boldsymbol{x}$. $\zeta(\boldsymbol{x})$ is formally the statistical evidence for making the decision to fire or not. In particular, it directly parameterizes the conditional probability of firing $p(y = 1|a \bullet \boldsymbol{x})$ given in Equation 1 and which is directly obtainable from $p(\boldsymbol{x},y|a)$:

$$p(y = 1 | \, a \bullet \boldsymbol{x}) = \frac{1}{1 + \exp[-\zeta(\boldsymbol{x})]} . \tag{1}$$

Therefore, the evidence $\zeta(\boldsymbol{x})$ is derived solely as a measurement statistic on its input code space X and $p(y = 1|a \bullet \boldsymbol{x})$ is a sigmoidal function of the rendered evidence. Equation 1 describes the operation of the system while $\mu = \left\langle \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{x} \,\middle|\, y = 1 \right\rangle$. describes how the system should adapt its decision threshold indicating that μ should evolve towards the average value of the induced potential $\boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{x}$ induced by input codes and as conditioned on the fact that the input $\boldsymbol{x}$ generates an action potential $y = 1$. The condition $y = 1$ is of course Hebb's rule [16] and arises as a natural consequence of the form of the distribution $p(\boldsymbol{x},y|a)$. Achieving this equilibrium guarantees the equalities $p(y|a) = 1 - p(\sim y|a) = ½$ and therefore $H(Y) = 1$ bit as borne out in simulations [14]. The actual algorithmic implementation used to achieve $\mu = \left\langle \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{x} \,\middle|\, y = 1 \right\rangle$. is rather unimportant and can for instance take the form of a simple windowed average for $\theta \in (0,1)$ of the form

$$\mu(t + \Delta t) = (1 - \theta)\mu(t) + \theta \boldsymbol{\lambda}^{T} \boldsymbol{x}(t) \tag{2}$$

which is executed by the system only under the Hebbian condition $y = 1$.

Determination of the optimal vector $\boldsymbol{\lambda}$ requires maximizing the mutual information between the system input X and output Y over $\boldsymbol{\lambda}$ [9]. A regularization constraint is introduced whereby $|\boldsymbol{\lambda}|^2 \le \gamma^2$. Architecturally, this says that the system has a limited resource that can be distributed across its synaptic inputs, and that this resource is judiciously used to selectively amplify pre-synaptic signals rendered to the system by the environment through the synaptic efficacies.

The resulting equilibrium adaptation condition that serves to maximize the acquisition rate and match it to its transmission rate dictates that the optimal $\boldsymbol{\lambda}$ should be a vector aligned with the largest eigenvector [14] of the covariance matrix $\mathbf{R}$ defined by

$$\mathbf{R} = \left\langle \left[\boldsymbol{x} - \left\langle \boldsymbol{x} \middle| y = 1 \right\rangle \right] \left[\boldsymbol{x} - \left\langle \boldsymbol{x} \middle| y = 1 \right\rangle \right]^{T} \;\middle|\; y = 1 \right\rangle \quad . \tag{3}$$

The matrix $\mathbf{R}$ is the conditional variation of the observed data about the conditional mean of the observed measurements $\boldsymbol{x}$ where the conditioning in each case is on the input code $\boldsymbol{x}$ generating an action potential $y = 1$ which again is Hebb's condition.

Regarding matched acquisition, a very efficient scalar and sequential computational algorithm already exists for dynamically finding the largest eigenvector of a matrix $\mathbf{R}$. In particular, Oja [17] proposed precisely the form of the algorithm needed. Oja's algorithm can be tailored to our needs and then executed under Hebbian gating constraint. In particular, Oja's algorithm becomes [9]

$$\boldsymbol{\lambda}(t + \Delta t) = \boldsymbol{\lambda}(t) + \pi \zeta(\boldsymbol{x})[\boldsymbol{x}(t) - \gamma^{-2} \zeta(\boldsymbol{x}) \boldsymbol{\lambda}(t)] \tag{4}$$

where γ^2 is the l^2 -norm constraint on $|\boldsymbol{\lambda}|^2$. This is a very simple adaptation equation and makes common use of the somatic evidence $\zeta(\boldsymbol{x})$ in conjunction with the adaptation algorithm for μ in Equation 2. The evidence $\zeta(\boldsymbol{x})$ has common use for the adaptation of inputs and output as well as system operation.

The adaptation algorithms given by Equations 2 and 4 have been tested [14] over a large variety of conditions and were found to consistently demonstrate learning equilibria that served to maximize the information throughput of the single-neuron systems. As an example, if input codes $\boldsymbol{x}$ are uniformly drawn from a larger training set, the model neuron consistently partitions the training set into two groups of equal size against which the neuron operationally fires on half the training set. This in effect ensures a decision rate of 1 bit. The adaptation process of the synaptic efficacies (Equation 4) creates a competition between synapses for a fixed resource that bounds achievable synaptic efficacies. Therefore, the strength of connections λ_i tied to a synaptic inputs never or rarely excited by a pre-synaptic potentials are actively attenuated as opposed to undergoing passive atrophy.

The temporal operation and adaptation of the single-neuron system can be derived in a like manner. Operationally, the system fires probabilistically according to $p(y=1|a\bullet x)$ as defined in Equation 1 thereby abiding by the measured evidence $\zeta(\boldsymbol{x})$. A very simple way to accomplish this is to provide additive non-specific neural inputs that do not convey information regarding Y, but rather only serve to modulate somatic noise levels [7] and the system error rate. Temporal adaptation is also essential to system operation. Pre-synaptic potentials must arrive and be integrated at a common location where decisions are made – a somatic decision point. Their influences must be coherently integrated to compute the evidence $\zeta(\boldsymbol{x})$ for firing at various instants in time. One can surmise this is done within the soma of biological neurons. Those input potentials having relevance to Y should synchronously arrive at the decision-generating location of the system. Ideally, then, the system should possess the capacity to modulate the transmission delays from the relevant pre-synaptic input sites to the system decision point. That is, the system should pursue a distributed temporal adaptation strategy whereby codes X_i having bearing on the question Y can be equalized regarding their arrival times at the decision point. It can easily be shown [14] that a simple and local Hebbian delay equalization algorithm corresponds to $\mathrm{d}\tau_i / \mathrm{d}t = -\lambda_i\, y(t)\, \mathrm{d}x_i(t-\tau_i)/\mathrm{d}t$. This expression contains a momentum-like factor having velocity $\mathrm{d}x_i/\mathrm{d}t$ and particle mass λ_i. This algorithm is also Hebbian and is cooperative with the adaptation of $\boldsymbol{\lambda}$ in that larger connection strengths force a more rapid equalization of those dendritic delays lying in correspondence with the larger synaptic efficacies. Thus, more relevant inputs have their delay times equalized at a faster rate. Including μ, each single-neuron system now has a total of $2N+1$ adaptive parameters that provide for its spatiotemporal adaptation using simple, biologically plausible, and sequential learning algorithms in its pursuit to optimize its system information throughput.

REFERENCES

1. G. Boole, Investigation of the Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities, London, 1854.
2. R. T. Cox, “Probability, Frequency, and Reasonable Expectation,” in Am. J. Physics **14**, 1-13, 1946.
3. R. T. Cox, Algebra of Probable Inference, Baltimore: Hopkins Press, 1961.
4. R. T., Cox, “Of Inference and Inquiry, An Essay in Inductive Logic” in Proc of the Maximum Entropy Formalism –1978, first maximized entropy workshop held at MIT, edited by D. Levine and M. Tribus, MIT Press, 1979, pp. 119-168.
5. R. L. Fry, Course Notes for “Maximum Entropy and Bayesian Methods” Johns Hopkins University, available from the author.
6. R. L. Fry, “The Engineering of Cybernetic Systems” in Proceedings of the 21st Workshop on Maximum Entropy and Bayesian Methods, New York: AIP 2002 (in press).
7. R. L. Fry, “Cybernetic Aspects of Neural Computation,” poster presentation at the Sixth International Conference on Cognitive and Neural Systems, Center for Adaptive Systems and Department of Cognitive and Neural Systems, Boston University, 2001 (available from the author).
8. R. L. Fry, “Neural processing of information,” in Proc. IEEE Int. Symp. on Info. Th, Trondheim, Norway (1994).
9. R. L. Fry, “Observer-Participant Models Of Neural Processing,” in IEEE Trans. Neural Networks **6**, 918-928, (1995).
10. R. L. Fry, “Maximized Mutual Information Using Macrocanonical Probability Distributions,” in Proc. IEEE/IMS Workshop on Information Theory and Statistics, Arlington, VA. (1994).

11. R. L. Fry, “Rational Neural Models Based On Information Theory,” Proc. Workshop On Maximum Entropy and Bayesian Methods, Sante Fe, NM, 1995.
12. R. L. Fry, “Rational Neural Models Based On Information Theory,” in Proc. Neural Information Processing Systems - Natural And Synthetic, Denver, CO, 1995.
13. R. L. Fry, “Neural Mechanics,” Proc. Int. Conf. Neur. Info. Proc. (ICONIP), Hong Kong, 1996.
14. R. L. Fry, and Sova, R. M., “A Logical Basis For Neural Network Design” in Techniques and Applications of Artificial Neural Networks **3**, Academic Press, 1998.
15. R. L. Fry, “Cybernetic Systems based on Inductive Logic” in Proceedings of the 20th Workshop on Maximum Entropy and Bayesian Methods, New York: AIP 2001
16. D. O. Hebb, Organization of Behavior, New York: Wiley, 1949.
17. E. Oja, “A Simplified Neuron Model as a Principal Component Analyzer,” in J. of Math. Biol. 15, 267-273 (1982).
18. J.E Shore and R.W. Johnson, “Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum-Cross Entropy,” in IEEE Transactions on Information Theory **26**, 26-37. (1980).
19. J. Skilling, "The Axioms of Maximum Entropy," in Maximum Entropy and Bayesian Methods in Science and Engineering, edited by G. J. Erickson and C. R. Smith, Dordrecht: Kluwer, 1988.
20. R. W. Yeung, “A New Outlook On Shannon’s Information Measures,” in IEEE Trans. Info. Th. **37**, 466-475 (1991).