Golden, R. M. (1996). Interpreting objective function minimization as intelligent inference. In J. Albus, A. Meystel, and R. Quintero (eds.) *Intelligent Systems: A Semiotic Perspective, Proceedings of An International Multidisciplinary Conference Volume I: Theoretical Semiotics.* U.S. Government Printing Office, Washington, D.C. [5 pages].

# Classical Methods for Interpreting Objective Function Minimization as Intelligent Inference

**Richard M. Golden**
Cognition & Neuroscience Program, GR41
School of Human Development
University of Texas at Dallas
Richardson, TX 75083-0688
golden@utdallas.edu

### Abstract

Most recognition algorithms and neural networks can be formally viewed as seeking a minimum value of an appropriate objective function during either classification or learning phases. The goal of this paper is to argue that in order to show a recognition algorithm is making intelligent inferences, it is not sufficient to show that that the recognition algorithm is computing (or trying to compute) the global minimum of some objective function. One must explicitly define a "relational system" for the recognition algorithm or neural network which identifies the: (i) sample space, (ii) the relevant sigma-field of events generated by the sample space, and (iii) the "relation" for that relational system. Only when such a "relational system" is properly defined, is it possible to formally establish the sense in which computing the global minimum of an objective function is an intelligent inference.

## OPTIMIZATION ALGORITHMS

In this discussion, algorithms will be defined from a dynamical systems perspective following the approach of Golden (forthcoming, Chapter 1). Let $S$ be a vector space and let $T$ be the set of non-negative integers. An *algorithm* will be defined in this paper as a function that maps an *initial state* $x \in S$, an *initial time* $t_0 \in T$, a *sequence of external inputs* $u \in U$, and a *final time* $t_f \in T$ into some *final state* in $x \in S$. Thus, an algorithm is a function $\Psi : S \times T \times T \times U \rightarrow S$. Some representative algorithms that can be viewed as minimizing an objective function are now presented.

### Fuzzy Logic and Boolean Logic Algorithms

Let $U$ be a set of assertions. For example, the assertion $u =$ *Linear algebra is fun* may be an element of $U$. Let $S = [0,1]$ be a set of *fuzzy truth values*.The algorithm $\Psi : [0,1] \times T \times T \times U \rightarrow [0,1]$ is a *Fuzzy logic algorithm* which maps assertions into *fuzzy truth values* such as *false* (0), *almost false* (0.3), *almost true* (0.7), *very very very true* (0.95), etc.

A function $\Psi$ designed to model the assignment of truth to assertions from the perspective of a student receiving the grade of "C" in a linear algebra class might have the property that: For all $x \in [0,1]$ and for all $t_0 \in T$, $\Psi(x, t_0, t_f, u) \rightarrow 0.3$ as $t_f \rightarrow \infty$ (which attempts to capture the idea that the assertion *linear algebra is fun* should be assigned the fuzzy truth value of *almost false*(0.3)). Another student who is receiving the grade of "A" in a linear algebra class might have the property that: For all $x \in [0,1]$ and for all $t_0 \in T$, $\Psi(x, t_0, t_f, u) \rightarrow 0.7$ as $t_f \rightarrow \infty$ (which attempts to capture the idea that the assertion *linear algebra is fun* should be assigned the fuzzy truth value of *almost true*(0.7)). The initial state $x$ does not influence the classification decision in this case because the assumption that a previous classification influences the current classification is not made. The initial time $t_0$ does not influence the classification decision because it is assumed that the classification strategy does not vary as a function of the initial stimulus presentation time.

A *Boolean logic algorithm* is a special case of a fuzzy logic algorithm which has the form $\Psi : \{0,1\} \times T \times T \times U \rightarrow \{0,1\}$ since the system state $x$ is restricted to the values of either *false* (0) or *true* (1).

One can view a fuzzy logic algorithm (and thus a Boolean logic algorithm as well) as seeking the minimum value of a particular objective function for classification. Let $V : [0,1] \times U \rightarrow \mathcal{R}$. The *classification objective function* for a fuzzy logic algorithm $\Psi$ given some assertion $u \in U$ is a function $V(\cdot; u) : [0,1] \rightarrow \mathcal{R}$. The computational goal of the fuzzy logic algorithm is to find a global minimum of $V(\cdot; u)$ on $[0,1]$ for a given $u \in U$. For example, suppose one wishes to set up the classification objective function so that the fuzzy truth value of $p_s \in [0,1]$ is assigned to the assertion $s \in U$. One possible choice of $V(\cdot; s)$ is to choose $V(p; s) = 1$ if $p \neq p_s$ and $V(p; s) = 0$ if $p = p_s$.

### Recurrent Artificial Neural Networks

The Cohen-Grossberg class of continuous-time networks include the continuous-time versions of the Hopfield (1982, 1984) and Anderson's BSB model (Anderson, Silverstein, Ritz, and Jones, 1977; Golden, 1986) as important special cases. Let $x_i$ be the activation of the $i$th unit in a $d$ neuron system. Let $\mathbf{W}$ be a $d$-dimensional matrix of symmetric connections among the $d$-units. The equation for one version of the Cohen-Grossberg network is:

$$dx_i/dt = z_i(x_i)[b_i(x_i) - \sum_{k=1}^{d} w_{ik} \mathcal{S}_k(x_k)]$$

where $x_i$ is the activation level of the $i$th neuron in the $d$-neuron system, $z_i$ is an arbitrary function of $x_i$ such that $z_i(x_i) > 0$ for all $x_i \in \mathcal{R}$. Let $F_{min}, F_{max} \in \mathcal{R}$ such that $F_{min} < F_{max}$. The sigmoidal function $\mathcal{S}_k$ :

$\mathcal{R} \rightarrow [F_{min}, F_{max}]$ is a continuous differentiable monotonically increasing function. The function $b_i : \mathcal{R} \rightarrow \mathcal{R}$ is an arbitrary continuous function. The real numbers $w_{ik} = w_{ki}$ for $i$ and $k$ can be represented as the real symmetric matrix $\mathbf{W} \in \mathcal{R}^{d \times d}$. Let $\mathbf{x} = [x_1, \ldots, x_d]$. Let $V : [F_{min}, F_{max}]^d \rightarrow \mathcal{R}$ be defined such that for all $\mathbf{x} \in [F_{min}, F_{max}]^d$:

$$V(\mathbf{x}) = -\sum_{i=1}^{d} \int^{x_i} b_i(u_i)\mathcal{S}_i'(u_i)du_i +$$

$$(1/2)\sum_{j=1}^{d}\sum_{k=1}^{d} w_{jk}\mathcal{S}_j(x_j)\mathcal{S}_k(x_k).$$

The notation $\mathcal{S}_i'(u_i)$ indicates the derivative of $\mathcal{S}_i$ evaluated at $u_i$.

The solution to the Cohen-Grossberg differential equation may be expressed as an algorithm of the form:

$$\mathbf{\Psi} : [F_{min}, F_{max}]^d \times T \times T \times \mathcal{R}^{d \times d} \rightarrow [F_{min}, F_{max}]^d.$$

It can be shown that under certain conditions, a Cohen-Grossberg algorithm

$$\mathbf{x}(t) = [x_1(t), \ldots, x_d(t)] = \mathbf{\Psi}(\mathbf{x}(t_0), t_0, t_f, \mathbf{W})$$

has the property that for all $t_0$ and for all $\mathbf{x}(t_0)$ near a strict local minimum $\mathbf{x}^*$ of $V$ that $\mathbf{x}(t_f) \rightarrow \mathbf{x}^*$ as $t_f \rightarrow \infty$.

### A General Class of Learning Algorithms

Consider the large class of classification algorithms that map an input vector $\mathbf{s} \in \mathcal{R}^m$ and a parameter vector $\mathbf{w} \in \mathcal{R}^q$ into a response vector $\mathbf{r} \in \mathcal{R}^k$. This class of classification algorithm includes as important special cases: linear regression and logistic regression algorithms, multilayer backpropagation learning algorithms (Rumelhart, Hinton, and Williams, 1986), and the classical Widrow-Hoff (1960) and perceptron (Rosenblatt, 1962) learning algorithms.

A member of this class of architectures may be denoted by a differentiable function $\mathbf{f} : \mathcal{R}^m \times \mathcal{R}^q \rightarrow \mathcal{R}^k$. Let $u_n = (\mathbf{s}^1, \mathbf{o}^1), \ldots (\mathbf{s}^n, \mathbf{o}^n)$ be a set of $n$ input/output pairs (i.e., the *training data*) where the $m$-dimensional real vector $\mathbf{s}^j$ is the $j$th *input vector* and the desired response of the classification algorithm to $\mathbf{s}^j$ is the $k$-dimensional real *target vector* $\mathbf{o}^j$ $(j = 1 \ldots n)$.

A learning algorithm $\mathbf{\Psi} : \mathcal{R}^q \times T \times T \times U \rightarrow \mathcal{R}^q$ maps an initial parameter vector $\mathbf{w}_0 \in \mathcal{R}^q$, an initial time $t_0 \in T$, a final time $t_f \in T$, and a set of training data $u_n \in U$ into a final parameter vector $\mathbf{w}^* \in \mathcal{R}^q$. The learning algorithm is designed (to the greatest extent possible) to compute a $\mathbf{w}^*$ which is a global minimum of some learning objective function $l(\cdot; u_n) : \mathcal{R}^q \rightarrow \mathcal{R}$. With respect to the statistical pattern recognition literature, the learning objective function is typically an expected risk measure.

## INTELLIGENT INFERENCE

In the previous section, a large class of representative algorithms were introduced within a common theoretical framework. It was shown that all of these algorithms may be viewed as seeking the minimum of some objective function. In some cases (such as the fuzzy logic algorithm), the sense in which a given algorithm is making an "intelligent inference" is reasonably clear, while such a notation is less clear for other algorithms (such as the Cohen-Grossberg artificial neural network). Moreover, even when it is clear that one algorithm may be a statistical pattern recognition algorithm while another algorithm is a Boolean logic algorithm, how can one develop a "science of intelligence" which shows explicitly how the intelligent inferences of a Boolean logic algorithm are related to the intelligent inferences of a linear regression algorithm?

These issues will be addressed in this section in the following manner. First, the general concept of a *relational system with a measure* is introduced as a scheme for representing preferences among sets of inferences. The concept of a relational system with a measure provides a unified framework for defining "systems of intelligence" for any algorithm whose computational goal involves minimizing some objective function.

### Relational Systems with Measures

The discussion in this section will follow the approach of Golden (forthcoming, Chapters 6 and 7) relatively closely. Note that the concept of an objective function is defined in terms of classical (crisp) set theory concepts, and so it will be convenient to continue with the classical approach and define relational systems in terms of crisp as opposed to fuzzy sets. Unless otherwise explicitly stated in the following discussion, all sets are assumed to be classical (crisp) sets.

**Inferences, events, and sigma-fields.** Let $S$ be a set of "potential inferences". Let $\mathcal{F}$ be the sigma-field generated by $S$. For example, if $S$ is a finite set, then $\mathcal{F}$ is the set of all subsets of $S$. An element of $\mathcal{F}$ (i.e., a subset of $S$) is called an *event*. Thus, an event is a set of possible inferences.

Consider the following example which involves a finite set $S$ which contains exactly two inferences. Let $s_1$ denote the inference

$s_1 =$ { *INFER : Ralph wins lottery on sunny day* },

and let $s_2$ denote the inference

$s_2 =$ { *INFER : Ralph wins lottery on windy day* }.

Define a sample space $S$ such that $S = \{s_1, s_2\}$, and so $S$ consists of two elements. The sample space $S$ generates the sigma-field $\mathcal{F}$ which is the set of all subsets of $S$ and which is given by the formula:

$$\mathcal{F} = \{\{\}, \{s_1\}, \{s_2\}, \{s_1, s_2\}\}.$$

Thus, the event $\{\} \in \mathcal{F}$ is a theoretical model of the set of inferences where one infers Ralph doesn't win the lottery on a sunny day and one infers Ralph doesn't win the lottery on a windy day. Another example of an event is the set of inferences where one infers Ralph wins the lottery on a sunny day. This event is $\{s_1\}$. The event $\{s_1, s_2\}$ corresponds to the set of inferences where one infers Ralph wins the lottery on *either* a sunny day *or* a windy day.

**Relational systems.** A relational system will now be defined with respect to the sigma-field $\mathcal{F}$ and the sample space $S$ which generated $\mathcal{F}$. The relational system defines a set of *inference preferences* for some intelligent agent $A$. A *relation* is a (crisp) set, $\omega$, of ordered pairs $\{(E_1, E_2)\}$ where both members of each ordered pair in the relation are elements of the sigma-field $\mathcal{F}$ generated by $S$. Although the semantic interpretation of the relation $\omega$ is essentially arbitrary, for the purposes of this paper if $(E_1, E_2) \in \omega$ then this means that the inferences specified by event $E_1$ are *more appropriate* than the inferences specified by event $E_2$. A *relational system* is a triplet $(S, \mathcal{F}, \omega)$.

For example, as before, let $S = \{s_1, s_2\}$ and let $\mathcal{F}$ be the sigma-field generated by $S$. Suppose that agent $A$ believes that the inference that Ralph will win the lottery on a sunny day is more appropriate than the inference that Ralph will win the lottery on a rainy day. Also assume agent $A$ believes that the inference Ralph will win the lottery on a rainy day is more appropriate than the inference Ralph will not win the lottery at all. Let the notation $(\{s_1\}, \{s_2\})$ indicate that agent $A$ believes that the event $\{s_1\}$ is more appropriate than the event $\{s_2\}$. A theoretical model of the beliefs of agent $A$ may be represented in terms of the relation: $\omega = \{(\{s_1\}, \{s_2\}), (\{s_1\}, \{\})\}$.

**Relational system with a measure.** A *relational system with a measure* $\mathcal{P} : \mathcal{F} \rightarrow \mathcal{R}$ is a relational system $(S, \mathcal{F}, \omega)$ with the property that for all $\mathbf{a}, \mathbf{b} \in \mathcal{F}$: $(\mathbf{a}, \mathbf{b}) \in \omega$ if and only if $\mathcal{P}(\mathbf{a}) \leq \mathcal{P}(\mathbf{b})$. It is important to note that given a relational system $(S, \mathcal{F}, \omega)$, it may be impossible to construct (or find) a measure for $(S, \mathcal{F}, \omega)$.

For example, as before, let $S = \{s_1, s_2\}$ and let $\mathcal{F}$ be the sigma-field generated by $S$. Let $\omega = \{(\{s_1\}, \{s_2\}), (\{s_1\}, \{\})\}$. Let $\mathcal{P}$ be defined such that: $\mathcal{P}(\{\}) = 100$, $\mathcal{P}(\{s_1\}) = 10$, $\mathcal{P}(\{s_2\}) = 20$, and $\mathcal{P}(\{s_1, s_2\}) = 0$. The function $\mathcal{P} : \mathcal{F} \rightarrow \mathcal{R}$ is thus one possible measure for the relational system $(S, \mathcal{F}, \omega)$.

On the other hand, assume

$$\omega^* = \{(\{s_1\}, \{s_2\}), (\{s_2\}, \{\})\}.$$

A measure for the relational system $(S, \mathcal{F}, \omega^*)$ can not be constructed. To see this, assume such a measure could be constructed. Since the function $\mathcal{P} : \mathcal{F} \rightarrow \mathcal{R}$ must satisfy: $\mathcal{P}(\{s_1\}) \leq \mathcal{P}(\{s_2\})$ and $\mathcal{P}(\{s_2\}) \leq \mathcal{P}(\{\})$ which implies: $\mathcal{P}(\{s_1\}) \leq \mathcal{P}(\{\})$. But if $\mathcal{P}(\{s_1\}) \leq \mathcal{P}(\{\})$, then $(\{s_1\}, \{\})$ must be a member of $\omega^*$ which contradicts the original definition of $\omega^*$.

**Properties of relational systems with measures.** As previously noted, relational systems which do happen to possess measures have some special properties. In this section, those special properties will be discussed.

Suppose that an assumption has been made that given any two events (i.e., sets of inferences) $E_1, E_2 \in \mathcal{F}$, an agent $A$ is able to decide whether or not the ordered pair $(E_1, E_2) \in \omega$. The agent $A$'s relational system $(S, \mathcal{F}, \omega)$ is then said to be *connected*.

Now consider any three events $E_1, E_2, E_3 \in \mathcal{F}$. In addition, suppose one assumes that if agent $A$ decides that $E_2$ is more appropriate than $E_1$ (i.e., $(E_2, E_1) \in \omega$) and agent $A$ decides that $E_3$ is more appropriate than $E_2$ (i.e., $(E_3, E_2) \in \omega$). If agent $A$ obeys the *transitivity axiom* of rational decision making, then agent $A$ should also decide that $E_3$ is more appropriate than $E_1$ (i.e., $(E_3, E_1) \in \omega$). A relation which satisfies the transitivity axion is said to be *transitive*. Thus, the relation $\omega = \{(E_2, E_1), (E_3, E_2)\}$ is not transitive but the relation $\omega = \{(E_2, E_1), (E_3, E_2), (E_3, E_1)\}$ is transitive.

From an engineering viewpoint, the identification of such axioms of rational decision making is highly desirable since critical constraints upon the computational goals of the decision making process are expressed by such axioms. On the other hand, considerable research in the field of human rational decision making (Tversky, 1969; Kahneman and Tversky, 1979; Wason, 1966; Johnson-Laird, Legrenzi and Legrenzi, 1972) has shown that it is possible to find situations where the systems of preference relationships used by humans are not consistent with the transitivity axiom as well as other classical axioms of rational decision making and logic.

Given these experimental findings, the viewpoint that an intelligent algorithm's relational system should be constrained to be rational may seem misleading to neuroscientists and psychologists. It is important to realize, however, that many intelligent algorithms never achieve their idealized rational computational goals due to their limited computational resources. Thus, intelligent systems which have been designed from a rational inference making perspective could still exhibit the classical violations of logic and transitivity observed in human subjects due to intrinsic computational limitations. Simon (1969) has proposed this explanation of irrationality in human performance.

**Probabilistic measures for relational systems.** A very important measure for a relational system is the *probabilistic measure*. Probabilistic measures have a different type of intelligence relative to ordinary measures. For example, Cox (1946; see Golden, forthcoming, Chapter 6, for a review) showed that a calculus of belief based upon the probability theory satisfies axioms of rational decision making such as: (i) consistency with the deductive logic (i.e., Boolean algebra), (ii) the degree to which $x$ is an appropriate inference depends upon the degree to which $x$ is not an appropriate inference, and (iii) the degree to which $x$ is an appropriate inference given inference $y$ depends upon the degree to which the conjunction of $x$ and $y$ is an appropriate inference and the degree to which $y$ is an appropriate inference.

## Constructing an Algorithm's Relational System.

The concept of an algorithm that makes intelligent inferences is now introduced. Suppose that some algorithm $\mathbf{\Psi} : S \times T \times T \times U \rightarrow S$ can be shown (in some sense) to seek a global minimum of the objective function $V : S \rightarrow \mathcal{R}$. The first part of this paper discussed a large class of algorithms with this property. It does not make sense to attribute "intelligence" to this optimization algorithm since $\mathbf{\Psi}$ is indeed merely an optimization algorithm.

On the other hand, suppose that the designer of the algorithm $\mathbf{\Psi}$ which seeks a global minimum of $V : S \rightarrow \mathcal{R}$ makes a *theoretical commitment* and defines some relational system $(S, \mathcal{F}, \omega)$ with a measure $\mathcal{P} : \mathcal{F} \rightarrow \mathcal{R}$ such that $\mathbf{\Psi}$ is seeking to minimize $\mathcal{P}$. Such a construction is always possible (for example) by defining $\mathcal{P}(\{x\}) = V(\mathbf{x})$ for all $\mathbf{x} \in S$ and then introducing any additional constraints on $\mathcal{P}$ such that $\mathcal{P}$ is completely defined on $\mathcal{F}$.

### Intelligent Inference Algorithms

If an algorithm can be shown to be seeking a global minimum of some measure $\mathcal{P} : \mathcal{F} \rightarrow \mathcal{R}$ with respect to some relational system $(S, \mathcal{F}, \omega)$, then that algorithm is seeking an inference which is more appropriate (in the sense defined by $\omega$) than any other inference. If the algorithm successfully converges to a global minimum of $\mathcal{P}$, then that algorithm is making intelligent inferences of a particular type.

For example, certain types of behavioral characteristics of birds may be viewed as intelligent within this theoretical framework. A bird which flies into a glass window might still be considered to be intelligent if it has selected the most appropriate inference (*infer no obstruction ahead*) with respect to it's "bird relational system". An outside agent using "human intelligence" might argue that the bird made an inappropriate inference but this argument is from the perspective of a "human relational system". Birds have adapted over time to be intelligent with respect to specific ecological niches.

Thus, a key feature of this definition of an "intelligent inference" is that there are many different types of intelligence. A taxonomy of these different systems of intelligence can be developed by exploring similarities and differences among classes of relational systems. The concept of an intelligent inference algorithm is now provided.

**DEFINITION: Intelligent inference algorithm.** Let $(S, \mathcal{F}, \omega)$ be a relational system with a measure $\mathcal{P} : \mathcal{F} \rightarrow \mathcal{R}$. Assume there exists an $s^* \in S$ such that $\{s^*\}$ is a global minimum of $\mathcal{P}$. Let $\mathbf{\Psi} : S \times T \times T \times U \rightarrow S$ be an algorithm with the property that for all $s \in I \subseteq S$, for all $t_0 \in T$, and for all $u \in R \subseteq U$: $\mathbf{\Psi}(s, t_0, t_f, u) \rightarrow s^*$ as $t_f \rightarrow \infty$. The algorithm $\mathbf{\Psi}$ is an *intelligent inference algorithm* with *$\omega$-type intelligence* with respect to $(S, \mathcal{F}, \omega)$, $\mathcal{P}$, $I$, and $R$.

The concept of an intelligent inference algorithm has thus been explicitly defined. It is important to note, however, that there are a variety of ways in which a given intelligent inference algorithm can fail to be intelligent. First, the initial conditions (i.e., the "retrieval cues") of the intelligent inference algorithm may not be sufficient to guarantee that the algorithm will converge to a global minimum of $\mathcal{P}$. Second, the numerical algorithm designed to compute the optimal inference may be suboptimal in some sense. Thus, the intelligent inference algorithm could fail to consistently make optimal inferences due to algorithm failure. Third, the relational system underlying the algorithm's intelligent behavior may not be appropriate to the information processing task. Or, in other words, failures in intelligent inference could result from flaws in the algorithm's knowledge of the problem. These ideas are discussed briefly in the next section within the context of a practical learning problem.

## AN APPLICATION OF THE THEORY

Consider the problem of estimating the parameters of a linear regression model which is defined by the formula:

$$\tilde{o} = ms + b + \tilde{n}$$

where the parameters $m$ and $b$ are real numbers, $s \in \mathcal{R}$ is a particular value of the predictor variable, $\tilde{n}$ is a Gaussian random variable with zero mean and variance parameter $\sigma^2$, and $\tilde{o}$ is a Gaussian random variable with mean $ms + b$ and variance $\sigma^2$. In particular, a linear regression model is a set of probability distributions whose elements are indexed by the three-dimensional real vector $[m, b, \sigma^2]$.

The problem of parameter estimation for the linear regression model may be solved using a member of the general class of learning algorithms which were previously discussed in this paper. In particular, let the *training data* $u_n = \{(s^1, o^1), \ldots, (s^n, o^n)\}$ be a set of $n$ input/output pairs where $s^i \in \mathcal{R}$ is a value of the "input" or *predictor* variable and $o^i \in \mathcal{R}$ is the "output" or *outcome* variable. A linear regression learning algorithm is an algorithm $\mathbf{\Psi} : \mathcal{R}^3 \times T \times T \times U \rightarrow \mathcal{R}^3$ that maps an initial parameter vector $[m_0, b_0, \sigma_0^2] \in \mathcal{R}^3$, training data $u_n \in U$ at some initial time $t_0 \in T$, and some final time $t_f \in T$, into the final parameter estimates $[\hat{m}_n, \hat{b}_n, \hat{\sigma}_n^2] \in \mathcal{R}^3$.

It can be shown that the final parameter estimates $[\hat{m}_n, \hat{b}_n, \hat{\sigma}_n^2]$ have the additional property that: Given $[\hat{m}_n, \hat{b}_n, \hat{\sigma}_n^2]$, the likelihood function (which measures the probability of the observed data $u_n \in U$ given a parameter vector $[w, b, \sigma^2] \in \mathcal{R}^3$) $p(\cdot|\cdot) : U \times \mathcal{R}^3 \rightarrow [0, \infty)$ has the property that the global maximum of $p(u_n|\cdot)$ is $[w, b, \sigma^2]$.

The linear regression learning algorithm $\mathbf{\Psi}$ is an intelligent inference algorithm once the relational system associated with the learning algorithm has been explicitly identified. Let an element of $S$ be an assertion of the form: $S = \{Infer[w, b, \sigma^2] : [w, b, \sigma^2] \in \mathcal{R}^3\}$. Let $\mathcal{F}$ be the sigma-field generated by $S$. Let $\omega$ define an appropriateness relation on $\mathcal{F}$ such that $p(u_n|\cdot)$ is a measure for the relational system $(S, \mathcal{F}, \omega)$.

Since the relational system $(S, \mathcal{F}, \omega)$ has a measure, then it immediately follows that all optimal inferences with respect to this relational system satisfy connectivity and transitivity axioms of rational decision making. The relational system $(S, \mathcal{F}, \omega)$ also has the property that if one assumes that the marginal a priori probability distribution $p([w, b, \sigma^2])$ exists and is uniform on the set of global minima of $p(u_n|\cdot)$, then $p(\cdot|u_n)$ is proportional to $p(u_n|\cdot)$. This implies that the relational system $(S, \mathcal{F}, \omega)$ also has the probabilistic measure $p(\cdot|u_n)$, and thus behaves according to a specific set of constraints.

These points are best illustrated with the following simple example. Consider a linear regression algorithm

whose computations are implemented so that the algorithm consistently makes optimal inferences. The linear regression algorithm is used to estimate the parameters of a regression line for each of the two data sets which are depicted in Figure 1. If one simply defines the linear regression learning algorithm as an algorithm (possibly implemented by a computer program), then it is not clear in what sense the linear regression algorithm is intelligent (or if it is intelligent at all). One could construct a relational system with a sum-squared error measure, and then show that the algorithm is minimizing that sum-squared error measure to find the parameter estimates. Such an analysis would reveal that the algorithm has intelligence in the sense that it makes inferences with respect to the set of relational systems which satisfy the connectivity and transitivity axioms. However, the linear regression learning algorithm may be viewed as seeking the global minimum of a specific probabilistic measure for a relational system. A relational system with a probabilistic measure makes inferences according to an even more stringent set of decision making axioms than a relational system whose measure is simply an arbitrary objective function (e.g., the sum-squared error function).

The different ecological niches for the linear and logistic regression algorithms arise from their distinctive probabilistic modeling assumptions. From a probabilistic modeling perspective, the linear regression model expects the conditional mean of $\tilde{o}$ given $s$, $E[\tilde{o}|s]$ to be a linear function of $s$. The logistic regression model, on the other hand, expects that the conditional mean of $\tilde{o}$ given $s$, $E[\tilde{o}|s]$ be a sigmoidal function of $s$. Each algorithm makes intelligent and appropriate inferences with respect to its own relational system regardless of the data set. However, a given relational system may be appropriate for one environment (i.e., type of data set) but may not be appropriate with respect to another environment (i.e., another type of data set).

## SUMMARY

Just as biological bird flying algorithms may have strong expectations that glass-like obstructions do not exist, both the linear and logistic regression algorithms have strong expectations about their respective classes of probabilistic environments as illustrated in Figure 1. Biological bird flying algorithms, linear regression learning algorithms, logistic regression learning algorithms, Cohen-Grossberg artificial neural networks, and other algorithms that seek to minimize some real-valued objective function are all intelligent algorithms (in some sense) with respect to their designated ecological niches. The challenge for the engineer concerned with the analysis and design of recognition algorithms which are truly intelligent is to explicitly identify the sense in which a given algorithm is intelligent, explicitly identify how the "intelligence" of one algorithm differs from another, and explicitly identify the conditions under which an algorithm with a particular relational system of intelligence successfully achieves its computational goals. Explicitly identifying relational systems and exploring similarities and differences among such relational systems in an algorithm-independent manner is an important step towards achieving a true science of intelligent recognition algorithms.

## References


[1977] Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review, 84*, 413–451.

[1983] Cohen, M. & Grossberg, S. Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE Transactions on Systems on Man and Cybernetics, SMC-13*, 815–826.

[1946] Cox, R. T. Probability, frequency, and reasonable expectation. *American Journal of Physics, 14*, 1–13.

[1986] Golden, R. M. The brain-state-in-a-box neural model is a gradient descent algorithm. *Journal of Mathematical Psychology, 30*, 73–80.

[Forthcoming] Golden, R. M. *Mathematical Methods for Neural Network Analysis and Design.* Cambridge, MA: MIT Press.

[1982] Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, USA, 79*, 2554–2558.

[1984] Hopfield, J. J. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences, USA, 81*, 3088–3092.

[1972] Johnson-Laird, P. N., Legrenzi, P., & Legrenzi, M. Reasoning and a sense of reality. *British Journal of Psychology, 63*, 395–400.

[1979] Kahneman, D. & Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263–291.

[1962] Rosenblatt, F. *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms.* Washington D. C.: Spartan Books.

[1986] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing. Volume 1: Foundations* (pp. 318–362). Cambridge, MA: MIT Press.

[1969] Simon, H. A. *The sciences of the artificial.* Cambridge, MA: MIT Press.

[1969] Tversky, A. Intransitivity of preferences. *Psychological Review, 76*, 31–48.

[1966] Wason, P. C. Reasoning. In B. Foss (Ed.), *New horizons in psychology.* Middlesex, England: Penguin.

Figure 1: Data points represented by open circles correspond to a raw data set which is more compatible with a logistic regression model as opposed to a linear regression model. Data points represented by filled-in circles correspond to a raw data set which is more compatible with a linear regression model as opposed to a logistic regression model. Both linear regression and logistic regression are equally intelligent algorithms with respect to their "ecological niches".