Hierarchical Mixtures of Experts and the EM Algorithm

Michael I. Jordan
Department of Brain and Cognitive Sciences,
Massachusetts Institute of Technology, Cambridge, MA 02139 USA

Robert A. Jacobs
Department of Psychology, University of Rochester,
Rochester, NY 14627 USA

We present a tree-structured architecture for supervised learning. The statistical model underlying the architecture is a hierarchical mixture model in which both the mixture coefficients and the mixture components are generalized linear models (GLMs). Learning is treated as a maximum likelihood problem; in particular, we present an Expectation-Maximization (EM) algorithm for adjusting the parameters of the architecture. We also develop an on-line learning algorithm in which the parameters are updated incrementally. Comparative simulation results are presented in the robot dynamics domain.

1 Introduction

The principle of divide-and-conquer is a principle with wide applicability throughout applied mathematics. Divide-and-conquer algorithms attack a complex problem by dividing it into simpler problems whose solutions can be combined to yield a solution to the complex problem. This approach can often lead to simple, elegant, and efficient algorithms. In this paper we explore a particular application of the divide-and-conquer principle to the problem of learning from examples. We describe a network architecture and a learning algorithm for the architecture, both of which are inspired by the philosophy of divide-and-conquer.

In the statistical literature and in the machine learning literature, divide-and-conquer approaches have become increasingly popular. The CART algorithm of Breiman et al. (1984), the MARS algorithm of Friedman (1991), and the I3S algorithm of Quinlan (1986) are well-known examples. These algorithms fit surfaces to data by explicitly dividing the input space into a nested sequence of regions, and by fitting simple surfaces (e.g., constant functions) within these regions. They have convergence times that are often orders of magnitude faster than gradient-based neural network algorithms.
Although divide-and-conquer algorithms have much to recommend them, one should be concerned about the statistical consequences of dividing the input space. Dividing the data can have favorable consequences for the bias of an estimator, but it generally increases the variance. Consider linear regression, for example, in which the variance of the estimates of the slope and intercept depends quadratically on the spread of data on the x-axis. The points that are the most peripheral in the input space are those that have the maximal effect in decreasing the variance of the parameter estimates.

The foregoing considerations suggest that divide-and-conquer algorithms generally tend to be variance-increasing algorithms. This is indeed the case and is particularly problematic in high-dimensional spaces where data become exceedingly sparse (Scott 1992). One response to this dilemma—that adopted by CART, MARS, and ID3, and also adopted here—is to utilize piecewise constant or piecewise linear functions. These functions minimize variance at a cost of increased bias. We also make use of a second variance-decreasing device: a device familiar in the neural network literature. We make use of "soft" splits of data (Bridle 1989; Nowlan 1991; Wahba et al. 1993), allowing data to lie simultaneously in multiple regions. This approach allows the parameters in one region to be influenced by data in neighboring regions. CART, MARS, and ID3 rely on "hard" splits, which, as we remarked above, have particularly severe effects on variance. By allowing soft splits the severe effects of lopping off distant data can be ameliorated. We also attempt to minimize the bias that is incurred by using piecewise linear functions, by allowing the splits to be formed along hyperplanes at arbitrary orientations in the input space. This lessens the bias due to high-order interactions among the inputs and allows the algorithm to be insensitive to the particular choice of coordinates used to encode the data (an improvement over methods such as MARS and ID3, which are coordinate-dependent).

The work that we describe here makes contact with a number of branches of statistical theory. First, as in our earlier work (Jacobs et al. 1991), we formulate the learning problem as a mixture estimation problem (cf. Cheeseman et al. 1988; Duda and Hart 1973; Nowlan 1991; Redner and Walker 1984; Titterington et al. 1985). We show that the algorithm that is generally employed for the unsupervised learning of mixture parameters—the Expectation–Maximization (EM) algorithm of Dempster et al. (1977)—can also be exploited for supervised learning. Second, we utilize generalized linear model (GLM) theory (McCullagh and Nelder 1983) to provide the basic statistical structure for the components of the architecture. In particular, the "soft splits" referred to above are modeled as multinomial logit models—a specific form of GLM. We also show that the algorithm developed for fitting GLMs—the iteratively reweighted least squares (IRLS) algorithm—can be usefully employed in our model, in particular as the M step of the EM algorithm. Finally, we show that these ideas can be developed in a recursive manner, yielding a tree-structured approach to estimation that is reminiscent of CART, MARS, and ID3.

The remainder of the paper proceeds as follows. We first introduce the hierarchical mixture-of-experts architecture and present the likelihood function for the architecture. After describing a gradient descent algorithm, we develop a more powerful learning algorithm for the architecture that is a special case of the general Expectation–Maximization (EM) framework of Dempster et al. (1977). We also describe a least-squares version of this algorithm that leads to a particularly efficient implementation. Both of the latter algorithms are batch learning algorithms. In the final section, we present an on-line version of the least-squares algorithm that in practice appears to be the most efficient of the algorithms that we have studied.

2 Hierarchical Mixtures of Experts

The algorithms that we discuss in this paper are supervised learning algorithms. We explicitly address the case of regression, in which the input vectors are elements of $\mathbb{R}^n$ and the output vectors are elements of $\mathbb{R}^m$. We also consider classification models and counting models in which the outputs are integer-valued. The data are assumed to form a countable set of paired observations $X = \{(x^i, y^i)\}$. In the case of the batch algorithms discussed below, this set is assumed to be finite; in the case of the on-line algorithms, the set may be infinite.

We propose to solve nonlinear supervised learning problems by dividing the input space into a nested set of regions and fitting simple surfaces to the data that fall in these regions. The regions have "soft" boundaries, meaning that data points may lie simultaneously in multiple regions. The boundaries between regions are themselves simple parameterized surfaces that are adjusted by the learning algorithm.

The hierarchical mixture-of-experts (HME) architecture is shown in Figure 1. The architecture is a tree in which the gating networks sit at the nonterminals of the tree. These networks receive the vector $x$ as input and produce scalar outputs that are a partition of unity at each point in the input space. The expert networks sit at the leaves of the tree. Each expert produces an output vector $\mu_y$ for each input vector. These output vectors proceed up the tree, being blended by the gating network outputs.

All of the expert networks in the tree are linear with a single output nonlinearity. We will refer to such a network as "generalized linear," borrowing the terminology from statistics (McCullagh and Nelder 1983).

---

1To simplify the presentation, we restrict ourselves to a two-level hierarchy throughout the paper. All of the algorithms that we describe, however, generalize readily to hierarchies of arbitrary depth. See Jordan and Xu (1993) for a recursive formalism that handles arbitrary hierarchies.
Mixtures of Experts and EM Algorithm

The gating networks are also generalized linear. Define intermediate variables $\xi_i$ as follows:

$$\xi_i = \mathbf{v}_i^T \mathbf{x}$$  \hspace{1cm} (2.2)

where $\mathbf{v}_i$ is a weight vector. Then the $i$th output of the top-level gating network is the “softmax” function of the $\xi_i$ (Bridle 1989; McCullagh and Nelder 1983):

$$g_i = \frac{e^{\xi_i}}{\sum_k e^{\xi_k}}$$  \hspace{1cm} (2.3)

Note that the $g_i$ are positive and sum to one for each $\mathbf{x}$. They can be interpreted as providing a “soft” partitioning of the input space.

Similarly, the gating networks at lower levels are also generalized linear systems. Define $\xi_{ij}$ as follows:

$$\xi_{ij} = \mathbf{v}_{ij}^T \mathbf{x}$$  \hspace{1cm} (2.4)

Then

$$g_{ij} = \frac{e^{\xi_{ij}}}{\sum_k e^{\xi_{ik}}}$$  \hspace{1cm} (2.5)

is the output of the $j$th unit in the $i$th gating network at the second level of the architecture. Once again, the $g_{ij}$ are positive and sum to one for each $\mathbf{x}$. They can be interpreted as providing a “soft” sub-partition of the input space nested within the partitioning provided by the higher-level gating network.

The output vector at each nonterminal of the tree is the weighted output of the experts below that nonterminal. That is, the output at the $i$th nonterminal in the second layer of the two-level tree is

$$\mathbf{u}_i = \sum_j g_{ij} \mathbf{u}_j$$

and the output at the top level of the tree is

$$\mathbf{u} = \sum_i g_i \mathbf{u}_i$$

Note that both the $g$'s and the $\mathbf{u}$'s depend on the input $\mathbf{x}$, thus the total output is a nonlinear function of the input.

\[\text{2.1 Regression Surface}\]

Given the definitions of the expert networks and the gating networks, the regression surface defined by the hierarchy is a piecewise blend of the regression surfaces defined by the experts. The gating networks provide a nested, “soft” partitioning of the input space and the expert networks provide local regression surfaces within the partition. There is overlap between neighboring regions. To understand the nature of the overlap, consider a one-level hierarchy with two
interpret the gating networks as modeling the input-dependent, multinomial probabilities associated with decisions at particular levels of resolution in a tree-structured model of the data.

Once a particular sequence of decisions has been made, resulting in a choice of regressive process \((i,j)\), output \(y\) is assumed to be generated according to the following statistical model. First, a linear predictor \(\eta_{ij}\) is formed:

\[
\eta_{ij} = \beta_i^j x
\]

The expected value of \(y\) is obtained by passing the linear predictor through the link function \(f^2\):

\[
\mu_{ij}^0 = f(\eta_{ij})
\]

The output \(y\) is then chosen from a probability density \(P\), with mean \(\mu_{ij}^0\) and "dispersion" parameter \(\phi_{ij}^0\). We denote the density of \(y\) as

\[
P(y|x, \theta_{ij}^0)
\]

where the parameter vector \(\theta_{ij}^0\) includes the weights \(\beta_i^j\) and the dispersion parameter \(\phi_{ij}^0\):

\[
\theta_{ij}^0 = \begin{bmatrix}
\beta_i^j \\
\phi_{ij}^0
\end{bmatrix}
\]

We assume the density \(P\) to be a member of the exponential family of densities (McCullagh and Nelder 1983). The interpretation of the dispersion parameter depends on the particular choice of density. For example, in the case of the \(n\)-dimensional gaussian, the dispersion parameter is the covariance matrix \(\Sigma_{ij}^0\).

Given these assumptions, the total probability of generating \(y\) from \(x\) is the mixture of the probabilities of generating \(y\) from each of the component densities, where the mixing proportions are multinomial probabilities:

\[
P(y|x, \theta^0) = \sum_i g_i(x, \psi_i^0) \sum_j g_{ij}(x, \psi_{ij}^0) P(y|x, \theta_{ij}^0)
\]

Note that \(\theta^0\) includes the expert network parameters \(\theta_{ij}^0\) as well as the gating network parameters \(\psi_i^0\) and \(\psi_{ij}^0\). Note also that we have explicitly

---

2We utilize the neural network convention in defining links. In GLIM theory, the convention is that the link function relates \(y\) to \(\mu\), thus, \(y = h(\mu)\), where \(h\) is equivalent to our \(f\).

3Not all exponential family densities have a dispersion parameter; in particular, the Bernoulli density discussed below has no dispersion parameter.
indicated the dependence of the probabilities $g_i$ and $g_{ij}$ on the input $x$ and on the parameters. In the remainder of the paper we drop the explicit reference to the input and the parameters to simplify the notation:

$$P(y|x, \theta^0) = \sum_i g_i^0 \sum_j g_{ij}^0 P_j(y)$$  \hspace{1cm} (2.9)

We also utilize equation 2.9 without the superscripts to refer to the probability model defined by a particular HME architecture, irrespective of any reference to a "true" model.

2.2.1 Example (Regression). In the case of regression the probabilistic component of the model is generally assumed to be gaussian. Assuming identical covariance matrices of the form $\sigma^2 I$ for each of the experts yields the following hierarchical probability model:

$$P(y|x, \theta) = \frac{1}{(2\pi)^{n/2} \sigma^n} \sum_i g_i \sum_j g_{ij} e^{-(1/2\sigma^2)(y - \mu_j)^2}$$

2.2.2 Example (Binary Classification). In binary classification problems the output $y$ is a discrete random variable having possible outcomes of "failure" and "success." The probabilistic component of the model is generally assumed to be the Bernoulli distribution (Cox 1970). In this case, the mean $\mu_j$ is the conditional probability of classifying the input as "success." The resulting hierarchical probability model is a mixture of Bernoulli densities:

$$P(y|x, \theta) = \sum_i g_i \sum_j g_{ij} (1 - \mu_j)^{1-y} \mu_j^y$$

2.3 Posterior Probabilities. In developing the learning algorithms to be presented in the remainder of the paper, it will prove useful to define posterior probabilities associated with the nodes of the tree. The terms "posterior" and "prior" have meaning in this context during the training of the system. We refer to the probabilities $g_i$ and $g_{ij}$ as prior probabilities, because they are computed based only on the input $x$, without knowledge of the corresponding target output $y$. A posterior probability is defined once both the input and the target output are known. Using Bayes' rule, we define the posterior probabilities at the nodes of the tree as follows:

$$h_i = \frac{g_i \sum_j g_{ij} P_j(y)}{\sum_i g_i \sum_j g_{ij} P_j(y)}$$  \hspace{1cm} (2.10)

and

$$h_{ij} = \frac{g_{ij} P_j(y)}{\sum_i g_{ij} P_j(y)}$$  \hspace{1cm} (2.11)

We will also find it useful to define the joint posterior probability $h_{ij}$, the product of $h_i$ and $h_{ij}$:

$$h_{ij} = \frac{g_{ij} P_j(y)}{\sum_i g_i \sum_j g_{ij} P_j(y)}$$  \hspace{1cm} (2.12)

This quantity is the probability that expert network $(i,j)$ can be considered to have generated the data, based on knowledge of both the input and the output. Once again, we emphasize that all of these quantities are conditional on the input $x$.

In deeper trees, the posterior probability associated with an expert network is simply the product of the conditional posterior probability along the path from the root of the tree to that expert.

2.4 The Likelihood and a Gradient Ascent Learning Algorithm. Jordan and Jacobs (1992) presented a gradient ascent learning algorithm for the hierarchical architecture. The algorithm was based on earlier work by Jacobs et al. (1991), who treated the problem of learning in mixture-of-experts architectures as a maximum likelihood estimation problem. The log likelihood of a data set $X = \{(x^{(i)}, y^{(i)})\}^N_i$ is obtained by taking the log of the product of $N$ densities of the form of equation 2.9, which yields the following log likelihood:

$$l(\theta; X) = \sum_i \ln \sum_j g_{ij} \sum_j g_{ij} P_j(y^{(i)})$$  \hspace{1cm} (2.13)

Let us assume that the probability density $P$ is gaussian with an identity covariance matrix and that the link function is the identity. In this case, by differentiating $l(\theta; X)$ with respect to the parameters, we obtain the following gradient ascent learning rule for the weight matrix $U_j$:

$$\Delta U_{ij} = \rho \sum_i h_{ij} P_j(y^{(i)} - \mu^{(i)}/x^{(i)})$$  \hspace{1cm} (2.14)

where $\rho$ is a learning rate. The gradient ascent learning rule for the $i$th weight vector in the top-level gating network is given by

$$\Delta \Theta_i = \rho \sum_i (h_i - \delta_i) x^{(i)}$$  \hspace{1cm} (2.15)

and the gradient ascent rule for the $j$th weight vector in the $i$th lower-level gating network is given by

$$\Delta \Theta_{ij} = \rho \sum_i (h_{ij} - \delta_{ij}) x^{(i)}$$  \hspace{1cm} (2.16)

Updates can also be obtained for covariance matrices (Jordan and Jacobs 1992).

The algorithm given by equations 2.14, 2.15, and 2.16 is a batch learning algorithm. The corresponding on-line algorithm is obtained by sim-
ply dropping the summation sign and updating the parameters after each stimulus presentation. Thus, for example,
\[ U_{ij}^{(t+1)} = U_{ij}^{(t)} + \rho U_{ij}^{(t)} (y_i^{(t)} - \mu_i^{(t)}) x_j^{(t)} \quad \text{for} \quad i, j \text{th} \text{ expert network} \text{ based on the } t \text{th} \text{ stimulus pattern.} \]

(2.17)

is the stochastic update rule for the weights in the \((i, j)\)th expert network on the \(t\)th stimulus pattern.

2.5 The EM Algorithm. In the following sections we develop a learning algorithm for the HME architecture based on the Expectation-Maximization (EM) framework of Dempster et al. (1977). We derive an EM algorithm for the architecture that consists of the iterative solution of a coupled set of iteratively-reweighted least-squares problems.

The EM algorithm is a general technique for maximum likelihood estimation. In practice EM has been applied almost exclusively to unsupervised learning problems. This is true of the neural network literature and machine learning literature, in which EM has appeared in the context of clustering (Cheeseman et al. 1988; Nowlan 1991) and density estimation (Specht 1991), as well as the statistics literature, in which applications include missing data problems (Little and Rubin 1987), mixture density estimation (Redner and Walker 1984), and factor analysis (Dempster et al. 1977). Another unsupervised learning application is the learning problem for Hidden Markov Models, in which the Baum-Welch reestimation formulas are a special case of EM. There is nothing in the EM framework that precludes its application to regression or classification problems; however, such applications have been few.4

EM is an iterative approach to maximum likelihood estimation. Each iteration of an EM algorithm is composed of two steps: an Estimation (E) step and a Maximization (M) step. The M step involves the maximization of a likelihood function that is redefined in each iteration by the E step. If the algorithm simply increases the function during the M step, rather than maximizing the function, then the algorithm is referred to as a Generalized EM (GEM) algorithm. The Boltzmann learning algorithm (Hinton and Sejnowski 1986) is a neural network example of a GEM algorithm. GEM algorithms are often significantly slower to converge than EM algorithms.

An application of EM generally begins with the observation that the optimization of the likelihood function \(l(\theta; X)\) would be simplified if only a set of additional variables, called “missing” or “hidden” variables, were known. In this context, we refer to the observable data \(X\) as the “incomplete data” and posit a “complete data” set \(Y\) that includes the missing variables \(Z\). We specify a probability model that links the fictive missing variables to the actual data: \(P(y, z|x, \theta)\). The logarithm of the density \(P\) defines the “complete-data likelihood,” \(l(\theta; Y)\). The original likelihood,

\[ l(\theta; X), \text{ is referred to in this context as the “incomplete-data likelihood.”} \]

It is the relationship between these two likelihood functions that motivates the EM algorithm. Note that the complete-data likelihood is a random variable, because the missing variables \(Z\) are in fact unknown. An EM algorithm first finds the expected value of the complete-data likelihood, given the observed data and the current model. This is the E step:

\[ Q(\theta, \theta^{(p)}) = E[l(\theta; Y)|X] \]

where \(\theta^{(p)}\) is the value of the parameters at the \(p\)th iteration and the expectation is taken with respect to \(\theta^{(p)}\). This step yields a deterministic function \(Q\). The M step maximizes this function with respect to \(\theta\) to find the new parameter estimates \(\theta^{(p+1)}\):

\[ \theta^{(p+1)} = \text{arg max} \ Q(\theta, \theta^{(p)}) \]

The E step is then repeated to yield an improved estimate of the complete likelihood and the process iterates.

An iterative step of EM chooses a parameter value that increases the value of \(Q\), the expectation of the complete likelihood. What is the effect of such a step on the incomplete likelihood? Dempster et al. proved that an increase in \(Q\) implies an increase in the incomplete likelihood:

\[ l(\theta^{(p+1)}; X) \geq l(\theta^{(p)}; X) \]

Equality obtains only at the stationary points of \(l\) (Wu 1983). Thus the likelihood \(l\) increases monotonically along the sequence of parameter estimates generated by an EM algorithm. In practice this implies convergence to a local maximum.

2.6 Applying EM to the HME Architecture. To develop an EM algorithm for the HME architecture, we must define appropriate “missing data” so as to simplify the likelihood function. We define indicator variables \(z_i\) and \(z_{ij}\), such that one and only one of the \(z_i\) is equal to one, and one and only one of the \(z_{ij}\) is equal to one. These indicator variables have an interpretation as the labels that correspond to the decisions in the probability model. We also define the indicator variable \(z_q\), which is the product of \(z_i\) and \(z_{ij}\). This variable has an interpretation as the label that specifies the expert (the regressive process) in the probability model. If the labels \(z_i, z_{ij}\), and \(z_q\) were known, then the maximum likelihood problem would decouple into a separate set of regression problems for each expert network and a separate set of multiway classification problems for the gating networks. These problems would be solved independently of each other, yielding a rapid one-pass learning algorithm. Of course, the missing variables are not known, but we can specify a probability model

4An exception is the “switching regression” model of Quandt and Ramsey (1972). For further discussion of switching regression, see Jordan and Xu (1993).
that links them to the observable data. This probability model can be written in terms of the $z_{ij}$ as follows:

$$P(y^{(i)}, z_{ij}^{(i)} | x^{(i)}, \theta) = \frac{g_{i}^{(i)} y_{j}^{(i)} P_{ij}(y^{(i)})}{\prod_{i} \prod_{j} (g_{i}^{(i)} y_{j}^{(i)} P_{ij}(y^{(i)}))^{z_{ij}^{(i)}}}$$

(2.18)

using the fact that $z_{ij}^{(i)}$ is an indicator variable. Taking the logarithm of this probability model yields the following complete-data likelihood:

$$l_{i}(\theta, Y) = \sum_{i} \sum_{j} z_{ij}^{(i)} \ln \left( \frac{g_{i}^{(i)} y_{j}^{(i)} P_{ij}(y^{(i)})}{\prod_{i} \prod_{j} (g_{i}^{(i)} y_{j}^{(i)} P_{ij}(y^{(i)}))^{z_{ij}^{(i)}}} \right)$$

(2.20)

$$= \sum_{i} \sum_{j} z_{ij}^{(i)} \left( \ln g_{i}^{(i)} + \ln y_{j}^{(i)} + \ln P_{ij}(y^{(i)}) \right)$$

(2.21)

Note the relationship of the complete-data likelihood in equation 2.21 to the incomplete-data likelihood in equation 2.13. The use of the indicator variables $z_{ij}$ has allowed the logarithm to be brought inside the summation signs, substantially simplifying the maximization problem. We now define the E step of the EM algorithm by taking the expectation of the complete-data likelihood:

$$Q(\theta, \theta^{(p)}) = \sum_{i} \sum_{j} h_{ij}^{(i)} \left( \ln s_{i}^{(i)} + \ln y_{j}^{(i)} + \ln P_{ij}(y^{(i)}) \right)$$

(2.22)

where we have used the fact that

$$E[z_{ij}^{(i)} | x] = P(z_{ij}^{(i)} = 1 | y^{(i)}, x^{(i)}, \theta^{(p)})$$

(2.23)

$$= \frac{P(y^{(i)} | x^{(i)}, \theta^{(p)}) P(z_{ij}^{(i)} = 1 | x^{(i)}, \theta^{(p)})}{P(y^{(i)} | x^{(i)}, \theta^{(p)})}$$

(2.24)

$$= \frac{\sum_{j} h_{ij}^{(i)} \ln P_{ij}(y^{(i)})}{\sum_{j} h_{ij}^{(i)} s_{i}^{(i)}}$$

(2.25)

$$= h_{ij}^{(i)}$$

(2.26)

(Note also that $E[z_{ij}^{(i)} | x] = h_{ij}^{(i)}$ and $E[z_{ij}^{(i)} | x] = h_{ij}^{(i)}$.

The M step requires maximizing $Q(\theta, \theta^{(p)})$ with respect to the expert network parameters and the gating network parameters. Examining equation 2.22, we see that the expert network parameters influence the $Q$ function only through the terms $h_{ij}^{(i)} \ln P_{ij}(y^{(i)})$, and the gating network parameters influence the $Q$ function only through the terms $h_{ij}^{(i)} \ln s_{i}^{(i)}$ and $h_{ij}^{(i)} \ln y_{j}^{(i)}$. Thus the M step reduces to the following separate maximization problems:

$$\theta_{y}^{(p+1)} = \arg \max_{\theta_{y}} \sum_{i} h_{ij}^{(i)} \ln P_{ij}(y^{(i)})$$

(2.27)

$$\nu_{y}^{(p+1)} = \arg \max_{\nu_{y}} \sum_{i} \sum_{k} h_{ik}^{(i)} \ln s_{k}^{(i)}$$

(2.28)

and

$$v_{y}^{(p+1)} = \arg \max_{v_{y}} \sum_{i} \sum_{j} h_{ij}^{(i)} \sum_{i} h_{ij}^{(i)} \ln s_{j}^{(i)}$$

(2.29)

Each of these maximization problems is itself a maximum likelihood problem. This is clearly true in the case of equation 2.27, which is simply a weighted maximum likelihood problem in the probability density $P_{ij}$.

Given our parameterization of $P_{ij}$, the log likelihood in equation 2.27 is a weighted log likelihood for a GLIM. An efficient algorithm known as iteratively reweighted least-squares (IRLS) is available to solve the maximum likelihood problem for such models (McCullagh and Nelder 1983). We discuss IRLS in Appendix A.

Equation 2.28 involves maximizing the cross-entropy between the posterior probabilities $h_{i}^{(i)}$ and the prior probabilities $s_{i}^{(i)}$. This cross-entropy is the log likelihood associated with a multinomial logit probability model in which the $h_{i}^{(i)}$ act as the output observations (see Appendix B). Thus the maximization in equation 2.28 is also a maximum likelihood problem for a GLIM and can be solved using IRLS. The same is true of equation 2.29, which is a weighted maximum likelihood problem with output observations $h_{ij}^{(i)}$ and observation weights $h_{ij}^{(i)}$.

In summary, the EM algorithm that we have obtained involves a calculation of posterior probabilities in the outer loop (the E step), and the solution of a set of IRLS problems in the inner loop (the M step). We summarize the algorithm as follows:

**Algorithm 1**

1. For each data pair $(x^{(i)}, y^{(i)})$, compute the posterior probabilities $h_{ij}^{(i)}$ and $h_{ij}^{(i)}$ using the current values of the parameters.

2. For each expert $(i, j)$, solve an IRLS problem with observations $\{(x^{(i)}, y^{(i)})\}^{N}_{i}$ and observation weights $\{h_{ij}^{(i)}\}^{N}_{i}$.

3. For each top-level gating network, solve an IRLS problem with observations $\{(x^{(i)}, h_{ij}^{(i)})\}^{N}_{i}$.

4. For each lower-level gating network, solve a weighted IRLS problem with observations $\{(x^{(i)}, h_{ij}^{(i)})\}^{N}_{i}$ and observation weights $\{h_{ij}^{(i)}\}^{N}_{i}$.

5. Iterate using the updated parameter values.

**2.7 A Least-Squares Algorithm**. In the case of regression, in which a gaussian probability model and an identity link function are used, the IRLS loop for the expert networks reduces to weighted least squares, which can be solved (in one pass) by any of the standard least-squares algorithms (Golub and van Loan 1989). The gating networks still require iterative processing. Suppose, however, that we fit the parameters of the
gating networks using least squares rather than maximum likelihood. In this case, we might hope to obtain an algorithm in which the gating network parameters are fit by a one-pass algorithm. To motivate this approach, note that we can express the IRLS problem for the gating networks as follows. Differentiating the cross-entropy (equation 2.28) with respect to the parameters \( \mathbf{v}_i \) (using the fact that \( \frac{\partial g_i}{\partial \delta} = g_i(b_{ij} - \hat{g}_i) \), where \( b_{ij} \) is the Kronecker delta) and setting the derivatives to zero yields the following equations:

\[
\sum_i (h_{ij}^{(i)} - g_i(x^{(i)}, \mathbf{v}_i)) x^{(i)} = 0 \tag{2.30}
\]

which are a coupled set of equations that must be solved for each \( i \). Similarly, for each gating network at the second level of the tree, we obtain the following equations:

\[
\sum_j h_{ij}^{(o)} (h_{ij}^{(o)} - g_{ij}(x^{(i)}, \mathbf{v}_j)) x^{(i)} = 0 \tag{2.31}
\]

which must be solved for each \( i \) and \( j \). There is one aspect of these equations that renders them unusual. Recall that if the labels \( z_i^{(o)} \) and \( z_j^{(o)} \) were known, then the gating networks would be essentially solving a set of multiway classification problems. The supervised errors \( (z_i^{(o)} - g_i^{(o)}) \) and \( (z_j^{(o)} - g_{ij}^{(o)}) \) would appear in the algorithm for solving these problems. Note that these errors are differences between indicator variables and probabilities. In equations 2.30 and 2.31, on the other hand, the error that drives the algorithm are the differences \( (h_{ij}^{(o)} - g_i^{(o)}) \) and \( (h_{ij}^{(o)} - g_{ij}^{(o)}) \), which are differences between probabilities. The EM algorithm effectively "fills in" the missing labels with estimated probabilities \( h_i \) and \( h_{ij} \). These estimated probabilities can be thought of as targets for the \( g_i \) and the \( g_{ij} \). This suggests that we can compute "virtual targets" for the underlying linear predictors \( \xi_i \) and \( \xi_{ij} \) by inverting the softmax function. (Note that this option would not be available for the \( z_i \) and \( z_j^{(o)} \), even if they were known, because zero and one are not in the range of the softmax function.) Thus the targets for the \( \xi_i \) are the values:

\[
\ln h_{ij}^{(o)} - \ln C
\]

where \( C = \sum_k e^{\xi_i} \) is the normalization constant in the softmax function. Note, however, that constants that are common to all of the \( \xi_i \) can be omitted, because such constants disappear when \( \xi_i \) are converted to \( g_i \) (cf. equation 2.3). Thus the values \( \ln h_{ij}^{(o)} \) can be used as targets for the \( \xi_i \). A similar argument shows that the values \( \ln h_{ij}^{(o)} \) can be used as targets for the \( \xi_{ij} \), with observation weights \( h_{ij}^{(o)} \).

The utility of this approach is that once targets are available for the linear predictors \( \xi_i \) and \( \xi_{ij} \), the problem of finding the parameters \( \mathbf{v}_i \) and \( \mathbf{v}_j \) reduces to a coupled set of weighted least-squares problems. Thus we obtain an algorithm in which all of the parameters in the hierarchy, both in the expert networks and the gating networks, can be obtained by solving least-squares problems. This yields the following learning algorithm:

**Algorithm 2**

1. For each data pair \( (x^{(i)}, y^{(i)}) \), compute the posterior probabilities \( h_{ij}^{(o)} \) and \( h_{ij}^{(o)} \) using the current values of the parameters.
2. For each expert \( (i,j) \), solve a weighted least-squares problem with observations \( \{(x^{(i)}, y^{(i)})\}^N \) and observation weights \( \{h_{ij}^{(o)}\}^N \).
3. For each top-level gating network, solve a least-squares problem with observations \( \{(x^{(i)}, \ln h_{ij}^{(o)})\}^N \).
4. For each lower-level gating network, solve a weighted least-squares problem with observations \( \{(x^{(i)}, \ln h_{ij}^{(o)})\}^N \) and observation weights \( \{h_{ij}^{(o)}\}^N \).
5. Iterate using the updated parameter values.

It is important to note that this algorithm does not yield the same parameter estimates as Algorithm 1; the gating network residuals \( (h_{ij}^{(o)} - g_{ij}^{(o)}) \) are being fit by least squares rather than maximum likelihood. The algorithm can be thought of as an approximation to Algorithm 1, an approximation based on the assumption that the differences between \( h_{ij}^{(o)} \) and \( g_{ij}^{(o)} \) are small. This assumption is equivalent to the assumption that the architecture can fit the underlying regression surface (a consistency condition) and the assumption that the noise is small. In practice we have found that the least-squares algorithm works reasonably well, even in the early stages of fitting when the residuals can be large. The ability to use least squares is certainly appealing from a computational point of view. One possible hybrid algorithm involves using the least-squares algorithm to converge quickly to the neighborhood of a solution and then using IRLS to refine the solution.

**2.8 Simulation Results.** We tested Algorithm 1 and Algorithm 2 on a nonlinear system identification problem. The data were obtained from a simulation of a four-joint robot arm moving in three-dimensional space (Fun and Jordan 1993). The network must learn the forward dynamics of the arm; a state-dependent mapping from joint torques to joint accelerations. The state of the arm is encoded by eight real-valued variables: four positions (rad) and four angular velocities (rad/sec). The torque was encoded as four real-valued variables (N·m). Thus there were 12 inputs to the learning system. Given these 12 input variables, the network must predict the four accelerations at the joints (rad/sec²). This
mapping is highly nonlinear due to the rotating coordinate systems and the interaction torques between the links of the arm.

We generated 15,000 data points for training and 5,000 points for testing. For each epoch (i.e., each pass through the training set), we computed the relative error on the test set. Relative error is computed as a ratio between the mean squared error and the mean squared error that would be obtained if the learner were to output the mean value of the accelerations for all data points.

We compared the performance of a binary hierarchy to that of a backpropagation network. The hierarchy was a four-level hierarchy with 16 expert networks and 15 gating networks. Each expert network had 4 output units and each gating network had 1 output unit. The backpropagation network had 60 hidden units, which yields approximately the same number of parameters in the network as in the hierarchy.

The HME architecture was trained by Algorithms 1 and 2, utilizing Cholesky decomposition to solve the weighted least-squares problems (Golub and van Loan 1989). Note that the HME algorithms have no free parameters. The free parameters for the backpropagation network (the learning rate and the momentum term) were chosen based on a coarse search of the parameter space. (Values of 0.00001 and 0.15 were chosen for these parameters.) There were difficulties with local minima (or plateaus) using the backpropagation algorithm: Five of 10 runs failed to converge to “reasonable” error values. (As we report in the next section, no such difficulties were encountered in the case of on-line backpropagation.) We report average convergence times and average relative errors only for those runs that converged to “reasonable” error values. All 10 runs for both of the HME algorithms converged to “reasonable” error values.

Figure 2 shows the performance of the hierarchy and the backpropagation network. The horizontal axis of the graph gives the training time in epochs. The vertical axis gives generalization performance as measured by the average relative error on the test set.

Table 1 reports the average relative errors for both architectures measured at the minima of the relative error curves. (Minima were defined by a sequence of three successive increases in the relative error.) We also report values of relative error for the best linear approximation, the CART algorithm, and the MARS algorithm. Both CART and MARS were run four times, once for each of the output variables. We combined the results from these four computations to compute the total relative error. Two versions of CART were run; one in which the splits were restricted to be parallel to the axes and one in which linear combinations of the input variables were allowed.

The MARS algorithm requires choices to be made for the values of two structural parameters: the maximum number of basis functions and the maximum number of interaction terms. Each basis function in MARS yields a linear surface defined over a rectangular region of the input space, corresponding roughly to the function implemented by a single expert in the HME architecture. Therefore we chose a maximum of 16 basis functions to correspond to the 16 experts in the four-level hierarchy. To choose the maximum number of interactions \((m_i)\), we compared the performance of MARS for \(m_i = 1, 2, 3, 6, \text{ and } 12\), and chose the value that yielded the best performance \((m_i = 3)\).

For the iterative algorithms, we also report the number of epochs required for convergence. Because the learning curves for these algorithms
generally have lengthy tails, we defined convergence as the first epoch at which the relative error drops within 5% of the minimum.

All of the architectures that we studied performed significantly better than the best linear approximation. As expected, the CART architecture with linear combinations performed better than CART with axis-parallel splits.\(^5\) The HME architecture yielded a modest improvement over MARS and CART. Backpropagation produced the lowest relative error of the algorithms tested (ignoring the difficulties with convergence).

These differences in relative error should be treated with some caution. The need to set tree parameters for some of the architectures (e.g., backpropagation) and the need to make structural choices (e.g., number of hidden units, number of basis functions, number of experts) make it difficult to match architectures. The HME architecture, for example, involves parameter dependencies that are not present in a backpropagation network. A gating network at a high level in the tree can “pinch off” a branch of the tree, rendering useless the parameters in that branch of the tree. Raw parameter count is therefore only a very rough guide to architecture capacity; more precise measures are needed (e.g., VC dimension) before definitive quantitative comparisons can be made.

The differences between backpropagation and HME in terms of convergence time are more definitive. Both HME algorithms reliably converge more than two orders of magnitude faster than backpropagation.

As shown in Figure 3, the HME architecture lends itself well to graphic investigation. This figure displays the time sequence of the distributions of posterior probabilities across the training set at each node of the tree. At Epoch 0, before any learning has taken place, most of the posterior probabilities at each node are approximately 0.5 across the training set. As the training proceeds, the histograms flatten out, eventually approaching bimodal distributions in which the posterior probabilities are either one or zero for most of the training patterns. This evolution is indicative of increasingly sharp splits being fit by the gating networks. Note that there is a tendency for the splits to be formed more rapidly at higher levels in the tree than at lower levels.

Figure 4 shows another graphic device that can be useful for understanding the way in which an HME architecture fits a data set. This figure, which we refer to as a “deviance tree,” shows the deviance (mean squared error) that would be obtained at each level of the tree if the tree were clipped at that level. We construct a clipped tree at a given level by replacing each nonterminal at that level with a matrix that is a weighted average of the experts below that nonterminal. The weights are the total prior probabilities associated with each expert across the training set. The error for each output unit is then calculated by passing the test set through the clipped tree. As can be seen in the figure, the deviance is substantially smaller for deeper trees (note that the ordinate of the plots is on a log scale). The deviance in the right branch of the tree is larger than in the left branch of the tree. Information such as this can be useful for purposes of exploratory data analysis and for model selection.

2.9 An On-Line Algorithm. The batch least-squares algorithm that we have described (Algorithm 2) can be converted into an on-line algorithm by noting that linear least squares and weighted linear least squares problems can be solved by recursive procedures that update the parameter estimates with each successive data point (Ljung and Söderström 1986). Our application of these recursive algorithms is straightforward; however, care must be taken to handle the observation weights (the posterior probabilities) correctly. These weights change as a function of the changing parameter values. This implies that the recursive least squares algorithm must include a decay parameter that allows the system to “forget” older values of the posterior probabilities.
Mixtures of Experts and EM Algorithm

where $R_{ij}$ is the inverse covariance matrix for expert network $(i,j)$. This matrix is updated via the equation:

$$R_{ij}^{(t)} = \lambda^{-1} R_{ij}^{(t-1)} - \lambda^{-1} \frac{R_{ij}^{(t-1)} x^{(t)} x^{(t)T} R_{ij}^{(t-1)}}{\lambda [h_{ij}^{(t)}]^{-1} + x^{(t)T} R_{ij}^{(t-1)} x^{(t)}}$$  \hspace{1cm} (2.33)

where $\lambda$ is the decay parameter.

It is interesting to note the similarity between the parameter update rule in equation 2.32 and the gradient rule presented earlier (cf. equation 2.14). These updates are essentially the same, except that the scalar $\rho$ is replaced by the matrix $R_{ij}^{(t)}$. It can be shown, however, that $R_{ij}^{(t)}$ is an estimate of the inverse Hessian of the least-squares cost function (Ljung and Söderström 1986), thus equation 2.32 is in fact a stochastic approximation to a Newton–Raphson method rather than a gradient method.\footnote{This is true for fixed values of the posterior probabilities. These posterior probabilities are also changing over time, however, as required by the EM algorithm. The overall convergence rate of the algorithm is determined by the convergence rate of EM, not the convergence rate of Newton–Raphson.}

Similar equations apply for the updates of the gating networks. The update rule for the parameters of the top-level gating network is given by the following equation (for the $i$th output of the gating network):

$$v_i^{(t+1)} = v_i^{(t)} + s_i^{(t)} (ln h_i^{(t)} - \xi_i^{(t)}) x_i^{(t)}$$  \hspace{1cm} (2.34)

where the inverse covariance matrix $S_i$ is updated by

$$s_i^{(t)} = \lambda^{-1} s_i^{(t-1)} - \lambda^{-1} \frac{S_i^{(t-1)} x_i^{(t)} x_i^{(t)T} S_i^{(t-1)}}{\lambda + x_i^{(t)T} S_i^{(t-1)} x_i^{(t)}}$$  \hspace{1cm} (2.35)

Finally, the update rule for the parameters of the lower-level gating network is as follows:

$$v_j^{(t+1)} = v_j^{(t)} + s_j^{(t)} (ln h_j^{(t)} - \xi_j^{(t)}) x_j^{(t)}$$  \hspace{1cm} (2.36)

where the inverse covariance matrix $S_j$ is updated by

$$s_j^{(t)} = \lambda^{-1} s_j^{(t-1)} - \lambda^{-1} \frac{S_j^{(t-1)} x_j^{(t)} x_j^{(t)T} S_j^{(t-1)}}{\lambda [h_j^{(t)}]^{-1} + x_j^{(t)T} S_j^{(t-1)} x_j^{(t)}}$$  \hspace{1cm} (2.37)

\subsection*{2.10 Simulation Results.} The on-line algorithm was tested on the robot dynamics problem described in the previous section. Preliminary simulations convinced us of the necessity of the decay parameter (\(\lambda\)). We also found that this parameter should be slowly increased as training proceeds—on the early trials the posterior probabilities are changing rapidly so that the covariances should be decayed rapidly, whereas on later trials the posterior probabilities have stabilized and the covariances should be decayed less rapidly. We used a simple fixed schedule: \(\lambda\) was...
initialized to 0.99 and increased a fixed fraction (0.6) of the remaining distance to 1.0 every 1000 time steps.

The performance of the on-line algorithm was compared to an on-line backpropagation network. Parameter settings for the backpropagation network were obtained by a coarse search through the parameter space, yielding a value of 0.15 for the learning rate and 0.20 for the momentum. The results for both architectures are shown in Figure 5. As can be seen, the on-line algorithm for backpropagation is significantly faster than the corresponding batch algorithm (cf. Fig. 2). This is also true of the on-line HME algorithm, which has nearly converged within the first epoch.

The minimum values of relative error and the convergence times for both architectures are provided in Table 2. We also provide the corresponding values for a simulation of the on-line gradient algorithm for the HME architecture (equation 2.17).

We also performed a set of simulations which tested the variety of different HME architectures. We compared a one-level hierarchy with 32 experts to hierarchies with five levels (32 experts), and six levels (64 experts). We also simulated two three-level hierarchies, one with branching factors of 4, 4, and 2 (proceeding from the top of the tree to the bottom), and one with branching factors of 2, 4, and 4. (Each three-level hierarchy contained 32 experts.) The results are shown in Figure 6. As can be seen, there was a significant difference between the one-level hierarchy and the other architectures. There were smaller differences among the multilevel hierarchies. No significant difference was observed between the two different 3-level architectures.

### 3 Model Selection

Utilizing the HME approach requires that choices be made regarding the structural parameters of the model, in particular the number of levels;
and the branching factor of the tree. As with other flexible estimation techniques, it is desirable to allow these structural parameters to be chosen based at least partly on the data. This model selection problem can be addressed in a variety of ways. In this paper we have utilized a test set approach to model selection, stopping the training when the error on the test set reaches a minimum. As is the case with other neural network algorithms, this procedure can be justified as a complexity control measure. As we have noted, when the parameters in the gating networks of an HME architecture are small, the entire system reduces to a single “averaged” GLIM at the root of the tree. As the training proceeds, the parameters in the gating networks begin to grow in magnitude and splits are formed. When a split is formed the parameters in the branches of the tree on either side of the split are decoupled and the effective number of degrees of freedom in the system increases. This increase in complexity takes place gradually as the values of the parameters increase and the splits sharpen. By stopping the training of the system based on the performance on a test set, we obtain control over the effective number of degrees of freedom in the architecture.

Other approaches to model selection can also be considered. One natural approach is to use ridge regression in each of the expert networks and the gating networks. This approach extends naturally to the on-line setting in the form of a “weight decay.” It is also worth considering Bayesian techniques of the kind considered in the decision tree literature by Buntine (1991), as well as the MDL methods of Quinlan and Rivest (1989).

4 Related Work

There are a variety of ties that can be made between the HME architecture and related work in statistics, machine learning, and neural networks. In this section we briefly mention some of these ties and make some comparative remarks.

Our architecture is not the only nonlinear approximator to make substantial use of GLIMs and the IRLS algorithm. IRLS also figures prominently in a branch of nonparametric statistics known as generalized additive models (GAMs; Hastie and Tibshirani 1990). It is interesting to note the complementary roles of IRLS in these two architectures. In the GAM model, the IRLS algorithm appears in the outer loop, providing an adjusted dependent variable that is fit by a backfitting procedure in the inner loop. In the HME approach, on the other hand, the outer loop is the E step of EM and IRLS is in the inner loop. This complementarity suggests that it might be of interest to consider hybrid models in which an HME is nested inside a GAM or vice versa.

We have already mentioned the close ties between the HME approach and other tree-structured estimators such as CART and MARS. Our approach differs from MARS and related architectures—such as the basis-function trees of Sanger (1991)—by allowing splits that are oblique with respect to the axes. We also differ from these architectures by using a statistical model—the multinomial logit model—for the splits. We believe that both of these features can play a role in increasing predictive ability—the use of oblique splits should tend to decrease bias, and the use of smooth multinomial logit splits should generally decrease variance. Oblique splits also render the HME architecture insensitive to the particular choice of coordinates used to encode the data. Finally, it is worth emphasizing the difference in philosophy behind these architectures. Whereas CART and MARS are entirely nonparametric, the HME approach has a strong flavor of parametric statistics, via its use of generalized linear models, mixture models, and maximum likelihood.

Similar comments can be made with respect to the decision tree methodology in the machine learning literature. Algorithms such as ID3 build trees that have axis-parallel splits and use heuristic splitting algorithms (Quinlan 1986). More recent research has studied decision trees with oblique splits (Murthy et al. 1993; Utgoff and Brodley 1990). None of these papers, however, has treated the problem of splitting data as a statistical problem, nor have they provided a global goodness-of-fit measure for their trees.

There are a variety of neural network architectures that are related to the HME architecture. The multiresolution aspect of HME is reminiscent of Moody's (1989) multiresolution CMAC hierarchy, differing in that Moody's levels of resolution are handled explicitly by separate networks. The “neural tree” algorithm (Strömberg et al. 1991) is a decision tree with multilayer perceptions (MLPs) at the nonterminals. This architecture can form oblique (or curvilinear) splits, however, the MLPs are trained by a heuristic that has no clear relationship to overall classification performance. Finally, Hinton and Nowlan (see Nowlan 1991) have independently proposed extending the Jacobs et al. (1991) modular architecture to a tree-structured system. They did not develop a likelihood approach to the problem, however, proposing instead a heuristic splitting scheme.

5 Conclusions

We have presented a tree-structured architecture for supervised learning. We have developed the learning algorithm for this architecture within the framework of maximum likelihood estimation, utilizing ideas from mixture model estimation and generalized linear model theory. The maximum likelihood framework allows standard tools from statistical theory to be brought to bear in developing inference procedures and measures of uncertainty for the architecture (Cox and Hinkley 1974). It also opens the door to the Bayesian approaches that have been found to be useful
in the context of unsupervised mixture model estimation (Cheeseman et al. 1988).

Although we have not emphasized theoretical issues in this paper, there are a number of points that are worth mentioning. First, the set of exponentially smoothed piecewise linear functions that we have utilized is clearly dense in the set of piecewise linear functions on compact sets in $\mathbb{R}^n$, thus it is straightforward to show that the hierarchical architecture is dense in the set of continuous functions on compact sets in $\mathbb{R}^n$. That is, the architecture is “universal” in the sense of Hornik et al. (1989). From this result it would seem straightforward to develop consistency results for the architecture (cf. Geman et al. 1992; Stone 1977). We are currently developing this line of argument and are studying the asymptotic distributional properties of fixed hierarchies. Second, convergence results are available for the architecture. We have shown that the convergence rate of the algorithm is linear in the condition number of a matrix that is the product of an inverse covariance matrix and the Hessian of the log likelihood for the architecture (Jordan and Xu 1993).

Finally, it is worth noting a number of possible extensions of the work reported here. Our earlier work on hierarchical mixtures of experts utilized the multilayer perceptron as the primitive function for the expert networks and gating networks (Jordan and Jacobs 1992). That option is still available, although we lose the EM proof of convergence (cf. Jordan and Xu 1993) and we lose the ability to fit the subnetworks efficiently with IRLS. One interesting example of such an application is the case where the experts are autoassociators (Bouland and Kamp 1988), in which case the architecture fits hierarchically nested local principal component decompositions. Another area in unsupervised learning worth exploring is the nonassociative version of the hierarchical architecture. Such a model would be a recursive version of classical mixture-likelihood clustering and may have interesting ties to hierarchical clustering models. Finally, it is also of interest to note that the recursive least squares algorithm that we utilized in obtaining an on-line variant of Algorithm 2 is not the only possible on-line approach. Any of the fast filter algorithms (Haykin 1991) could also be utilized, giving rise to a family of on-line algorithms. Also, it is worth studying the application of the recursive algorithms to PRESS-like cross-validation calculations to efficiently compute the changes in likelihood that arise from adding or deleting parameters or data points.

Appendix A: Iteratively Reweighted Least Squares

The iteratively reweighted least squares (IRLS) algorithm is the inner loop of the algorithm that we have proposed for the HME architecture. In this section, we describe the IRLS algorithm, deriving it as a special case of the Fisher scoring method for generalized linear models. Our presentation derives from McCullagh and Nelder (1983).

IRLS is an iterative algorithm for computing the maximum likelihood estimates of the parameters of a generalized linear model. It is a special case of a general algorithm for maximum likelihood estimation known as the Fisher scoring method (Finney 1973). Let $l(\beta; x)$ be a log likelihood function—a function of the parameter vector $\beta$—and let $\frac{\partial l}{\partial \beta \beta'}$ denote the Hessian of the log likelihood. The Fisher scoring method updates the parameter estimates $\beta$ as follows:

$$
\beta_{r+1} = \beta_r - \left( E \left[ \frac{\partial l}{\partial \beta \beta'} \right] \right)^{-1} \frac{\partial l}{\partial \beta}
$$

(5.1)

where $\beta_r$ denotes the parameter estimate at the $r$th iteration and $\frac{\partial l}{\partial \beta}$ is the gradient vector. Note that the Fisher scoring method is essentially the same as the Newton–Raphson algorithm, except that the expected value of the Hessian replaces the Hessian. There are statistical reasons for preferring the expected value of the Hessian—and the expected value of the Hessian is often easier to compute—but Newton–Raphson can also be used in many cases.

The likelihood in generalized linear model theory is a product of densities from the exponential family of distributions. This family is an important class in statistics and includes many useful densities, such as the normal, the Poisson, the binomial, and the gamma. The general form of a density in the exponential family is the following:

$$
P(y; \eta, \phi) = \exp \{ (y \eta - b(\eta)) / \phi + c(y, \phi) \}
$$

(5.2)

where $\eta$ is known as the “natural parameter” and $\phi$ is the dispersion parameter.\footnote{We restrict ourselves to scalar-valued random variables to simplify the presentation, and describe the (straightforward) extension to vector-valued random variables at the end of the section.}

Example (Bernoulli Density). The Bernoulli density with mean $\pi$ has the following form:

$$
P(y; \pi) = \pi^y (1 - \pi)^{1-y}
= \exp \{ \ln(\pi^y (1 - \pi)^{1-y}) \}
= \exp \{ y \eta - \ln(1 + e^\eta) \}
$$

(5.3)

where $\eta = \ln(\pi/1 - \pi)$ is the natural parameter of the Bernoulli density. This parameter has the interpretation as the log odds of "success" in a random Bernoulli experiment.

In a generalized linear model, the parameter $\eta$ is modeled as a linear function of the input $x$:

$$
\eta = \beta^T x
$$
where $\beta$ is a parameter vector. Substituting this expression into equation 5.2 and taking the product of $N$ such densities yields the following log likelihood for a data set $X = \{(x^{(i)}, y^{(i)})\}_{i=1}^N$:

$$l(\beta, X) = \sum_i \left\{ (\beta^T x^{(i)}) y^{(i)} - b(\beta^T x^{(i)}) / \phi + c(y^{(i)}, \phi) \right\}$$

The observations $y^{(i)}$ are assumed to be sampled independently from densities $P(y, \eta, \phi)$, where $\eta^{(i)} = \beta^T x^{(i)}$.

We now compute the gradient of the log likelihood:

$$\frac{\partial l}{\partial \beta} = \sum_i (y^{(i)} - b(\beta^T x^{(i)})) x^{(i)} / \phi$$

and the Hessian of the log likelihood:

$$\frac{\partial^2 l}{\partial \beta \partial \beta^T} = -\sum_i b''(\beta^T x^{(i)}) x^{(i)} x^{(i)T} / \phi$$

These quantities could be substituted directly into equation 5.1, however, there is additional mathematical structure that can be exploited. First note the following identity, which is true of any log likelihood:

$$E \left[ \frac{\partial l}{\partial \beta} \right] = 0$$

(This fact can be proved by differentiating both sides of the identity $\int P(y, \beta, \phi) dy = 1$ with respect to $\beta$.) Because this identity is true for any set of observed data, including all subsets of $X$, we have the following:

$$E[y^{(i)}] = b'(\beta^T x^{(i)})$$

for all $i$. This equation implies that the mean of $y^{(i)}$, which we denote as $\mu^{(i)}$, is a function of $\eta^{(i)}$. We therefore include in the generalized linear model the link function, which models $\mu$ as a function of $\eta$:

$$\mu^{(i)} = f(\eta^{(i)})$$

**Example (Bernoulli Density).** Equation 5.3 shows that $b(\eta) = \ln(1 + e^\eta)$ for the Bernoulli density. Thus

$$\mu = b'(\eta) = \frac{e^\eta}{1 + e^\eta}$$

which is the logistic function. Inverting the logistic function yields $\eta = \ln(\mu / (1 - \mu))$; thus, $\mu$ equals $\pi$, as it must.

The link function $f(\eta) = b'(\eta)$ is known in generalized linear model theory as the canonical link. By parameterizing the exponential family density in terms of $\eta$ (cf. equation 5.2), we have forced the choice of the canonical link. It is also possible to use other links, in which case $\eta$

no longer has the interpretation as the natural parameter of the density. There are statistical reasons, however, to prefer the canonical link (McCullagh and Nelder 1983). Moreover, by choosing the canonical link, the Hessian of the likelihood turns out to be constant (cf. equation 5.5), and the Fisher scoring method therefore reduces to Newton–Raphson.

To continue the development, we need an additional fact about log likelihoods. By differentiating the identity $\int P(y, \beta, \phi) dy = 1$ twice with respect to $\beta$, the following identity can be established:

$$E \left[ \frac{\partial^2 l}{\partial \beta \partial \beta^T} \right] = -E \left[ \frac{\partial l}{\partial \beta} \right] \left[ \frac{\partial l}{\partial \beta} \right]^T$$

This identity can be used to obtain a relationship between the variance of $\eta$ and the function $b''(\eta)$ in the exponential family density. Beginning with equation 5.5, we have:

$$-E \left[ \sum_i b''(\beta^T x^{(i)}) x^{(i)} x^{(i)T} / \phi \right]$$

$$= E \left[ \frac{\partial l}{\partial \beta \partial \beta^T} \right]$$

$$= -E \left[ \frac{\partial l}{\partial \beta} \right] \left[ \frac{\partial l}{\partial \beta} \right]^T$$

$$= -\frac{1}{\phi^2} E \left[ \sum_i (y^{(i)} - b'(\beta^T x^{(i)})) x^{(i)} \sum_i (y^{(i)} - b'(\beta^T x^{(i)})) x^{(i)T} \right]$$

$$= -\frac{1}{\phi^2} E \left[ \sum_i (y^{(i)} - b'(\beta^T x^{(i)}))^2 x^{(i)} x^{(i)T} \right]$$

$$= -\frac{1}{\phi^2} \sum_i \text{Var}[y^{(i)}] x^{(i)} x^{(i)T}$$

where we have used the independence assumption in the fourth step. Comparing equation 5.5 with the last equation, we obtain the following relationship:

$$\text{Var}[y^{(i)}] = \phi b'''(\beta^T x^{(i)})$$

Moreover, because $f'(\eta) = b'(\eta)$, we have

$$\text{Var}[y^{(i)}] = \phi f'(\beta^T x^{(i)})$$

(5.6)

We now assemble the various pieces. First note that equation 5.6 can be utilized to express the Hessian (equation 5.5) in the following form:

$$\frac{\partial^2 l}{\partial \beta \partial \beta^T} = -\sum_i x^{(i)} x^{(i)T} w^{(i)}$$

---

*Whether or not the canonical link is used, the results presented in the remainder of this section are correct for the Fisher scoring method. If noncanonical links are used, then Newton–Raphson will include additional terms (terms that vanish under the expectation operator).*
where the weight \( w^{(t)} \) is defined as follows:

\[
        w^{(t)} = \frac{f'(\eta^{(t)})}{\text{Var}[\eta^{(t)}]}
\]

In matrix notation we have

\[
    \frac{\partial l}{\partial \beta \beta^T} = -X^TWX
\]

(5.7)

where \( X \) is the matrix whose rows are the input vectors \( x^{(t)} \) and \( W \) is a diagonal matrix whose diagonal elements are \( w^{(t)} \). Note also that the Hessian is a constant, thus the expected value of the Hessian is also \( X^TWX \).

Similarly, equation 5.6 can be used to remove the dependence of the gradient (equation 5.4) on \( \phi \):

\[
    \frac{\partial l}{\partial \beta} = \sum_{t} (y^{(t)} - \mu^{(t)}_{\phi})x^{(t)}w^{(t)}f'(\eta^{(t)})
\]

This equation can be written in matrix notation as follows:

\[
    \frac{\partial l}{\partial \beta} = X^T We
\]

(5.8)

where \( e \) is the vector whose components are

\[
    e^{(t)} = (y^{(t)} - \mu^{(t)}_{\phi})f'(\eta^{(t)})
\]

Finally, substitute equation 5.7 and equation 5.8 into equation 5.1 to obtain

\[
    \beta_{t+1} = \beta_t + (X^TWX)^{-1}X^TWe
\]

(5.9)

\[
    = (X^TWX)^{-1}X^TWe
\]

(5.10)

where \( z = X\beta + e^{(t)} \). These equations are the normal equations for a weighted least squares problem with observations \( [(x^{(t)}, z^{(t)})]_{t=1}^{m} \) and observation weights \( w^{(t)} \). The weights change from iteration to iteration, because they are a function of the parameters \( \beta_t \). The resulting iterative algorithm is known as iteratively reweighted least squares (IRLS).

It is easy to generalize the derivation to allow additional fixed observation weights to be associated with the data pairs. Such weights simply multiply the iteratively varying weights \( w^{(t)} \), leading to an iteratively reweighted weighted least squares algorithm. Such a generalization is in fact necessary in our application of IRLS: the EM algorithm defines observation weights in the outer loop that IRLS must treat as fixed during the inner loop.

Finally, it is also straightforward to generalize the derivation in this section to the case of vector outputs. In the case of vector outputs, each row of the weight matrix (e.g., \( U \) for the expert networks) is a separate parameter vector corresponding to the vector \( \beta \) of this section. These row vectors are updated independently and in parallel.

---

**Appendix B: Multinomial Logit Models**

The multinomial logit model is a special case of the generalized linear model in which the probabilistic component is the multinomial density or the Poisson density. It is of particular interest to us because the gating networks in the HME architecture are multinomial logit models.

Consider a multiway classification problem on \( n \) variables \( y_1, y_2, \ldots, y_n \). A natural probability model for multiway classification is the multinomial density:

\[
    P(y_1, y_2, \ldots, y_n) = \frac{m!}{(y_1!)^1(y_2!)^2 \cdots (y_n!)^n} p_1^{y_1} p_2^{y_2} \cdots p_n^{y_n}
\]

where the \( p_i \) are the multinomial probabilities associated with the different classes and \( m = \sum_{i=1}^{n} y_i \) is generally taken to equal one for classification problems. The multinomial density is an member of the exponential family and can be written in the following form:

\[
    P(y_1, y_2, \ldots, y_n) = \exp \left\{ \ln \frac{m!}{(y_1!)^1(y_2!)^2 \cdots (y_n!)^n} + \sum_{i=1}^{n} y_i \ln p_i \right\}
\]

(6.1)

Taking the logarithm of both sides, and dropping the terms that do not depend on the parameters \( p_i \), we see that the log likelihood for the multinomial logit model is the cross-entropy between the observations \( y_i \) and the parameters \( p_i \).

Implicit in equation 6.1 is the constraint that the \( p_i \) sum to one. This constraint can be made explicit by defining \( p_n \) as follows: \( p_n = 1 - \sum_{i=1}^{n-1} p_i \) and rewriting equation 6.1:

\[
    P(y_1, y_2, \ldots, y_n) = \exp \left\{ \ln \frac{m!}{(y_1!)^1(y_2!)^2 \cdots (y_n!)^n} + \sum_{i=1}^{n-1} y_i \ln p_i - n \ln p_n \right\}
\]

(6.2)

The natural parameters in an exponential family density are those quantities that appear linearly in the \( y_i \) (cf. equation 5.2), thus we define

\[
    \eta_i = \ln \frac{p_i}{p_n}
\]

(6.3)

Using \( p_n = 1 - \sum_{i=1}^{n-1} p_i \) implies

\[
    p_n = \frac{1}{1 + \sum_{i=1}^{n-1} e^{\eta_i}}
\]

and therefore equation 6.3 can be inverted to yield

\[
    p_i = \frac{e^{\eta_i}}{1 + \sum_{j=1}^{n-1} e^{\eta_j}} = \frac{e^{\eta_i}}{\sum_{j=1}^{n} e^{\eta_j}}
\]

(6.4)

using \( \eta_n = 0 \) from equation 6.3. This latter expression is the "softmax" function (Bridle 1989).
Finally, note that equation 6.2 implies that $b(\eta)$ must be defined as follows (cf. equation 5.2):

$$b(\eta) = n \ln \left( \sum_{i=1}^{n} e^{\eta_i} \right)$$

which implies

$$\mu_i = \frac{\partial b(\eta)}{\partial \eta_i} = \frac{n e^{\eta_i}}{\sum_{i=1}^{n} e^{\eta_i}} = n p_i \quad (6.5)$$

The fitting of a multinomial logit model proceeds by IRLS as described in Appendix A, using equations 6.4 and 6.5 for the link function and the mean, respectively.

Acknowledgments

We want to thank Geoffrey Hinton, Tony Robinson, Mitsuo Kawato, Carlotta Domeniconi, and Daniel Wolpert for helpful comments on the manuscript. This project was supported in part by a grant from the McDonnell-Pew Foundation, by a grant from ATR Human Information Processing Research Laboratories, by a grant from Siemens Corporation, by Grant IRI-9013991 from the National Science Foundation, and by Grant N00014-90-J-1942 from the Office of Naval Research. The project was also supported by NSF Grant ASC-9217041 in support of the Center for Biological and Computational Learning at MIT, including funds provided by DARPA under the HPCPC program, and NSF Grant ECS-9216531 to support an initiative in Intelligent Control at MIT. Michael I. Jordan is an NSF Presidential Young Investigator.

References


Received February 22, 1993; accepted July 9, 1993.