Neural Networks, Vol. 1, pp. 325–337, 1988
Printed in the USA. All rights reserved.

0893-6080/88 $3.00 + .00

ORIGINAL CONTRIBUTION

Conjunctoids: Statistical Learning Modules for Binary Events

ROBERT J. JANNARONE, KAI F. YU, AND YOSHIYASU TAKEFUJI

University of South Carolina

(*Received January* 1988; *revised and accepted May* 1988)

Abstract—*A general family of fast and efficient neural network learning modules for binary events is introduced. The family subsumes probabilistic as well as functional event associations; subsumes all levels of input/output association; yields truly parallel learning processes; provides for optimal parameter estimation; points toward a workable description of optimal model performance; and yields procedures that are simple and fast enough to be serious candidates for reflecting both neural functioning and real time machine learning. Examples as well as operational details are provided.*

Keywords—Conjunctive measurement, Machine learning, Parallel distributed processing, Statistical pattern recognition, Nonlinear neural networks, Binary neural networks, Neurocomputing.

INTRODUCTION

Scope

Forty years ago the psychologist T. L. Kelley began his *Fundamentals of Statistics* with the compelling premise that, "An isolated fact is an unthinkable phenomenon" (Kelley, 1947). More recently the emerging neural network learning (NNL) movement (Grossberg, 1988a; Rumelhart & McClelland, 1986) has drawn credibility from the converse premise that all thought is based on associations among component facts. During the years following Kelley's book the statistics movement has refined a framework for describing and evaluating associations among component facts or events, which has taken centuries to develop. During its shorter history the NNL movement has in turn produced many neural models and modular learning "machines" for developing and utilizing associations among component events. Thus, both the statistics and the NNL movements have been based on evaluating associations among component variables. However, the NNL focus has been on primitive learning and performance structures, whereas the statistical focus has been on efficient estimation (learning) and decision making (performance) procedures.

Curiously, in developing its learning models the NNL movement has so far made little use of the associative framework that statistics has already developed (see Amari, 1988 and Anderson & Abrahams, 1987 for notable exceptions). This has perhaps been due to either a scarcity of active researchers in both fields, or a lack until recently of some adequate statistical procedures for NNL applications, or both. In either case the present seems like a good time for NNL modelers to make more use of existing statistical concepts. As one attempt to supply the NNL movement with a broader inferential footing, this report provides a statistical inference solution to an unsolved NNL problem: how to construct a family of machines that can *quickly and efficiently*: (a) learn from experience how *any* "input" set of binary (true or false) events is related to any other "output" binary event set; and (b) use the associations learned in (a) to choose the best possible output event set for each input set.

Purpose

The purpose of this report is to introduce a general family of fast and efficient NNL learning modules for binary events called "conjunctoids," by employing an appropriate framework from probability theory; adapting a class of recently developed conjunctive models from psychometric theory; tailoring sound statistical estimation and evaluation schemes to fit NNL learning needs; and presenting a detailed functional description of the required conjunctoid circuitry.

This is an abridged version of a detailed report, which is available from the authors upon request.

Requests for reprints should be sent to Robert J. Jannarone, Department of Electrical and Computer Engineering, University of South Carolina, Columbia, SC 29208.

OVERVIEW

Some Learning Task Examples

All of the models that we will present are based on associations among M binary input variables, $\mathbf{x} = (x_1, \ldots, x_M)$, and N binary output variables, $\mathbf{y} = (y_1, \ldots, y_N)$. To fix ideas, we will use two examples throughout this section: learning to recognize the parity of an M-variate binary vector, and learning to recognize any of 2^N distinct stimuli from a visual display broken down into M binary (e.g., presence or absence) sectors.

Basic Concepts: Multinomial Conjunctoids

Conjunctoids are functional NNL modules that are based on a probability framework, which treats each observed $(\mathbf{x}, \mathbf{y})$ combination as a realization of a multivariate (binary) random variable,

$$\underset{1\times(M+N)}{\mathbf{W}} = (\underset{1\times M}{\mathbf{X}}, \underset{1\times N}{\mathbf{Y}}), \tag{1}$$

where N may be either 1 as in the parity example or greater than 1 as in the pattern recognition example. The probability framework also assumes the existence of specific likelihood functions for samples based on $\mathbf{W}$. These likelihoods include estimable parameters that can be used to both evaluate and utilize $(\mathbf{X}, \mathbf{Y})$ associations, hence reflecting machine learning and performing functions, respectively. When $M + N$ is small and reflecting all possible associations among $\mathbf{X}$ and $\mathbf{Y}$ is necessary, it is convenient to assume that $\mathbf{W}$ has the *multinomial* likelihood,

$$Pr\{\mathbf{W} = \mathbf{w} \mid \underset{1\times(M+N)}{\alpha}\} = \frac{\alpha_{\mathbf{w}}}{\sum_{\mathbf{u}\in\mathcal{B}^{M+N}} \alpha_{\mathbf{u}}}, \quad \mathbf{w} \in \mathcal{B}^{M+N},$$
$$= 0 \quad \text{elsewhere}, \tag{2}$$

where $\mathcal{B}^{M+N} = \{\mathbf{w}: w_k = 0, 1; k = 1, \ldots, 2^{M+N}\}$ and the parameter vector α satisfies $0 \le \alpha_{\mathbf{u}} \le 1$ $(\mathbf{u} \in \mathcal{B}^{M+N})$. Multinomial conjunctoid *learning* occurs during a series of learning trials, when $\mathbf{W}$ values are observed and α values are estimated.

A useful consequence of the parametric probability framework is that conditional output probabilities for given input values can also be described by estimable parameters. For the multinomial case these probabilities take the form,

$$Pr\{\mathbf{Y} = \mathbf{y} \mid \mathbf{X} = \mathbf{x}; \alpha\} = \frac{\alpha_{(\mathbf{x},\mathbf{y})}}{\sum_{\mathbf{v}\in\mathcal{B}^N} \alpha_{(\mathbf{x},\mathbf{v})}}. \tag{3}$$

Multinomial machine *performance* occurs when an action represented by a specific $\mathbf{y}$ value is selected, based on a specific input $\mathbf{x}$ value along with estimated α values from previous learning trials.

In addition to assuming a parametric likelihood for observable $(\mathbf{x}, \mathbf{y})$ values, it is useful to include in the probability framework a Bayes model for likelihood parameters. For the multinomial case, imposing a Bayes structure entails treating the appropriate α for each multinomial learning application as a realization from a second random variable distinct from $\mathbf{W}$. Imposing a Bayes structure also involves assuming a reasonable "prior" probability model for α, in a way that will be described later.

Figure 1 illustrates how a multinomial machine learns parity in the case where $M = 3$ and "almost" no Bayes structure is used ("almost," because defining probabilities before the first learning trial requires a weak Bayes prior, *q.v.*). Initially, the conjunctoid assigns a probability of .5 to both $y = 1$ and $y = 0$ for each possible $\mathbf{x}$ value, in lieu of any experience that would point toward the correct y values. This is indicated by the value of .5 for the 16 estimated output probability graphs in Figure 1 at learning trial 0. The top graph in Figure 1 shows a sequence of 14 hypothetical $(\mathbf{x}, y)$ learning trial values. The effect of the first trial value, $(\mathbf{x}, y) = (\mathbf{0}, 0)$, is shown in the estimated output probability graph for $\mathbf{x} = \mathbf{0}$. During the first learning trial the estimate of $Pr\{Y = 1 \mid \mathbf{X} = \mathbf{0}\}$ shifts from .5 to 0, whereas the estimated $Pr\{Y = 0 \mid \mathbf{X} = \mathbf{0}\}$ shifts from .5 to 1. Other learning trials have similar effects on appropriate y probabilities, as Figure 1 shows.

To illustrate performance functioning for the multinomial learning sequence in Figure 1, the bottom Figure 1 graph plots the likelihood of correctly choosing output y values as a function of learning trial number (assuming equally likely input $\mathbf{x}$ values). Before the first learning trial, the machine will choose arbitrarily among the two equally likely y values for each possible $\mathbf{x}$ value, yielding an expected correct guess rate of .5. Between the first and second learning trials the machine will correctly guess the y value when $\mathbf{x} = \mathbf{0}$, but it will guess randomly when $\mathbf{x} \neq \mathbf{0}$. At that point the correct guess probability will be

$$(1) \times (1/8) + (.5) \times (7/8) = 9/16,$$

and so on for the next 13 trials as indicated in the graph.

Moving finally to a NNL circuitry description, Figure 2 contains a schematic diagram for a multinomial conjunctoid module. The diagram is made up of interconnections among several functional units, called elementary processing units, that include 2^{M+N} parameter estimators, a parameter multiplexer, 2^N output pattern accumulators, and an output comparator. As Figure 2 indicates, multinomial conjunctoid circuitry can also be grouped into larger "experience" and "performance" segments, which function as follows. During each learning cycle the experience segment receives prior/learning data and sends them to the parameter estimators, which in turn send current parameter estimates to the performance segment. At the start of each learning cycle a unit of prior/learning data—consisting of a single $(\mathbf{x}, \mathbf{y})$ observation, $\mathbf{w}$, along with a positive learning importance weight, L—is passed to

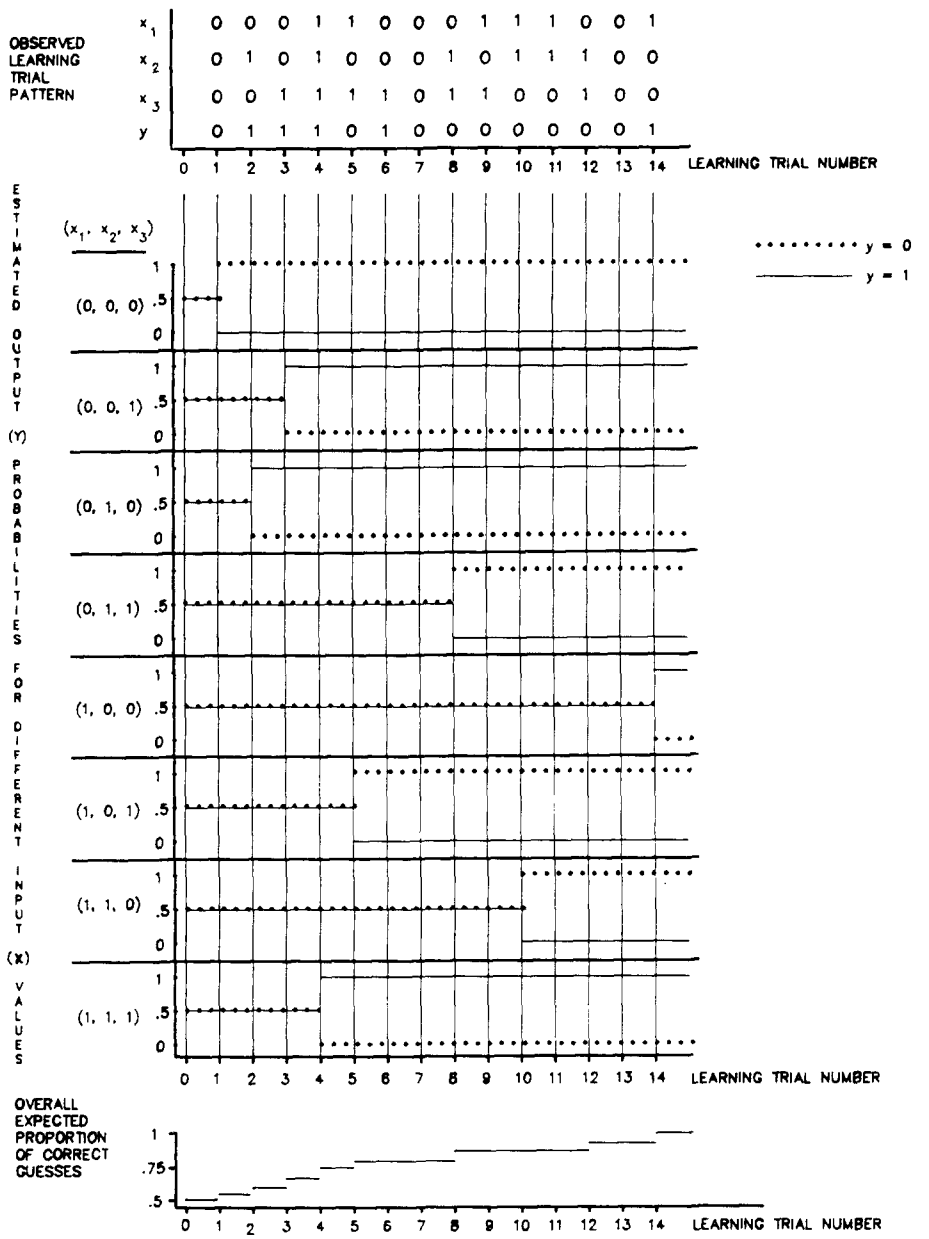

FIGURE 1. A parity learning illustration.

each parameter estimator. Next, each parameter estimator—consisting of an input indicator, an estimate updator, and an output register—performs two functions: (a) the input indicator sets a flag, u, to 1, if $\mathbf{w}$ matches the parameter for its estimator unit, and to 0 otherwise; and (b) the parameter updator modifies the output register by setting

$$\hat{\alpha}_{\text{new}} = \frac{\hat{\alpha}_{\text{old}} + L\mathbf{u}}{1 + L}. \tag{4}$$

Before any prior or learning trials occur, $\hat{\alpha}$ values are set to an initial value of .5 for each parameter. Also, each parameter estimator performs separately from and simultaneously with all others, so that each learning cycle is very fast.

Regarding performance unit functioning, just as the experience segment of Figure 2 executes one learning cycle for each input $(\mathbf{w}, L)$ learning unit, the performance segment executes one behavior cycle for each input $\mathbf{x}$ value. At the beginning of each behavior cycle, the parameter multiplexer uses the input $\mathbf{x}$ value to admit only the 2^N parameter estimates, $\{\hat{\alpha}_{(\mathbf{x},\mathbf{u})}, \mathbf{u} \in \mathcal{B}^N\}$ associated with the input $\mathbf{x}$ value, among the 2^{M+N} parameter estimates coming from the experience segment. Next, each output pattern accumulator selects and stores the single estimate coming from the parameter multiplexer that corresponds to its associated $\mathbf{y}$ value. Finally, the output comparator identifies the single output pattern having the highest estimated parameter value and outputs its $\mathbf{y}$ value. Thus, as with the

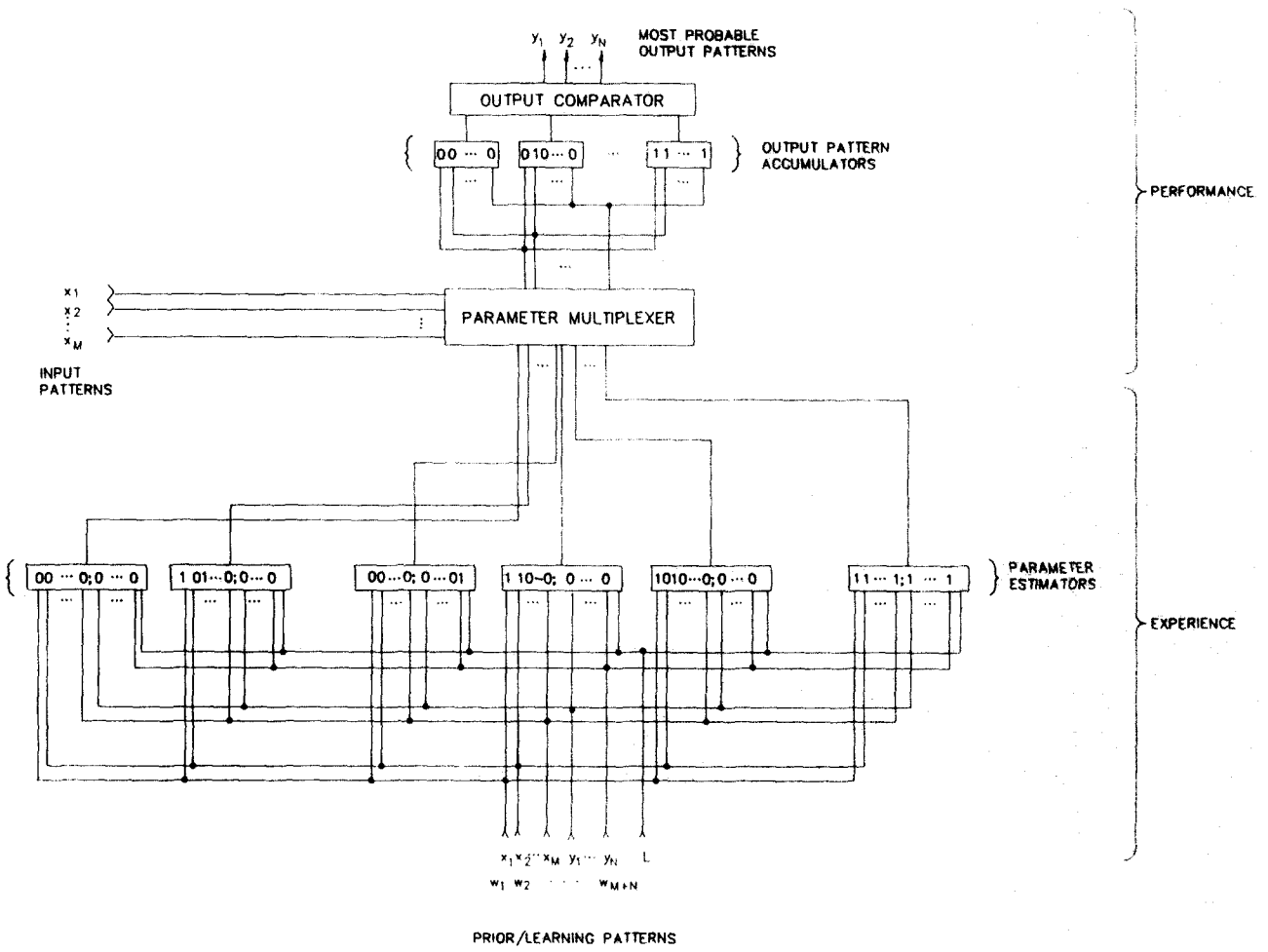

FIGURE 2. A multinomial conjunctoid.

experience segment the performance segment functions quite simply. Moreover, each behavior cycle is quick because the parameter multiplexer performs no computations, the output pattern accumulators function simultaneously, and the output comparator's sole task is to locate the address containing the largest value among 2^N words of storage.

The Conjunctoid Family

The conjunctoids that we will describe next have three distinct advantages over the multinomial version. First, they require far fewer than the necessary $2^{M+N} + 2^N + 2$ elementary processing units associated with Figure 2. Second, they produce parameter estimates that can directly suggest the simplest underlying $\mathbf{x}$, $\mathbf{y}$ associations. Finally, for many applications nonmultinomial conjunctoids require far fewer learning trials to produce a given level of performance accuracy, because they estimate far fewer parameters. We will begin by introducing a representative—so-called third-order—conjunctoid and follow with an overview of the general family.

As in the multinomial case the probability model for third-order machines assumes the existence of a K-variate random variable, $\mathbf{W}$, where $K = M + N$. However, in place of multinomial probabilities third-order conjunctive probabilities take the form,

$$Pr\{\mathbf{W} = \mathbf{w} \mid \underset{1\times K}{\beta^{(1)}}, \underset{1\times\binom{K}{2}}{\beta^{(2)}}, \underset{1\times\binom{K}{3}}{\beta^{(3)}}\}$$

$$= \nu(\beta^{(1)}, \beta^{(2)}, \beta^{(3)}) \exp\{\sum_{k=1}^{K} \beta_k^{(1)} w_k + \sum_{k=1}^{K-1} \sum_{m=k+1}^{K} \beta_{km}^{(2)} w_k w_m + \sum_{k=1}^{K-2} \sum_{m=k+1}^{K-1} \sum_{n=m+1}^{K} \beta_{kmn}^{(3)} w_k w_m w_n\}, \quad (5)$$

where the estimable parameters in $\beta^{(1)}$, $\beta^{(2)}$, and $\beta^{(3)}$ are real-valued and the positive normalizing function ν ensures that all probabilities will sum to **1**. The term "third-order" implies that the probabilities defined by (5) are third-degree polynomials in the observable binary events, w_1 through w_K. Also, since the elements of $\mathbf{w}$ are binary the probabilities in (5) may be considered as third-order conjunctive functions, in that they depend on third order conjuncts among the elements of $\mathbf{w}$.

As with the multinomial version, third-order machines perform by using conditional probabilities corresponding to (5) rather than using (5) directly. The

pertinent conditional probabilities may be expressed as,

$$Pr\{\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}; \beta^{(1)}, \beta^{(2)}, \beta^{(3)}\} = \pi(\mathbf{x}, \beta^{(1)}, \beta^{(2)}, \beta^{(3)}) \times \exp\{\sum_{k=1}^{M}\sum_{m=M+1}^{k-1}\sum_{n=m+1}^{k} x_k y_m y_n + \sum_{n=M+1}^{K}(\sum_{k=1}^{M}\beta_{kn}^{(2)}x_k + \sum_{k=1}^{M-1}\sum_{m=k+1}^{M}\beta_{kmn}^{(3)}x_k x_m)y_n\}\exp\{\sum_{n=M+1}^{K}\beta_n^{(1)}y_n + \sum_{m=M+1}^{K-1}\sum_{n=M+1}^{K}\beta_{mn}^{(2)}y_m y_n + \sum_{k=M+1}^{K-2}\sum_{m=k+1}^{K-1}\sum_{n=m+1}^{K}\beta_{kmn}^{(3)}y_k y_m y_n\}, \quad (6)$$

where π is another normalizing function.

For a random sample of L learning trials satisfying (5), it can be shown that the joint likelihood based on (5) is monotonically related to,

$$\sum_{k=1}^{K}\beta_k^{(1)}s_k^{(1)} + \sum_{k=1}^{K-1}\sum_{m=k+1}^{K}\beta_{km}^{(2)}s_{km}^{(2)} + \sum_{k=1}^{K-2}\sum_{m=k+1}^{K-1}\sum_{n=m+1}^{K}\beta_{kmn}^{(3)}s_{kmn}^{(3)}, \quad (7)$$

where the statistics $s_k^{(1)}$, $s_{km}^{(2)}$, and $s_{kmn}^{(3)}$ are proportions of the L trials for which their corresponding first, second and third-order conjuncts were 1.

Bayes structure for the third-order case closely follows the multinomial case. Bayes structure can easily be imposed on (5) by replacing any statistic based on L learning trials in (7), say s_L, with

$$s_{\text{posterior}} = \frac{Is_{\text{prior}} + Ls_L}{I + L}. \quad (8)$$

I in (8) is the "prior sample size," s_{prior} is the proportion of times that the "prior statistic corresponding to s_L occurred in the prior sample," and $s_{\text{posterior}}$ is the resulting composite statistic. Conjunctoid functioning for the third-order case also parallels the multinomial case. Functioning for both cases can be broken down into experience and performance segments, with experience resulting in learning *via* parameter estimation and performance yielding behavior in the form of selecting most likely **y** values given **x**.

Third-order conjunctoids estimate parameters by a conditional maximum likelihood (CML) method. The CML method finds an estimate for each β value in (6) based on its corresponding sample s value in (5) and conditional upon all other concurrent s values in (5). The advantage of the CML approach is that estimation for each parameter does not involve other parameters in the model. Instead, separate CML functions for each parameter—depending only on that parameter and its corresponding statistics—are used to find each CML parameter estimate. Also, each CML function is simple, well-behaved, and amenable to an elementary line search method (see the Estimation Details section). *Most importantly, the CML estimation method is consistent over learning trials and can in principle be implemented in only one read-only-memory (ROM) fetch cycle* (Yu & Jannarone, 1987).

Third-order and multinomial conjunctoids are two members of the large, general conjunctoid family. Each family member may be indexed by a set of subscripts defining both its parameters and its statistics. For third-order machines the indexing set is,

$$\mathcal{P}_3 = \{1, 2, \ldots, K, (1, 2), (1, 3), \ldots, (K-1, K), (1, 2, 3), (1, 2, 4), \ldots, (K-2, K-1, K)\}, \quad (9)$$

indicating that all first, second, and third-order parameters and their conjuncts appear in the third-order probability model (5). Indexing-sets for all conjunctoid family members are restrictions, $\mathcal{R}$, of the fully parameterized conjunctoid that is indexed by the so-called power set, $\mathcal{F}$, which includes all possible subsets of $\{1, 2, \ldots, K\}$. The family may also be described as including all Pth-order conjunctoids, $\mathcal{P}_P(P = 1, \ldots, K)$, as well as all of their special cases that could be obtained by fixing some parameters at 0 (or equivalently removing the parameters and their conjuncts from the model).

Conjunctoid Hardware Summary

Figure 3 contains a schematic diagram for a third-order module. The diagram shows the same types of elementary processing units—as well as the same experience and performance segments—that Figure 2 shows for the multinomial case. Each estimator in Figure 3 consists of an input indicator, a statistic updator, a bounds comparator, and an estimate updator. Each input indicator begins every learning cycle by setting an indicator flag, u, to 1 if the learning trial value of **w** "covers" its corresponding parameter. Next, the statistic updator modifies its statistic register by setting,

$$s_{\text{new}} = \frac{s_{\text{old}} + L\mathbf{u}}{1 + L}, \quad (10)$$

where L plays the same weighting role as in the multinomial case. After statistic values have been updated during a learning cycle, the necessary values for computing upper and lower bounds are sent from each statistic register to all appropriate bounds evaluation registers. Finally, after their upper and lower bounds have been evaluated the estimate updators fetch new CML parameter estimates from appropriate ROM locations according to current s, $\underline{s}$, and $\bar{s}$ values (see *Conditional Probabilities* below).

The third-order performance segment indicated in Figure 3 is nearly the same as its multinomial counterpart in Figure 2, although functioning is more detailed in the third-order case. The role of the Figure 3 parameter multiplexer is to send appropriate "joint input–output parameters" for a given **x** value to the output pattern accumulators, in accordance with the conditional probabilities in (6). Once each output accu-

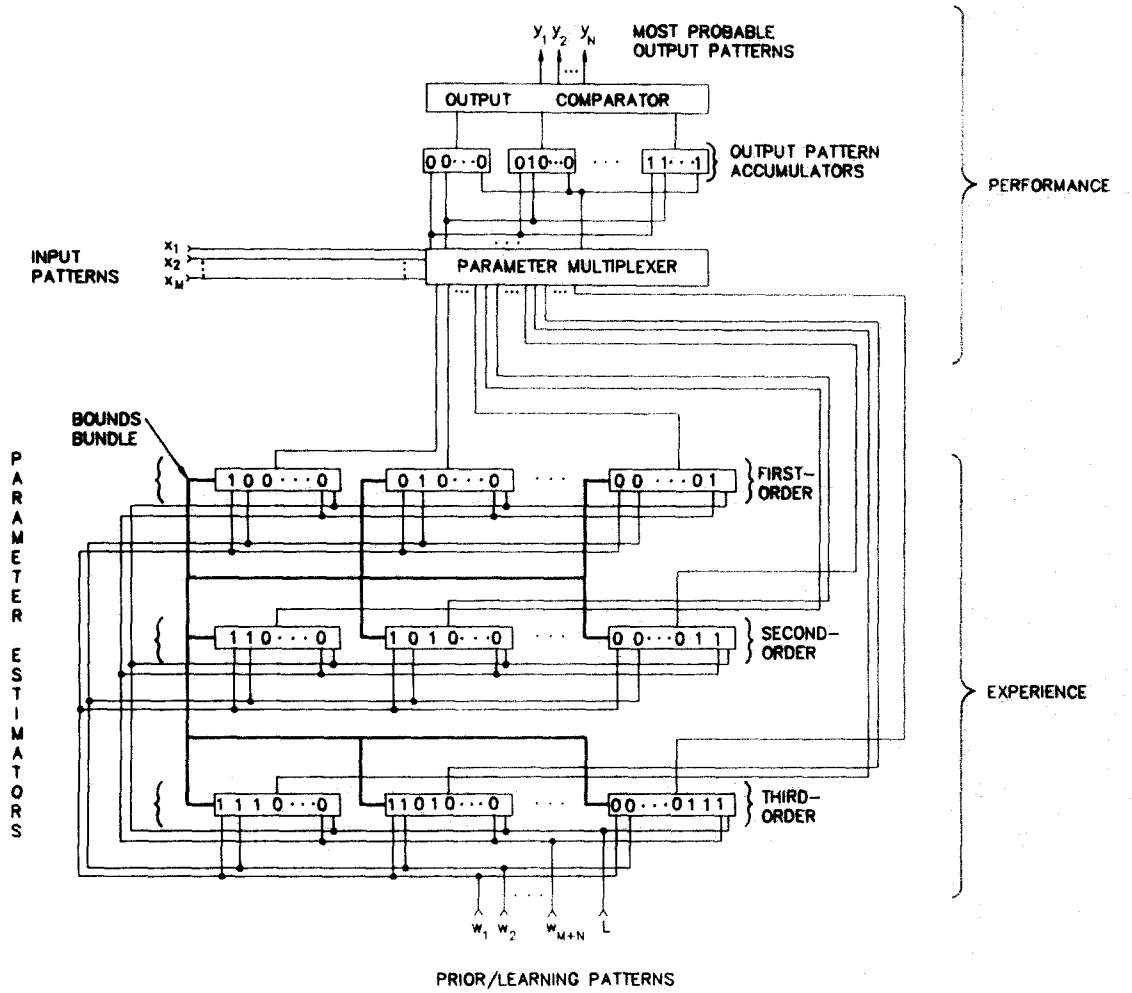

FIGURE 3. A third-order conjunctoid.

mulator has received all of its appropriate parameter values, it simply sums them up. Finally, the third-order output comparator functions precisely as in the multinomial case, by finding the largest output pattern accumulator value at the end of each behavior cycle.

Turning next to third-order hardware efficiency, each probability estimator in Figure 3 requires storage for its statistic value, each of its potential lower and upper bound statistics, its parameter estimate value, and its ROM. Also, each third-order pattern accumulator requires storage for each of its parameter estimates and for its summing circuitry. Otherwise, the memory requirements for third-order elementary processing units are the same as for their multinomial counterparts. Regarding execution time, since all third-order estimators function simultaneously the execution time for each estimator is the same as the time for an entire learning cycle. The third-order statistic updator takes the same amount of time as the entire multinomial learning cycle. In addition, the third-order estimator must transfer its bound statistics, identify their most restrictive values, and locate its CML estimate value. All three of these additional functions can be performed quite quickly, however. All other third-order functions take the same time to perform as their multinomial counterparts, with the exception of the third-order output accumulator's functioning—its parameter summing takes slightly more time. In sum: third-order storage requirements are much smaller overall—though larger per elementary processing unit—than those for multinomial machines; and third-order functioning is slower than multinomial, although not much slower.

Functioning for Pth-order machines, with $P = 2, 4, 5, \ldots, K$, is similar to third-order conjunctoid functioning. However, bounds identification becomes quite complicated for high-order cases (see the *Conditional Probabilities* section below). Other conjunctoid family members may be constructed as Pth-order versions by excluding selected parameter values. Parameter values may be effectively excluded by fixing their parameter estimates at 0 and removing their connections to other parameter estimators.

Related Work

In this section we will briefly review some conjunctoid-related results within the NNL, statistical pattern recognition, mathematical statistics, psychometrics, and biometrics fields. For excellent reviews see Grossberg (1988a; 1987) and Rumelhart and McClelland (1986). From the conjunctoid perspective the key NNL results have been (a) a focus on fast and parallel processing (Rumelhart & McClelland, 1986), (b) the perceptron learning algorithm, and (c) modern attempts, especially in the form of sigma-pi units (Feldman & Ballard, 1982; Rumelhart, Hinton, & McClelland, 1986) to overcome perceptron limitations.

Noniterative processing is essential for neural modeling, because neurons simply function too slowly and humans respond too quickly for serial processing to be feasible (Crick & Asanuma, 1986; Grossberg, 1982). This simple fact not only rules out the entire von Neumann (traditional serial subroutine) paradigm as a basis for much of neural information processing; it also provides much of the driving force for the paradigm shift that is currently underway toward distributed models of cognition (Grossberg, 1982; McClelland, Rumelhart, & Hinton, 1986).

Perceptrons (Feldman, 1982; Minsky & Papert, 1969; Rumelhart & Zipster, 1986) were the first serious models for fast, parallel processing. They included many features that appear in current NNL models, including an error-correction approach rather than a traditional statistical approach to machine learning. This early NNL emphasis on error correction is not surprising, because a statistical approach based on the standard parameter estimation methods at that time would have required prohibitively slow iterations at each learning trial. Also, perceptrons were analogous to second-order conjunctoids in that they would learn only if the relationship between input and output variables was linear.

Sigma-pi units (Amari, 1977; Feldman, 1981; Feldman & Ballard, 1982; Grossberg, 1969; Grossberg, 1987b; Kohonen, 1977; Rumelhart, Hinton, & McClelland, 1986), can reflect all forms of conjunctive logic. Like perceptrons, sigma-pi units use error-correction as a means for learning. However, sigma-pi learning schemes are necessarily more complicated than the perceptron learning algorithm, requiring a process called "back propagation" (Rumelhart, Hinton, & Williams, 1986). Back propagation involves adjusting learning weights associated with so-called "hidden units," and leads to some additional sigma-pi unit problems (Rumelhart, Hinton, & Williams, 1986). These include: no provisions for representing the optimal configuration of hidden units associated with a given learning task; a potentially long, iterative process of weight adjustment and **y** estimation that must proceed until estimated and actual **y** values coincide (Ono & Fushikida, 1987; Sejnowski & Rosenberg, 1987); no guarantee that suboptimal solutions (local optima) will not result during parameter estimation; no guarantee that sigma-pi back propagation units are sufficiently general to reflect all learning situations; and no provisions for *gradual* learning over a series of trials.

Conjunctoids are potentially more powerful than sigma-pi units utilizing back propagation, because they do not require iterative updating and they use sound statistical procedures rather than error correction as a basis for learning. Conjunctoids have a further advantage over sigma-pi units that is quite important. Unlike perceptrons, sigma-pi units carry no guarantee of convergence to proper learning states as the number of learning trials increases. Indeed, much attention is currently being given to this limitation and ways of resolving it. By sharp contrast, the statistical theory of exponential families guarantees that conjunctoid estimation procedures are consistent. Finally, as indicated in the preceding parity example, conjunctoids include a natural mechanism for retaining and incorporating prior learned information. A similar mechanism has not yet been presented for sigma-pi units.

Conjunctoids include many other underlying concepts that are similar to existing ideas in the NNL literature. These include potential provisions for "unlearning" (Hinton & Sejnowski, 1986; Hopfield, Feinstein, & Palmer, 1983) by simply by making L negative; existing back propagation provisions for prior weighting (learning rates) that are quite similar to (4) and (8) (Rumelhart, Hinton, & Williams, 1986); existing NNL models that are similar to multinomial conjunctoids (e.g., the so-called probabilistic conjunctive encoders—Hinton, McClelland, & Rumelhart, 1986—see also Anderson & Abrahams, 1987; 1986); models called Boltzman machines (Hinton & Sejnowski, 1986) that have some probabilistic features like conjunctoids but severe difficulties associated with back propagation; and many other similarities—too many to list here.

Turning next to statistical pattern recognition, conjunctoids are natural pattern recognizers as one of the examples for this report illustrates. In that regard they closely resemble the wide variety of statistical pattern recognizers that have been studied (Devijver & Kittler, 1982). Existing pattern recognition jargon includes terms to describe many of the concepts that have been introduced here, including "features" (independent variables), "training/design," "contextual information" (e.g., using Markov models to focus on spatial proximity), and "nearest neighbor decision rules" (e.g., choosing the most probable **y** value given **x**). Indeed, statistical pattern recognition is more similar to conjunctoid theory than any alternatives that have been discussed up until now. Some key differences exist for statistical pattern recognition models as well, however.

Most notably, statistical pattern recognition has not yet produced models with the generality, speed, and computing compatibility of conjunctoids. (Some special cases seem quite close, however—see Marroquin, Mitter, & Poggio, 1987; Pickard, 1987.) Finally, conjunctoids have the potential for reflecting much more than pattern learning abilities. Their potential includes modeling the learning of associations among *any* binary variables, including logical variables that could reflect a variety of expertise, knowledge, and attitudes, as well as resulting choices and other behaviors.

Regarding related work from psychometrics, the statistical theory of mental tests (Lord & Novick, 1968) is fundamentally similar to NNL theory, in that both have been primarily concerned with associations among binary events. In the psychometric setting the binary events correspond to pass versus fail scores on test items, whereas in the NNL setting the binary events correspond to dependent and independent logical variable values. Psychometric test theory has differed, however, in that it has traditionally attempted to explain all binary event associations in terms of only one causal (ability) variable. On the other hand, the recent introduction of conjunctive item response theory (Jannarone, 1986; 1988; Jannarone, Laughlin, & Yu, 1988) has provided psychometrics with a much broader class of models and methods for reflecting binary event associations. It is from this class of models and methods that conjunctoids have been conceived.

Turning finally to related developments in mathematical statistics, the power of statistical theory lies in its formalization of decision making processes based on uncertain information. Modern advances include the Neyman-Pearson estimation and hypothesis testing theory (Neyman, 1967; Lehmann, 1983, 1986), Bayesian theory (Box & Tiao, 1973; Savage, 1954), and a general decision framework that includes Neyman-Pearson models, Bayes models, and other concepts as well (Ferguson, 1967; Wald, 1950). In its most general form, statistical decision theory assigns costs (or utilities) to different decisions based on observed random variable values. For each possible data value, loss (or utility) functions are formulated that specify the cost associated with each resulting decision about "states of nature," given the true "states of nature."

Loss functions are typically formulated in reasonable ways, so that if a decision accurately reflects nature's true state then its loss value will be zero; otherwise positive loss values are assigned that reflect how severe the discrepancies are between decisions about nature and nature's actual states. For example, in pattern recognition cases nature's true states take the form of actual stimuli (dependent variables) that are presented; data take the form of independent variable values that are generated by actual stimulus parameters (the data can be random in that the stimuli can be presented randomly and the same stimuli can lead to different perceptions/independent variable values); and simple loss functions can be formulated such that if the learning machine guesses the correct stimulus then the loss value will be 0—otherwise the loss value will be 1.

At its best, statistical decision theory points toward optimal decision strategies in the face of uncertainty. Because of the uncertainty aspect, however, criteria for optimality must be described in probabilistic terms. For example, most reasonable pattern recognition models are formulated such that two distinct stimuli can sometimes produce the same perceptions. In this case, no matter what kind of decision rule is formulated it is possible that the rule will sometimes yield incorrect decisions. That is, no decision procedure can be provided that will be absolutely perfect. Instead, the only reasonable optimality criteria in such cases must include probabilistic notions such as minimizing *expected* loss, maximizing *expected* utility, and so on.

A further notion from statistical decision theory that pertains to conjunctoids is the concept of asymptotic optimality. For the conjunctoid case the major asymptotic optimality consideration is whether a given conjunctoid and underlying estimation procedure will have optimal expected loss as the number of learning trials increases. As it happens, this type of optimality is guaranteed by the consistency of CML estimates (Yu & Jannarone, 1987). (However, conjunctoids based on alternative procedures such as unconditional maximum likelihood estimation may be more asymptotically *efficient*—see Lehmann, 1983.)

In sum, statistical decision theory has much to offer theories of machine learning, because it provides a straightforward framework for representing optimal decisions under uncertainty and for subsequently identifying optimal procedures. However, several criteria for optimality—both finite and asymptotic—will need to be considered in order to do the machine learning problem justice.

Other related results from mathematical statistics include specific statistical (decision making) procedures that are currently available and related to conjunctoid procedures. These include linear discriminant analysis for continuous variables (Anderson, 1984), linear and nonlinear discriminant analysis for discrete variables (Goldstein & Dillon, 1978; Lachenbruch, 1975), linear and nonlinear regression (Draper & Smith, 1966), CML estimation (Andersen, 1980; Barndorff-Nielsen, 1978), and conjugate Bayes estimation (Bickel & Doksum, 1977; Novick & Jackson, 1974). The results in this report offer no new formulations relative to these existing statistical results, except the two new results that have already been cited (Jannarone, Laughlin, & Yu, 1988; Yu & Jannarone, 1987). Instead, our emphasis here has been on selecting the combination of existing results from statistics and psychometrics that have resulted in general as well as fast conjunctoids. Finally, Anderson and Abrahams (1987, 1986) have

introduced a general probability framework for NNL, along with an outline for nonparametric estimation. Conjunctoids may be viewed as a family of special cases, each having a viable parameter, sufficient statistic, estimation, and real-time hardware implementation structure.

PROBABILITY DETAILS

Joint Probabilities for Binary Events

We will begin this section by formulating a general class of conjunctive probability models, after which we will focus on some special cases. First, consider a K-variate random variable, $\mathbf{W}$, satisfying

$$Pr_{\mathcal{R}}\{\mathbf{W} = \mathbf{w}|\boldsymbol{\beta}\} = \nu_{\mathcal{R}}(\boldsymbol{\beta})\exp\{\sum_{(k_1,\ldots,k_s)\in\mathcal{R}} \beta_{k_1,\ldots,k_s} w_{k_1}\cdots w_{k_s}\}, \quad \mathbf{w}\in\mathcal{B}^K, = 0 \quad \text{elsewhere}, \tag{11}$$

where

$$\mathcal{R}\subseteq\mathcal{F} = \{(k_1,\ldots,k_s), k_m = 1,\ldots,K, m = 1,\ldots,s, 1\le k_1<k_2<\cdots<k_s, s = 1,\ldots,K\},$$

$$\boldsymbol{\beta} = \{\beta_{k_1\cdots k_s}, (k_1,\ldots,k_s)\in\mathcal{R}\},$$

$$\mathcal{B}^K = \{\underset{1\times K}{\mathbf{w}} : w_k = 0, 1, k = 1,\ldots,K\},$$

and

$$\nu_{\mathcal{R}}(\boldsymbol{\beta}) = [\sum_{\mathbf{u}\in\mathcal{B}^K}\exp\{\sum_{(k_1,\ldots,k_s)\in\mathcal{R}}\beta_{k_1\cdots k_s}u_{k_1}\cdots u_{k_s}\}]^{-1}.$$

It follows that for a sequence of L independent observations, $\mathbf{w}_1,\ldots,\mathbf{w}_L$, that are identically distributed according to (11), their joint likelihood is,

$$Pr_{\mathcal{R}}\{\mathbf{W}_1 = \mathbf{w}_1,\ldots,\mathbf{W}_L = \mathbf{w}_L|\boldsymbol{\beta}\} = [\nu(\boldsymbol{\beta})]^L\exp\{\sum_{(k_1,\ldots,k_s)\in\mathcal{R}}\beta_{k_1\cdots k_s}\sum_{i=1}^{L} w_{ik_1}\cdots w_{ik_s}\}. \tag{12}$$

It also follows from exponential family theory (Lehmann, 1983) that the vector of observed conjunct proportions,

$$\mathbf{s}_{\mathcal{R}}(\mathbf{w}_1,\ldots,\mathbf{w}_L) = \left(\left(\frac{1}{L}\sum_{i=1}^{L} w_{ik_1}\cdots w_{ik_s}\right)\right), \quad (k_1,\ldots,k_s)\in\mathcal{R}, \tag{13}$$

is a sufficient statistic for $\boldsymbol{\beta}$, whence the term "conjunctive probability models."

It will sometimes be useful to label the elements of $\boldsymbol{\beta}$, $\mathbf{s}$, and the like sequentially from 1 to R, the number of elements in $\mathcal{R}$. In such cases a single subscript j will be used in place of each $k_1\cdots k_s$ subscript, where $j:\mathcal{R}\to\{1,\ldots,R\}\to j(1)<j(2)<\cdots<j(K)<j(1,2)<j(1,3)<\cdots<j(1,2,\ldots,K)$. Thus, the likelihood (12) can be expressed as,

$$Pr_{\mathcal{R}}\{\underset{1\times R}{\mathbf{S}} = \underset{1\times R}{\mathbf{s}} \mid \underset{1\times R}{\boldsymbol{\beta}}, L\} = [\nu_{\mathcal{R}}(\boldsymbol{\beta})]^L\exp\{L\sum_{r=1}^{R}\beta_r s_r\}, \quad \mathbf{s}\in\mathcal{S}_L(\mathcal{R}), = 0 \quad \text{elsewhere}, \tag{14}$$

where

$$\mathcal{S}_L(\mathcal{R}) = \{\underset{1\times R}{\mathbf{S}} = \mathbf{s}_{\mathcal{R}}(\mathbf{u}_1,\ldots,\mathbf{u}_L), \mathbf{u}_i\in\mathcal{B}^K, i = 1,\ldots,L\}, L = 1,\ldots.$$

The first special cases of (11) to consider are the so-called Pth-order conjunctive probability models defined by,

$$\mathcal{R} = \mathcal{P}^P = \{1,\ldots,K,(1,2),(1,3),\ldots,(K-1,K),\ldots,(1,\ldots,P),\ldots,(K-P+1,\ldots,K)\},$$

and yielding likelihoods of the form,

$$Pr_{\mathcal{P}^P}\{\mathbf{W}_1 = \mathbf{w}_1,\ldots,\mathbf{W}_L = \mathbf{w}_L|\boldsymbol{\beta}\} = [\nu_{\mathcal{P}^P}(\boldsymbol{\beta})]^L\exp\{\sum_{k=1}^{K}\beta_k\sum_{i=1}^{L}w_{ik} + \sum_{k_1=1}^{K-1}\sum_{k_2=k_1+1}^{K}\beta_{k_1k_2}\sum_{i=1}^{L}w_{ik_1}w_{ik_2} + \cdots + \sum_{k_1=1}^{K-P+1}\cdots\sum_{k_P=k_{P-1}+1}^{K}\beta_{k_1\cdots k_P}\sum_{i=1}^{L}w_{i1}\cdots w_{iP}\}. \tag{15}$$

When expressed in terms of sufficient statistics (15) becomes,

$$Pr_{\mathcal{P}^P}\{\underset{1\times Q}{\mathbf{S}} = \underset{1\times Q}{\mathbf{s}} \mid \underset{1\times Q}{\boldsymbol{\beta}}\} = h_L(s)[\nu_{\mathcal{P}^P}(\boldsymbol{\beta})]^L\exp\{L\sum_{j=1}^{Q}\beta_j s_j\}, \quad \mathbf{s}\in\mathcal{S}_L(\mathcal{P}^P),$$

where the $h_L(s)$ are defined below and

$$Q = \sum_{p=1}^{P}\binom{K}{p}.$$

The Kth-order special case of (15) is equivalent to the multinomial case, which is simpler to formulate as follows. For a parameter space defined by,

$$\{\underset{1\times 2^K}{\boldsymbol{\alpha}} : 0\le\alpha_{\mathbf{u}}\le 1, \mathbf{u}\in\mathcal{B}^K, \sum_{\mathbf{u}\in\mathcal{B}^K}\alpha_{\mathbf{u}} = 1\},$$

we have

$$Pr_{\mathcal{M}}\{\mathbf{W} = \mathbf{w}|\boldsymbol{\alpha}\} = \alpha_{\mathbf{w}}, \quad \mathbf{w}\in\mathcal{B}^K = 0 \quad \text{elsewhere}, \tag{16}$$

so that for a random sample of size L,

$$Pr_{\mathcal{M}}\{\mathbf{W}_1 = \mathbf{w}_1,\ldots,\mathbf{W}_L = \mathbf{w}_L|\boldsymbol{\alpha}\} = \prod_{i=1}^{L}\alpha_{\mathbf{w}_i} = \prod_{\mathbf{u}\in\mathcal{B}^K}\alpha_{\mathbf{u}}^{Ls_{\mathbf{u}}(\mathbf{w}_1,\ldots,\mathbf{w}_L)} = \prod_{j=1}^{F}\alpha_j^{Ls_j}, \tag{17}$$

where $s_{\mathbf{u}}$ is the proportion of $\mathbf{u}$-valued patterns among $\mathbf{w}_1$ through $\mathbf{w}_L$, j is the decimal equivalent of the binary value corresponding to $\mathbf{u}$, and $F = 2^K$.

Other special cases of interest include the Markov models, the simplest of which is the first-order one-dimensional Markov field defined by,

$$\mathcal{R} = \{1, \ldots, K, (1, 2), (2, 3), \ldots, (K-1), K\}.$$

Conjugate Bayes Structure

Since conjunctive probability models are members of the exponential family, known results (e.g., Bickel & Doksum, 1977, Prop. 2.4.1) can be used to form workable conjugate conjunctive structures. In particular, $\forall \mathcal{R} \subseteq \mathcal{F}$ proper conjugate prior densities can be obtained by setting

$$g_{\mathcal{R}}(\boldsymbol{\beta} | \mathbf{b}, I) \propto [\nu_{\mathcal{R}}(\boldsymbol{\beta})]^I \exp\{I \sum_{(k_1,\ldots,k_s)\in\mathcal{R}} \beta_{k_1\cdots k_s} b_{k_1\cdots k_s}\},$$

or equivalently,

$$g_{\mathcal{R}}(\underset{1\times R}{\boldsymbol{\beta}} | \underset{1\times R}{\mathbf{b}}, I) \propto [\nu_{\mathcal{R}}(\boldsymbol{\beta})]^I \exp\{\sum_{r=1}^{R} \beta_r b_r\}, \quad (18)$$

provided that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [\nu_{\mathcal{R}}(\boldsymbol{\beta})]^I \exp\{\sum_{r=1}^{R} \beta_r b_r\} d\beta_1 \cdots d\beta_R < \infty. \quad (19)$$

If (19) is not satisfied then conjugate prior densities will be improper (Bickel & Doksum, 1980). In either case, the prior density in (18) and the likelihood in (14) will result in posterior densities of the form,

$$h_{\mathcal{R}}(\underset{1\times R}{\boldsymbol{\beta}} | \underset{1\times R}{\mathbf{b}}, I, \underset{1\times R}{\mathbf{s}}, L) \propto [\nu_{\mathcal{R}}(\boldsymbol{\beta})]^{I+L} \exp\{\sum_{r=1}^{R} \beta_r (Ib_r + Ls_r)\}. \quad (20)$$

A useful consequence of (18) through (20) is that if I and $\mathbf{b}$ are chosen such that

$$\mathbf{b} \in \mathcal{S}_I(\mathcal{R}) \quad (21)$$

then the posterior expression (20) will be proportional to the likelihood (14). It follows that any workable likelihood maximization estimation method based on (14) could also yield posterior maximizing estimates based on (20). All posterior distributions to be considered in the sequel will be of the (20) type with $\mathbf{b}$ and I satisfying (21).

The multinomial version of the posterior expression (20) is,

$$h_{\mathcal{M}}(\boldsymbol{\alpha} | \underset{1\times F}{\mathbf{a}}, I, \underset{1\times F}{\mathbf{s}}, L) \propto \prod_{j=1}^{F} \alpha_j^{Ia_j + Ls_j}. \quad (22)$$

(The requirement that I and L be integer-valued will be dropped in the sequel—given that the b_r and s_r are proportions, all of the above results apply $\forall I, L > 0$ without loss of generality.)

Conditional Probabilities

Two conditional probability model types will be described: models for some of the w_k probabilities given the others, and models for each of the s_r probabilities given all others. Beginning with the first type, for

$$\underset{1\times K}{\mathbf{W}} = (\underset{1\times M}{\mathbf{X}}, \underset{1\times N}{\mathbf{Y}}),$$

(11) may also be expressed as,

$$Pr_{\mathcal{R}}\{(\mathbf{X}, \mathbf{Y}) = (\mathbf{x}, \mathbf{y}) | \boldsymbol{\beta}\}$$
$$= \nu_{\mathcal{R}}(\boldsymbol{\beta}) \exp\{\sum_{(m_1,\ldots,m_s)\in\mathcal{R}_{\mathbf{x}}} \beta_{m_1\cdots m_s} x_{m_1}\cdots x_{m_s}$$
$$+ \sum_{(n_1,\ldots,n_t)\in\mathcal{R}_{\mathbf{y}}} \beta_{n_1\cdots n_t} y_{n_1}\cdots y_{n_t}$$
$$+ \sum_{(m_1,\ldots,m_s,n_1,\ldots,n_t)\in\mathcal{R}_{\mathbf{xy}}} \beta_{m_1\cdots m_s n_1\cdots n_t} x_{m_1}\cdots x_{m_s} y_{n_1}\cdots y_{n_t}\},$$
$$\mathbf{x} \in \mathcal{B}^M, \quad \mathbf{y} \in \mathcal{B}^N,$$
$$= 0 \quad \text{elsewhere}, \quad (23)$$

where

$$\mathcal{R}_{\mathbf{x}} = \{(m_1, \ldots, m_s) \in \mathcal{R}, m_1, \ldots, m_s \le M\},$$
$$\mathcal{R}_{\mathbf{y}} = \{(n_1, \ldots, n_t) \in \mathcal{R}, M+1 \le n_1, \ldots, n_t \le N\},$$

and

$$\mathcal{R}_{\mathbf{xy}} = \mathcal{R} - \mathcal{R}_{\mathbf{x}} - \mathcal{R}_{\mathbf{y}}.$$

Thus,

$$Pr_{\mathcal{R}}\{\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}; \boldsymbol{\beta}\} = \frac{Pr\{(\mathbf{X}, \mathbf{Y}) = (\mathbf{x}, \mathbf{y}) | \boldsymbol{\beta}\}}{\sum_{v\in\mathcal{B}^N} Pr\{(\mathbf{X}, \mathbf{Y}) = (\mathbf{x}, v) | \boldsymbol{\beta}\}}$$
$$= \frac{\exp\{\sum_{(n_1,\ldots,n_t)\in\mathcal{R}_{\mathbf{y}}} \beta_{n_1\cdots n_t} y_{n_1}\cdots y_{n_t} + \sum_{(m_1,\ldots,m_s,n_1,\ldots,n_t)\in\mathcal{R}_{\mathbf{xy}}} \beta_{m_1\cdots m_s n_1\cdots n_t} x_{m_1}\cdots x_{m_s} y_{n_1}\cdots y_{n_t}\}}{\sum_{v\in\mathcal{B}^N} \exp\{\sum_{(n_1,\ldots,n_t)\in R_{\mathbf{y}}} \beta_{n_1\cdots n_t} v_{n_1}\cdots v_{n_t} + \sum_{(m_1,\ldots,m_s,n_1,\ldots,n_t)\in R_{\mathbf{xy}}} \beta_{m_1\cdots m_s n_1\cdots n_t} x_{m_1}\cdots x_{m_s} v_{n_1}\cdots v_{n_t}\}}. \quad (24)$$

For the multinomial case we have,

$$Pr_{\mathcal{M}}\{\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}; \boldsymbol{\alpha}\} = \frac{\alpha_{(\mathbf{x},\mathbf{y})}}{\sum_{\mathbf{u}\in\mathcal{B}^N} \alpha_{(\mathbf{x},\mathbf{u})}}, \quad \mathbf{y} \in \mathcal{B}^N, \quad \mathbf{x} \in \mathcal{B}^M,$$
$$= 0 \quad \text{elsewhere}. \quad (25)$$

For the conjunctive case it follows that **Y** is independent of **X** if and only if $\forall(m_1, \ldots, m_s; n_1, \ldots, n_t) \in \mathcal{R}_{\mathbf{xy}}$, $\beta_{m_1 \ldots m_s n_1 \ldots n_t} = 0$. It also follows that $Pr_{\mathcal{R}}\{\mathbf{Y}|\mathbf{X}; \boldsymbol{\beta}\}$ is of the same order in **X** as the largest t value in (24). Finally, it follows that the Y_n are mutually independent if and only if $\beta_{n_1 \ldots n_t} = 0\ \forall n_1, \ldots, n_t \in \mathcal{R}_y - \{M+1, M+2, \ldots, N\}$.

Turning next to conditional likelihoods, $\forall \mathbf{s} \in \mathcal{S}_L(\mathcal{R})$ and $q = 1, \ldots, R$, we have,

$$Pr\{S_q = s_q | S_1 = s_1, \ldots, S_{q-1} = s_{q-1}, S_{q+1} = s_{q+1}, \ldots, S_R = s_R\}$$

$$= \frac{h_L(\mathbf{s})\exp\{L\sum_{r=1}^{R}\beta_r s_r\}}{\sum_{v \in \mathcal{S}(\mathcal{R}|\mathbf{s},q)} h_L(s_1, \ldots, s_{q-1}, v, s_{q+1}, \ldots, s_R)\exp\{L\beta_q v + L\sum_{\substack{r=1 \\ r\neq q}}^{R}\beta_r s_r\}}$$

$$= \frac{h_L(\mathbf{s})\exp\{L\beta_q s_q\}}{\sum_{v \in \mathcal{S}(\mathcal{R}|\mathbf{s},q)} h_L(s_1, \ldots, s_{q-1}, v, s_{q+1}, \ldots, s_R)\exp\{L\beta_q v\}}, \quad s_q \in \mathcal{S}\{R|\mathbf{s}, q\},$$

$$= 0 \quad \text{elsewhere,} \tag{26}$$

where $h_L(\mathbf{s})$ is the number of ways that **s** can be observed in samples of size L and

$$\mathcal{S}(\mathcal{R}|\mathbf{s}, q) = \{s_q: \mathbf{s} \in \mathcal{S}_L(\mathcal{R})\}.$$

Turning next to "conditional posterior likelihoods" based on (20), we consider conditional probabilities associated with each of the "posterior sufficient statistic" values in

$$\mathbf{t} = \frac{1}{I+L}(I\mathbf{b} + L\mathbf{s})$$

and conditional upon the others. We have,

$$Pr\{T_q = t_q | T_1 = t_1, \ldots, T_{q-1} = t_{q-1}, T_{q+1} = t_{q+1}, \ldots, T_R = t_R; \beta_q\}$$

$$= \frac{h_L(\mathbf{t})\exp\{(I+L)\beta_q t_q\}}{\sum_{v \in \mathcal{S}(\mathcal{R}|\mathbf{t},q)} h_L(t_1, \ldots, t_{q-1}, v, t_{q+1}, \ldots, t_r)\exp\{(I+L)\beta_q v\}},$$

$$q = 1, \ldots, R, t_q \in \mathcal{S}(\mathcal{R}|\mathbf{t}, q),$$

$$= 0 \quad \text{elsewhere.} \tag{27}$$

The important consequence of (26) and (27) is that conditional probabilities for any sufficient statistic given the others depend only on the single parameter associated with that sufficient statistic.

Finally, expressions for the $\mathcal{S}(\mathcal{R}|\mathbf{s}, q)$ based on Pth-order conjunctive models will be provided without proof. For all $(k_1, \ldots, k_t) \in \mathcal{P}^P$ we have,

$$\mathcal{S}(\mathcal{P}^P|\mathbf{s}, k_1, \ldots, k_t) = \{\mathbf{s}_{k_1 \ldots k_t}(\mathbf{s}) + j/L,$$

$$j = 0, 1, \ldots, L(\bar{s}_{k_1 \ldots k_t}(\mathbf{s}) - \underline{s}_{k_1 \ldots k_t}(\mathbf{s}))\},$$

where the $\underline{s}_{k_1 \ldots k_t}$ and $\bar{s}_{k_1 \ldots k_t}$ depend on **s** as follows: the $\underline{s}_{k_1 \ldots k_t}$ are maxima among terms that include 0 when $t = P$, $\{s_{k_1 \ldots k_{t+1}}: s_{k_1 \ldots k_{r-1}k_{r+1} \ldots k_{t+1}} = s_{k_1 \ldots k_t}, r = 1, \ldots, t+1\}$ when $t < P$, and $\{s_{k_1 \ldots k_r} + s_{k_{r+1} \ldots k_t} - 1, [(k_1, \ldots, k_r), (k_{r+1}, \ldots, k_t)]$ is a partition of $\{k_1, \ldots, k_t\}, r = 1, \ldots, t\}$ when $P > 1$; the $\underline{s}_{k_1 \ldots k_t}$ are minima among terms that include 1 when $P = 1$, $\{1 + s_{k_1 \ldots k_t m_1 \ldots m_r} - s_{m_1 \ldots m_r}, r = 1, \ldots, P-t\}$ when $t = 1, \ldots, P-1$ and $P = 2, \ldots, k$, and $\{S_{k_1 \ldots k_{r-1}k_{r+1} \ldots k_t}, r = 1, \ldots, t\}$ when $t = 2, \ldots, P$ and $P = 2, \ldots, K$.

ESTIMATION

Conditional Likelihood Maximization

Because conjunctive models having conditional likelihoods given by (26) are in the exponential family, it follows (Yu & Jannarone, 1987) that given mild regularity conditions (a) unique maximum likelihood estimates $\tilde{\beta}_q$ exist for each β_q based on s_q and conditional upon $s_1, \cdots, s_{q-1}, s_{q+1}, \cdots, s_R$, except when s_q takes on boundary values (e.g., $\underline{s}_q$ and $\bar{s}_q$ for Pth-order models), (b) the CML log-likelihoods corresponding to (26) are concave, and (c) the resulting CML estimates are consistent.

For Pth-order conjunctoids conditional likelihoods take the form,

$$\frac{h_L(\mathbf{s})\exp\{\beta_q L s_q\}}{\sum_{v=L\underline{s}_q}^{L\bar{s}_q} h_L(s_1, \ldots, s_{q-1}, v/L, s_{q+1}, \ldots, s_R)\exp\{\beta_q L v\}}. \tag{28}$$

Details associated with obtaining accurate CML estimates based on (28) are beyond this article's scope

and will be given elsewhere, although the general conditions can be outlined at this point. Although (28) depends on sample size as well as **s** values the conjunctoid scheme requires that adequate CML estimates for any sample size be stored in a ROM. Our approach entails treating each learning trial as if the effective (prior plus data) sample associated with current **s** values were large and using accurate approximations based on (28) accordingly. These approximations utilize the consistency of (28); the fact that if $\hat{\beta}_q$ values maximize monotone functions of (28) then they are also CML estimates; limiting integral forms of the denominator in (28); and limiting forms of the $h_L(\mathbf{s})$ in (28) based on Stirling's approximation.

Conditional Posterior Likelihoods

Just as conditional posterior likelihoods are formally equivalent to conditional likelihoods, so are conditional posterior MLEs equivalent to conditional MLEs. Therefore no further developments are necessary for conditional posterior likelihood maximization. Instead, in this section we will outline a procedure for generating "posterior sufficient statistics" based on prior belief along with likelihood data. First, any type of prior belief for a given conjunctive model can be expressed in terms of β_0, a vector of prior values for β. Second, expected prior sufficient statistic values corresponding to β_0 can be obtained by evaluating,

$$\mathbf{b} = \mathcal{E}(\mathbf{S}|\beta_0). \tag{29}$$

Finally, posterior sufficient statistics can be formed by setting,

$$\mathbf{t} = \frac{1}{L_{\text{prior}} + L_{\text{sample}}} [L_{\text{prior}}\mathbf{b} + L_{\text{sample}}\mathbf{s}], \tag{30}$$

with L_{prior} being chosen to reflect prior belief strength relative to L_{sample}.

Conditional posterior likelihood maximization can be useful in avoiding boundary value problems (Jannarone, Laughlin, & Yu, 1988). It can be shown that if **b** is chosen by setting

$$\mathbf{b} = \mathcal{E}(\mathbf{S}|\beta_0 = \mathbf{0})$$

and if $L_{\text{prior}} > 0$ then no boundary values of **t** will occur.

The Multinomial Case

Conditional likelihood maximization is not necessary for multinomial conjunctoids, because multinomial parameters can be separately estimated through a simpler (unconditional) maximum likelihood approach. Maximum likelihood estimates of the α_j in (17) are merely their corresponding s_j proportions, which pose no boundary problems.

Posterior Bayes structure is also easier to implement for multinomial machines. In the multinomial case prior belief as reflected by $\alpha_0 = \mathbf{a}$ can also be directly reflected by prior sufficient statistic values of **a**. Also, relative prior belief strengths can be specified by setting the prior sample size L_{prior} at appropriate values relative to L_{sample}, as in the conjunctive case. Posterior maximum likelihood estimates can then be obtained by simply setting the posterior sufficient statistics to

$$\mathbf{t} = \frac{1}{L_{\text{prior}} + L_{\text{sample}}} (L_{\text{prior}}\mathbf{a} + L_{\text{sample}}\mathbf{s}). \tag{31}$$

SUMMARY

A general family of learning modules for binary events has been introduced that is based on: probabilistic notions including random variables, conditional probabilities, and conjugate Bayes structures; psychometric formulations that feature conjuncts among component events as sufficient statistics; conditional maximum likelihood and posterior modal estimation schemes; and modern computer design features including VLSI ROM technology. The resulting modules have been shown to be general—all relationships among binary events are special cases as are many different kinds of learning schemes; fast—noniterative parameter estimation is the key to practically real-time learning potential; and realistic—feasible models for a variety of machine and neural learning functions are special cases. Finally, a variety of necessary steps for future development of the learning models has been proposed.

REFERENCES

Amari, S. A. (1988). Statistical neurodynamics of associative memory. *Neural Networks*, **1**, 63–74.

Amari, S. A. (1977). A mathematical approach to neural systems. In J. Metzler (Ed.), *Systems neuroscience* (pp. 67–117). New York: Academic Press.

Andersen, E. B. (1980). *Discrete statistical models with social science applications*. Amsterdam: North Holland.

Anderson, C. H., & Abrahams, E. (1987). The Bayes connection. In M. Caudill & C. Butler (Eds.), *Proceedings of the First International Conference on Neural Networks*. New York: IEEE Press.

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis* (2nd ed.). New York: Wiley.

Barndorff-Nielsen, O. (1978). *Information and exponential families*. New York: Wiley.

Bickel, P. J., & Doksum, K. A. (1977). *Mathematical statistics: Basic ideas and concepts*. San Francisco: Holden-Day.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, MA. Addison-Wesley.

Crick, F., & Asanuma, C. (1986). Certain aspects of the anatomy and physiology of the cerebral cortex. In Rumelhart, D. E. & McClelland, J. C. (Eds.), *Parallel distributed processing, Vol. I*. Cambridge, MA: MIT Press.

Devijver, P. A., & Kittler, J. (1982). *Pattern recognition: A statistical approach*. Englewood Cliffs, NJ: Prentice Hall.

Draper, N. R., & Smith, H. (1966). *Applied linear regression*. New York: Wiley.

Feldman, J. A. (1981). A connectionist model of vision memory. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory*. Hillsdale, NJ: Erlbaum.

Feldman, J. A. (1982). Dynamic connections in neural networks. *Biological Cybernetics,* **46,** 27–39.

Feldman, J. A., & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science,* **6,** 205–254.

Ferguson, T. S. (1967). *Mathematical statistics: A decision theoretic approach.* New York: Academic Press.

Goldstein, M., & Dillon, W. R. (1978). *Discrete discriminant analysis.* New York: Wiley.

Grossberg, S. (1969). On learning and energy-entropy dependence in recurrent and nonrecurrent signed networks. *Journal of Statistical Physics,* **1,** 319–350.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding, I: Parallel development and coding of neural feature detectors. *Biological Cybernetics,* **23,** 187–202.

Grossberg, S. (1982). *Studies of mind and brain: Neural principles of learning, perception, development, cognition and motor control.* Boston: Reidel Press.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science,* **11,** 23–63.

Grossberg, S. (1988). Nonlinear neural networks: Principles, mechanism, and architectures. *Neural Networks,* **1,** 17–62.

Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzman machines. In Rumelhart, D. E. & McClelland, J. C. (Eds.), *Parallel distributed processing, Vol. I.* Cambridge, MA: MIT Press.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In Rumelhart, D. E. & McClelland, J. C. (Eds.), *Parallel distributed processing, Vol. I.* Cambridge, MA: MIT Press.

Hopfield, J. J., Feinstein, D. I., & Palmer, R. G. (1983). "Unlearning" has a stabilizing effect in collective memories. *Nature,* **304,** 158–159.

Jannarone, R. J. (1986). Conjunctive item response theory kernels. *Psychometrika,* **50,** 357–373.

Jannarone, R. J. (1988). Locally dependent models for reflecting learning abilities. *Psychometrika,* in review.

Jannarone, R. J., Laughlin, J. E., & Yu, K. F. (1988). Easy Bayes estimation for Rasch-type models. To appear in *Psychometrika.*

Jannarone, R. S., Yu, K. F., & Takefuji, Y. (1987). *Conjunctoids: Statistical learning models for binary events* (Tech. Rep. No. 87-65). University of South Carolina, Center for Machine Intelligence.

Kelley, T. J. (1947). *Fundamentals of statistics.* Cambridge, MA: Harvard University Press.

Kohonen, T. (1977). *Associative memory: A system theoretical approach.* New York: Springer.

Lachenbruch, P. A. (1975). *Discriminant analysis.* New York: Hafner.

Lashley, K. S. (1950). In search of the engram. In *Society of Experimental Biology Symposium No. 4: Psychological mechanisms in animal behavior* (pp. 478–505). London: Cambridge University Press.

Lehmann, E. L. (1983). *Theory of point estimation.* New York: Wiley.

Lehmann, E. L. (1986). *Testing statistical hypotheses* (2nd ed.). New York: Wiley.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Marroquin, J., Mitter, S., & Poggio, T. (1987). Probabilistic solution of ill-posed problems in computational vision. *Journal of the American Statistical Association,* **82,** 76–89.

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The appeal of NNL. In Rumelhart, D. E. & McClelland, J. C. (Eds.), *Parallel distributed processing, Vol. I.* Cambridge, MA: MIT Press.

Minsky, M., & Papert, S. (1969). *Perceptrons.* Cambridge, MA: MIT Press.

Neyman, J. (1967). *Joint statistical papers of J. Neyman and E. S. Pearson.* Berkeley: University of California Press.

Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research.* New York: McGraw-Hill.

Ono, & Fushikida (1987, September). *Text to phoneme conversions based on neural nets.* Paper presented at the 35th Information Processing National Conference, Hokkaido, Japan. [In Japanese]

Pickard, D. K. (1987). Inference for discrete Markov fields: The simplest nontrivial case. *Journal of the American Statistical Association,* **82,** 90–95.

Rumelhart, D. E., & McClelland, J. L. (Eds.). (1986). *Parallel distributed processing, Vols. I & II.* Cambridge, MA: MIT Press.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In Rumelhart, D. E. & McClelland, J. C. (Eds.), *Parallel distributed processing, Vol. I.* Cambridge, MA: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representation by error propagation. In Rumelhart, D. E., & McClelland, J. C. (Eds.), *Parallel distributed processing, Vol. I.* Cambridge, MA: MIT Press.

Rumelhart, D. E., & Zipster, D. (1986). Feature Discovery by competitive learning. In Rumelhart, D. E. & McClelland, J. C. (Eds.), *Parallel distributed processing, Vol. I.* Cambridge, MA: MIT Press.

Savage, L. J. (1954). *The foundations of statistics.* New York: Wiley.

Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce English texts. *English systems,* **1,** 145–168.

Wald, A. (1950). *Statistical decision functions.* New York: Wiley.

Yu, K. F., & Jannarone, R. J. (1987). *Conjunctoid conditional maximum likelihood estimation.* Statistics Department, University of South Carolina Technical Report.