An Exposition of Hierarchical Mixtures of Experts

Yuki Zhang

August 25, 2024

Abstract

We present the probability model for the Hierarchical Mixtures of Experts architecture and discuss an implementation using the EM algorithm. We see how the maximization step of the EM algorithm in this application can be solved by Iteratively Reweighted Least Squares.

Contents

1 Introduction

Hierarchical Mixtures of Experts (HME) is an architecture for supervised learning. It approaches the problem with a tree structure in which "soft decisions" are made at the non-terminal nodes (called *gating networks*), which are used to form an average (or mixture) of the predictions made at the terminal nodes (called *experts*), which is then output as the model's final judgment.

To start introducing the concepts and to clarify the use of the phrase "soft decisions", we take a look at the CART algorithm, which, as we explain below, utilizes "hard decisions" at the non-terminal nodes. The example we will consider is a simple regression tree with one non-terminal node and two terminal nodes. Both the input and output are scalar-valued. This is depicted in the following figure.

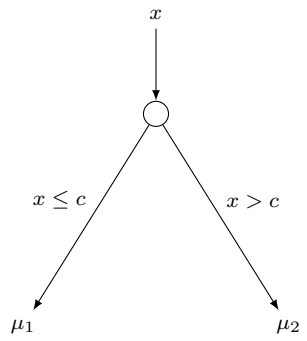

Figure 1: A CART Regression Tree

For a given x, if $x \leq c$, the model outputs μ_1. On the other hand, if $x > c$, the predicted value is μ_2. We can thus write the output of the model in the following way

$$\mu(x) = I(x \leq c)\mu_1 + I(x > c)\mu_2 \tag{1.1}$$

We will refer to this type of decisions as "hard decisions", emphasizing the fact that the result comes from either the left leaf or the right leaf.

Now let us consider a different way of determining the output given an input x. Specifically, let $\Delta(x)$ be a Bernoulli random variable with $P(\Delta(x) = 1) = p(x)$ where $0 \leq p(x) \leq 1$. Consider the random variable defined by

$$Y = \begin{cases} \mu_1 & \text{if } \Delta(x) = 1 \\ \mu_2 & \text{if } \Delta(x) = 0 \end{cases} \tag{1.2}$$

In contrast to (1.1), we define the model output to be

$$\begin{aligned}\mu(x) &= E(Y) \\ &= p(x)\mu_1 + (1 - p(x))\mu_2\end{aligned} \tag{1.3}$$

This represents a weighted average, or mixture, of the values μ_1 and μ_2 and is what we meant by a "soft decision". The HME model utilizes this type of decisions when constructing a response given an input. This is illustrated in the following Figure.

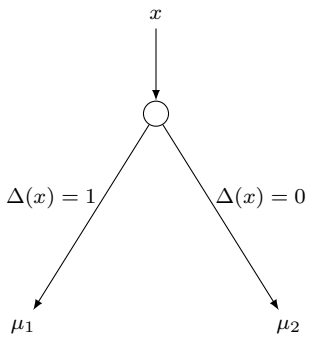

Figure 2: An HME Tree. The Bernoulli random variable $\Delta(x)$ is used to determine an "average" of the results at the leaves.

2 The HME Model

In this paper, for simplicity, we consider only HME structures with two levels of gating networks. We now describe the model in more detail.

Let $\mathbf{x}$ be a given input. To determine the corresponding output, we will introduce three groups of random variables whose distributions depend on $\mathbf{x}$, which, as we will see, correspond to the nodes in a tree representation of an HME model.

The first group correspond to what is called the top gating networks of the model. We denote by $\Delta_1(\mathbf{x}), \cdots, \Delta_n(\mathbf{x})$ random variables from a multinomial distribution with $\Delta_1(\mathbf{x}) + \cdots + \Delta_n(\mathbf{x}) = 1$ and class probabilities $p_1(\mathbf{x}), \cdots, p_n(\mathbf{x})$.

Secondly, for each $1 \leq i \leq n$, let $\Delta_{1|i}(\mathbf{x}), \cdots, \Delta_{m|i}(\mathbf{x})$ be random variables from a multinomial distribution with $\Delta_{1|i}(\mathbf{x}) + \cdots + \Delta_{m|i}(\mathbf{x}) = 1$ and class probabilities $p_{1|i}(\mathbf{x}), \cdots, p_{m|i}(\mathbf{x})$.

Finally, let $\mathbf{Y}_{ij}$, $1 \leq i \leq n$, $1 \leq j \leq m$ be random vectors from known distributions with means $\mu_{ij}(\mathbf{x})$ and probability functions f_{ij}.

The predicted value of the output given the input $\mathbf{x}$ is given by

$$\mu(\mathbf{x}) = \sum_i p_i(\mathbf{x}) \sum_j p_{j|i}(\mathbf{x}) \mu_{ij}(\mathbf{x}) \tag{2.1}$$

To see the rationale of this expression, let us consider the following construction. We define a random vector $\mathbf{Y}$ by

$$\mathbf{Y} = \sum_i \Delta_i(\mathbf{x}) \sum_j \Delta_{j|i}(\mathbf{x}) Y_{ij} \tag{2.2}$$

or equivalently

$$\mathbf{Y} = \mathbf{Y}_{ij} \text{ if } \Delta_i j(\mathbf{x}) = 1 \text{ and } \Delta_{j|i}(\mathbf{x}) = 1 \tag{2.3}$$

It is straight forward to see that the distribution of Y is given by

$$f(\mathbf{y} \mid \mathbf{x}) = \sum_i p_i(\mathbf{x}) \sum_j p_{j|i}(x) f_{ij}(\mathbf{y} \mid \mathbf{x}) \tag{2.4}$$

and that the expectation of $\mathbf{Y}$ is

$$\begin{aligned} E(\mathbf{Y}) &= E(\sum_i \Delta_i(\mathbf{x}) \sum_j \Delta_{j|i} \mathbf{Y}_{ij}) \\ &= \sum_i p_i(\mathbf{x}) \sum_j p_{j|i}(\mathbf{x}) \mu_{ij}(\mathbf{x}) \end{aligned} \tag{2.5}$$

which is the same as (2.1).

Therefore, we see that the output of an HME model is the expectation of an average of the results from the experts (the random variables $\mathbf{Y}_{ij}$) weighted by the decisions made at the gating networks (the random variables Δ_i, $\Delta_{j|i}$). The case with $n = 2$ and $m = 3$ is illustrated in Figure 3.

How do the probability functions p_i, $p_{j|i}$ and f_{ij} depend on $\mathbf{x}$? The distributions are assumed to be generalized linear models with respect to $\mathbf{x}$. That is,

$$g(E(U \mid \mathbf{x})) = \boldsymbol{\beta}^\intercal \mathbf{x} \tag{2.6}$$

where g is called a link function and $\mathbf{x} = (1, x_1, \cdots, x_{p-1})$ is the input augmented with 1 in the first column to account for the intercept.

At the gating networks, the link function g is the multinomial analogue of the logit function, i.e.

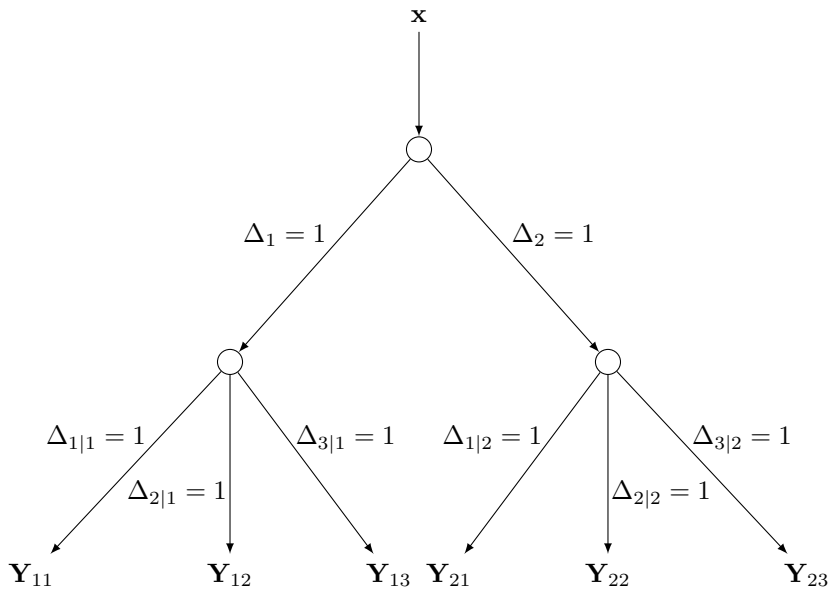

Figure 3: A two-level HME tree. The top gating network has 2 classes and the lower gating networks have 3 classes, respectively.

$$\log \frac{p_i}{p_n} = \boldsymbol{\beta}_i^\intercal \mathbf{x}$$
$$\log \frac{p_{j|i}}{p_{n|i}} = \boldsymbol{\beta}_{j|i}^\intercal \mathbf{x}$$

The link function at the expert nodes depend on the problem at hand. For regression problems, g is generally taken to be the identity function. On the other hand, for binary classification problems, g is generally taken to be the logistic function.

3 Fitting the Model

Since the distributions at all nodes are assumed to be generalized linear and since the functions f_{ij} are known, fitting the model amounts to estimating the parameters involved in the distributions and can be viewed as a maximum likelihood problem. This is considered in[1].

Jordan and Jacobs [1] discuss a gradient descent algorithm for estimating the parameters. In this paper, we will focus on an algorithm that is an application of the EM algorithm to the HME architecture.

3.1 The EM Algorithm

In this section, we briefly discuss the EM algorithm. Our presentation is based on [2].

The EM algorithm is a procedure for maximum likehood problems with unobserved (also called latent) data. The name EM stand for expectation-maximization which are the two iterated steps in the algorithm.

We start with some notations. Denote by $\mathbf{Z}$ the observed data, $\mathbf{Z}^m$ the unobserved (or missing) data and $\mathbf{T} = (\mathbf{Z}, \mathbf{Z}^m)$ the complete data. Denote by $\boldsymbol{\theta}$ the parameters that the probability function of $\mathbf{T}$ depend on. By definition, we have

$$P(\mathbf{Z}^m \mid \mathbf{Z}, \boldsymbol{\theta}') = \frac{P(\mathbf{Z}, \mathbf{Z}^m \mid \boldsymbol{\theta}')}{P(\mathbf{Z} \mid \boldsymbol{\theta}')} \tag{3.1}$$

In terms of log-likelihoods, this gives rise to

$$l(\boldsymbol{\theta}'; \mathbf{Z}) = l(\boldsymbol{\theta}'; \mathbf{T}) - l(\boldsymbol{\theta}'; \mathbf{Z}^m \mid \mathbf{Z}) \tag{3.2}$$

Taking expectation with respect to the conditional distribution of $\mathbf{T}$ on $\mathbf{Z}$ with parameters $\boldsymbol{\theta}$, we have

$$\begin{aligned} l(\boldsymbol{\theta}'; \mathbf{Z}) &= E(l(\boldsymbol{\theta}'; \mathbf{T}) \mid \mathbf{Z}, \boldsymbol{\theta}) - E(l(\boldsymbol{\theta}'; \mathbf{Z}^m \mid \mathbf{Z}) \mid \mathbf{Z}, \boldsymbol{\theta}) \\ &= Q(\boldsymbol{\theta}', \boldsymbol{\theta}) - R(\boldsymbol{\theta}', \boldsymbol{\theta}) \end{aligned}$$

where

$$\begin{aligned} Q(\boldsymbol{\theta}', \boldsymbol{\theta}) &= E(l(\boldsymbol{\theta}'; \mathbf{T}) \mid \mathbf{Z}, \boldsymbol{\theta}) \\ R(\boldsymbol{\theta}', \boldsymbol{\theta}) &= E(l(\boldsymbol{\theta}'; \mathbf{Z}^m \mid \mathbf{Z}) \mid \mathbf{Z}, \boldsymbol{\theta}) \end{aligned}$$

If $\boldsymbol{\theta}'$ maximizes $Q(\boldsymbol{\theta}', \boldsymbol{\theta})$, we have

$$Q(\boldsymbol{\theta}', \boldsymbol{\theta}) - Q(\boldsymbol{\theta}, \boldsymbol{\theta}) \geq 0 \tag{3.3}$$

On the other hand, we have

$$\begin{aligned} R(\boldsymbol{\theta}', \boldsymbol{\theta}) - R(\boldsymbol{\theta}, \boldsymbol{\theta}) &= E(l(\boldsymbol{\theta}'; \mathbf{Z}^m \mid \mathbf{Z}) - l(\boldsymbol{\theta}; \mathbf{Z}^m \mid \mathbf{Z}) \mid \mathbf{Z}, \boldsymbol{\theta}) \\ &= E(\log(\frac{L(\boldsymbol{\theta}'; \mathbf{Z}^m \mid \mathbf{Z})}{L(\boldsymbol{\theta}; \mathbf{Z}^m \mid \mathbf{Z})}) \mid \mathbf{Z}, \boldsymbol{\theta}) \end{aligned}$$

Note that the above expectation is taken with respect to $L(\boldsymbol{\theta}; \mathbf{Z}^m \mid \mathbf{Z})$ viewed as a probability function. To make further progress, we need the following lemma.

Lemma 3.1. *Let g, h be probability functions. We have that*

$$E_h(\log \frac{g}{h})$$

where the expectation is taken with respect to h is maximized when $g = h$.

Proof. Since the function $f(x) = -\log x$ is convex, Jensen's inequality implies that

$$\begin{aligned} E(-\log \frac{g}{h}) &\geq -\log E(\frac{g}{h}) \\ &= -\log \int \frac{g}{h} h \\ &= -\log \int g = 0 \end{aligned}$$

So we have $E(\log \frac{g}{h}) \leq 0 = E(\log \frac{h}{h})$, which completes the proof. □

This lemma and the note before it imply that

$$R(\boldsymbol{\theta}', \boldsymbol{\theta}) - R(\boldsymbol{\theta}, \boldsymbol{\theta}) \leq 0 \tag{3.4}$$

Combining (3.3) and (3.4) together, we have

$$l(\boldsymbol{\theta}'; \mathbf{Z}) - l(\boldsymbol{\theta}; \mathbf{Z}) \geq 0 \tag{3.5}$$

So the choice of $\boldsymbol{\theta}' = \text{argmax}_{\boldsymbol{\theta}^*} Q(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ does not make the log-likelihood of the observed data $l(\boldsymbol{\theta}'; \mathbf{Z})$ smaller. This gives an explanation of why the EM algorithm presented as follows works in approximating the maximum likelihood estimates.

Algorithm 3.2. *Let $\boldsymbol{\theta}_0$ be initial guesses of the parameters. Starting from $i = 0$, iterate the following steps until convergence.*

> *Expectation Step: Take the expectation of the total likelihood $l(\boldsymbol{\theta}'; \mathbf{T})$ with respect to the conditional distribution of $\mathbf{T}$ given $\mathbf{Z}$ and parameters $\boldsymbol{\theta}_i$*
>
> $$Q(\boldsymbol{\theta}', \boldsymbol{\theta}) = E(l(\boldsymbol{\theta}'; \mathbf{T}) \mid \mathbf{Z}, \boldsymbol{\theta}_i) \tag{3.6}$$
>
> *Maximization Step: Let $\boldsymbol{\theta}_{i+1}$ be the maximizer of $Q(\boldsymbol{\theta}', \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}'$*
>
> $$\boldsymbol{\theta}_{i+1} = \underset{\boldsymbol{\theta}'}{\text{argmax}}\, Q(\boldsymbol{\theta}', \boldsymbol{\theta}_i) \tag{3.7}$$

The above discussion shows that, for each $i \geq 0$, we have

$$l(\boldsymbol{\theta}_{i+1}; \mathbf{Z}) \geq l(\boldsymbol{\theta}_i; \mathbf{Z}) \tag{3.8}$$

3.2 Applying the EM Algorithm To HME

From (2.4), the log-likelihood of a sample $(\mathbf{Y}, \mathbf{X})$ is given by

$$l(\boldsymbol{\theta}; \mathbf{Y}|\mathbf{X}) = \sum_t \log \left[\sum_i p_i(\mathbf{x}^{(t)}) \sum_j p_{j|i}(\mathbf{x}^{(t)}) f_{ij}(\mathbf{y}^{(t)} \mid \mathbf{x}^{(t)}) \right] \tag{3.9}$$

To apply the EM algorithm, we introduce latent variables as follows. Let $\Delta_i^{(t)}$, $i = 1, \cdots, n$ be random variables from a multinomial distribution with class probabilities $p_i(\mathbf{x}^{(t)})$ and $\Delta_1^{(t)} + \cdots + \Delta_n^{(t)} = 1$. Similarly, for each $1 \leq i \leq n$, let $\Delta_{j|i}^{(t)}$, $j = 1, \cdots, m$ be random variables from a multinomial distribution with class probabilities $p_{j|i}(\mathbf{x}^{(t)})$ and $\Delta_{1|i}^{(t)} + \cdots + \Delta_{m|i}^{(t)} = 1$. The log likelihood of the complete data $\mathbf{T} = (\mathbf{Y}, \Delta_i, \Delta_{j|i})$ is given by

$$l(\boldsymbol{\theta}; \mathbf{T}|\mathbf{X}) = \sum_t \sum_i \sum_j \Delta_i^{(t)} \Delta_{j|i}^{(t)} \log \left[p_i(\mathbf{x}^{(t)}) p_{j|i}(\mathbf{x}^{(t)}) f_{ij}(\mathbf{y}^{(t)} \mid \mathbf{x}^{(t)}) \right] \tag{3.10}$$

Comparing (3.10) with(3.9), we see that the introduction of the latent variables $\Delta_i^{(t)}$ and $\Delta_{j|i}^{(t)}$ allowed the logarithm to be brought inside the sum over i and j. We now examine the two steps of the EM algorithm in this case.

Expectation Step

If we take the expectation of $l(\boldsymbol{\theta}; \mathbf{T}|\mathbf{X})$ with respect to the conditional distribution given $\mathbf{Y}$, we obtain

$$E(l(\boldsymbol{\theta}; \mathbf{T}|\mathbf{X}) \mid \mathbf{Y}, \boldsymbol{\theta}) = \sum_t \sum_i \sum_j h_i^{(t)} h_{j|i}^{(t)} \log \left[p_i(\mathbf{x}^{(t)}) p_{j|i}(\mathbf{x}^{(t)}) f_{ij}(\mathbf{y}^{(t)} \mid \mathbf{x}^{(t)}) \right] \tag{3.11}$$

where

$$\begin{aligned} h_i^{(t)} &= P(\Delta_i^{(t)} = 1 \mid \mathbf{Y}, \boldsymbol{\theta}) \\ &= \frac{\sum_j P(\Delta_i^{(t)} = 1, \Delta_{j|i}^{(t)} = 1, \mathbf{y}^{(t)})}{P(\mathbf{y}^{(t)})} \\ &= \frac{\sum_j P(\mathbf{y}^{(t)} \mid \Delta_i^{(t)} = 1, \Delta_{j|i}^{(t)} = 1)}{P(\mathbf{y}^{(t)})} \\ &= \frac{p_i(\mathbf{x}^{(t)}) \sum_j p_{j|i}(\mathbf{x}^{(t)}) f_{ij}(\mathbf{y}^{(t)} \mid \mathbf{x}^{(t)})}{\sum_i p_i(\mathbf{x}^{(t)}) \sum_j p_{j|i}(\mathbf{x}^{(t)}) f_{ij}(\mathbf{y}^{(t)} \mid \mathbf{x}^{(t)})} \end{aligned}$$

Similarly, we have

$$
\begin{aligned}
h_{j|i}^{(t)} &= P(\Delta_{j|i}^{(t)} = 1 \mid \Delta_i^{(t)} = 1, \mathbf{Y}, \boldsymbol{\theta}) \\
&= \frac{P(\Delta_i^{(t)} = 1, \Delta_{j|i}^{(t)} = 1, \mathbf{y}^{(t)})}{P(\mathbf{y}^{(t)}, \Delta_i^{(t)} = 1)} \\
&= \frac{P(\mathbf{y}^{(t)} \mid \Delta_i^{(t)} = 1, \Delta_{j|i}^{(t)} = 1)}{P(\mathbf{y}^{(t)} | \Delta_i^{(t)} = 1) P(\Delta_i^{(t)} = 1)} \\
&= \frac{p_i(\mathbf{x}^{(t)}) p_{j|i}(\mathbf{x}^{(t)}) f_{ij}(\mathbf{y}^{(t)} \mid \mathbf{x}^{(t)})}{p_i(\mathbf{x}^{(t)}) \sum_j p_{j|i}(\mathbf{x}^{(t)}) f_{ij}(\mathbf{y}^{(t)} \mid \mathbf{x}^{(t)})}
\end{aligned}
$$

It is also convenient to define the constants

$$
\begin{aligned}
h_{ij}^{(t)} &= P(\Delta_i^{(t)} = 1, \Delta_{j|i}^{(t)} = 1 \mid \mathbf{Y}, \boldsymbol{\theta}) \\
&= h_i^{(t)} h_{j|i}^{(t)} \\
&= \frac{p_i(\mathbf{x}^{(t)}) p_{j|i}(\mathbf{x}^{(t)}) f_{ij}(\mathbf{y}^{(t)} \mid \mathbf{x}^{(t)})}{\sum_i p_i(\mathbf{x}^{(t)}) \sum_j p_{j|i}(\mathbf{x}^{(t)}) f_{ij}(\mathbf{y}^{(t)} \mid \mathbf{x}^{(t)})}
\end{aligned}
$$

Maximization Step

If we denote the parameters that the probabilities $p_i, p_{j|i}, p_{ij}$ depend on by $\boldsymbol{\beta}^{(1)}$, $\boldsymbol{\beta}^{(2|i)}$ and $\boldsymbol{\theta_{ij}}$, we see that the maximization of (3.11) is equivalent to three separate problems, namely,

$$
\boldsymbol{\beta}^{(1)} = \underset{\boldsymbol{\beta}^{(1)}}{\operatorname{argmax}} \sum_t \sum_i h_i^{(t)} \log p_i(\mathbf{x}^{(t)}) \tag{3.12}
$$

$$
\boldsymbol{\beta}^{(2,i)} = \underset{\boldsymbol{\beta}^{(2,i)}}{\operatorname{argmax}} \sum_t h_i^{(t)} \sum_j \log p_{j|i}(\mathbf{x}^{(t)}) \tag{3.13}
$$

$$
\boldsymbol{\theta_{ij}} = \underset{\boldsymbol{\theta_{ij}}}{\operatorname{argmax}} \sum_t h_{ij}^{(t)} \log f_{ij}(\mathbf{y}^{(t)} \mid \mathbf{x}^{(t)}) \tag{3.14}
$$

Problem (3.14) is a weighted maximum likelihood problem for the probability function f_{ij}. Problem (3.12) is a maximum likelihood problem for the multinomial logit model with observations $h_i^{(t)}$. Problem (3.13) is a weighted maximum likelihood problem for the multinomial logit model with observations $h_{j|i}^{(t)}$ and observation weights $h_i^{(t)}$.

Due to our parameterization of the probabilities in in terms of $\mathbf{x}$, all three problems are maximum likelihood problems for a generalized linear model and can be solved using the Iteratively Reweighted Least Squares (IRLS) algorithm, which we will now discuss.

4 Iteratively Reweighted Least Squares

In this section, we discuss the Iteratively Reweighted Least Squares (IRLS) algorithm. IRLS is an algorithm for computing maximum likelihood estimates of the parameters in a generalized linear model. It is an application of the Fisher Scoring algorithm, which is an algorithm for general maximum likelihood problems, to the case of generalized linear models. We will see that, in this case, Fisher Scoring reduces to an iterative algorithm in which each step is a weighted least squares problem and hence the name of the algorithm. Our presentation is based on and is a natural generalization of the dicussion in [1] in the following two ways.

1. The underlying distribution of the maximum likelihood problem is that of a random vector.

2. In the maximum likelihood problem, we allow observation weights.

As detailed in the next section, with these two generalizations, IRLS can be applied in the Maximization Step of the EM algorithm used to fit the HME model.

4.1 The Exponential Family

The distribution of a random variable Y is said to be a member of the exponential family if the probability function can be written in the following form

$$f(y, \eta, \phi) = \exp\left[\frac{\eta y - b(\eta)}{\phi} + c(y, \phi)\right] \tag{4.1}$$

where η is called the natural parameter, ϕ is called the dispersion parameter and b and c are functions. This can be generalized to a family of distributions of random vectors. To simplify the development, we will consider only a subclass of this family. Namely, we will restrict our attention to distributions whose probability functions can be written in the following form

$$f(\mathbf{y}, \boldsymbol{\eta}) = \exp\left[\boldsymbol{\eta}^{\mathsf{T}}\mathbf{y} - b(\boldsymbol{\eta}) + c(\mathbf{y})\right] \tag{4.2}$$

In the case of a random variable, (4.2) is the same as (4.1) with $\phi = 1$, i.e. the dispersion parameter is 1.

4.2 Fisher Scoring

Let $l(\boldsymbol{\beta}; \mathbf{Y})$ be a log-likelihood. The Fisher scoring method is an iterative algorithm for estimating $\boldsymbol{\beta}$ in which, in each iteration, the parameters are updated in the following way

$$\boldsymbol{\beta}_{r+1} = \boldsymbol{\beta}_r - \left(E\left[\frac{\partial l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathsf{T}}}\right]\right)^{-1} \frac{\partial l}{\partial \boldsymbol{\beta}} \tag{4.3}$$

where the expectation is taken with respect to the distribution of $\mathbf{T}$ with parameters $\boldsymbol{\beta}_r$ We note the similarity between (4.3) and the update criterion of the Newton-Raphson algorithm. Indeed, Fisher scoring replaces the Hessian $\frac{\partial l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\intercal}$ used in Newton-Raphson with its expectation.

Let $\mathbf{y}$ be a random vector whose distribution is given by (4.2). Generalized linear models assume that $\boldsymbol{\eta}$ can be written in the form

$$\boldsymbol{\eta} = \boldsymbol{\beta}^\intercal \mathbf{x} \tag{4.4}$$

where $\boldsymbol{\beta}$ is a parameter vector. Note that the intercept term is included in the above expression, i.e. $x = (1, x_1, \cdots, x_{p-1})$.

Thus, for a random sample $\mathbf{Y} = (\mathbf{y}^{(t)}$ and corresponding input $\mathbf{X} = (\mathbf{x}^{(t)}$, the log-likelihood of the sample is given by

$$l(\boldsymbol{\beta}; \mathbf{Y}|\mathbf{X}) = \sum_{t=1}^{N} \left[\sum_{i=1}^{n} \boldsymbol{\beta}_i^\intercal \mathbf{x}^{(t)} y_i^{(t)} - b(\boldsymbol{\beta}_1^\intercal \mathbf{x}^{(t)}, \cdots, \boldsymbol{\beta}_n^\intercal \mathbf{x}^{(t)}) + c(\mathbf{y}^{(t)}) \right] \tag{4.5}$$

where n is the length of the random vector $\mathbf{y}$ and N is the number of observations in the sample.

For each $1 \leq i \leq n$, we have

$$\frac{\partial l}{\partial \boldsymbol{\beta}_i} = \sum_t (y_i^{(t)} - \frac{\partial b}{\partial \eta_i}) \mathbf{x}^{(t)}. \tag{4.6}$$

For each $1 \leq i, j \leq n$, we have

$$\frac{\partial l}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_i^\intercal} = \sum_t (\frac{\partial^2 b}{\partial \eta_i \partial \eta_j}) \mathbf{x}^{(t)} \mathbf{x}^{(t)\intercal}. \tag{4.7}$$

To continue, we need two identities for log-likelihoods.

Lemma 4.1. *Let $L(\boldsymbol{\beta}; \mathbf{Y})$ be a likelihood function and let $l(\boldsymbol{\beta}; \mathbf{Y}) = \log L(\boldsymbol{\beta}; \mathbf{Y})$ be the log-likelihood. Then we have*

$$E(\frac{\partial l}{\partial \boldsymbol{\beta}}) = \mathbf{0}$$
$$E(\frac{\partial l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\intercal}) = -E(\frac{\partial l}{\partial \boldsymbol{\beta}} \frac{\partial l}{\partial \boldsymbol{\beta}^\intercal})$$

Proof. For each $1 \leq i \leq n$, we have

$$\begin{aligned} 0 = \frac{\partial}{\partial \beta_i} \int L = \int \frac{\partial L}{\partial \beta_i} &= \int \frac{\frac{\partial L}{\partial \beta_i}}{L} L \\ &= \int \frac{\partial l}{\partial \beta_i} L = E(\frac{\partial l}{\partial \beta_i}). \end{aligned}$$

For each $1 \le i, j \le n$, we have

$$\begin{aligned} 0 &= \frac{\partial}{\partial \beta_j} \int \frac{\partial l}{\partial \beta_i} L \\ &= \int \frac{\partial^2 l}{\partial \beta_i \partial \beta_j} L + \int \frac{\partial l}{\partial \beta_i} \frac{\partial L}{\partial \beta_i} \\ &= E(\frac{\partial^2 l}{\partial \beta_i \partial \beta_j}) + \int \frac{\partial l}{\partial \beta_i} \frac{\frac{\partial L}{\partial \beta_i}}{L} L \\ &= E(\frac{\partial^2 l}{\partial \beta_i \partial \beta_j}) + E(\frac{\partial l}{\partial \beta_i} \frac{\partial l}{\partial \beta_j}). \end{aligned}$$

This completes the proof. □

This first identity in the lemma and (4.6) imply that, for any $1 \le i \le n$

$$E(y_i^{(t)}) = \frac{\partial b}{\partial \eta_i}. \tag{4.8}$$

Using the second identity, (4.6) and (4.7) imply that

$$\begin{aligned} -\sum_t (\frac{\partial^2 b}{\partial \eta_i \partial \eta_j}) \mathbf{x}^{(t)} \mathbf{x}^{(t)\intercal} &= E(\frac{\partial l}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_i^\intercal}) \\ &= -E(\frac{\partial l}{\partial \beta_i} \frac{\partial l}{\partial \beta_j}) \\ &== E\left(\sum_t (y_i^{(t)} - \frac{\partial b}{\partial \eta_i}) \mathbf{x}^{(t)} \sum_s (y_i^{(s)} - \frac{\partial b}{\partial \eta_j}) \mathbf{x}^{(s)\intercal} \right) \\ &= -E\left(\sum_t (y_i^{(t)} - \frac{\partial b}{\partial \eta_i})(y_j^{(t)} - \frac{\partial b}{\partial \eta_j}) \mathbf{x}^{(t)} \mathbf{x}^{(t)\intercal} \right) \\ &= -\sum_t \mathrm{Cov}(y_i^{(t)}, y_j^{(t)}) \mathbf{x}^{(t)} \mathbf{x}^{(t)\intercal} \end{aligned} \tag{4.9}$$

where, in the second to last step, we used the independence of the sample $\mathbf{Y}$, which implies that $E((y_i^{(t)} - \frac{\partial b}{\partial \eta_i})(y_j^{(s)} - \frac{\partial b}{\partial \eta_j})) = 0$ for $t \ne s$. Comparing the coefficients of $\mathbf{x}^{(t)} \mathbf{x}^{(t)\intercal}$ on both sides, we have

$$\mathrm{Cov}(y_i^{(t)}, y_j^{(t)}) = \frac{\partial^2 b}{\partial \eta_i \partial \eta_j}$$

For each $1 \le i, j \le n$, let W_{ij} be the diagonal matrix whose (t,t)th element is $\mathrm{Cov}(y_i^{(t)}, y_j^{(t)})$. From the derivation of (4.9), we have

$$
\begin{aligned}
E(\frac{\partial l}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_j^\intercal}) &= -\sum_t \mathrm{Cov}(y_i^{(t)}, y_j^{(t)}) \mathbf{x}^{(t)} \mathbf{x}^{(t)\intercal} \\
&= -\mathbf{X}^\intercal W_{ij} \mathbf{X}
\end{aligned} \tag{4.10}
$$

It follows that

$$
\begin{aligned}
E(\frac{\partial l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\intercal}) &= E \begin{pmatrix} \frac{\partial l}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^\intercal} & \cdots & \frac{\partial l}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_n^\intercal} \\ \vdots & \ddots & \vdots \\ \frac{\partial l}{\partial \boldsymbol{\beta}_n \partial \boldsymbol{\beta}_1^\intercal} & \cdots & \frac{\partial l}{\partial \boldsymbol{\beta}_n \partial \boldsymbol{\beta}_n^\intercal} \end{pmatrix} \\
&= \begin{pmatrix} -\mathbf{X}^\intercal W_{11} \mathbf{X} & \cdots & -\mathbf{X}^\intercal W_{1n} \mathbf{X} \\ \vdots & \ddots & \vdots \\ -\mathbf{X}^\intercal W_{n1} \mathbf{X} & \cdots & -\mathbf{X}^\intercal W_{nn} \mathbf{X} \end{pmatrix} \\
&= -\begin{pmatrix} \mathbf{X}^\intercal & & \\ & \ddots & \\ & & \mathbf{X}^\intercal \end{pmatrix} \begin{pmatrix} W_{11} & \cdots & W_{1n} \\ \vdots & \ddots & \vdots \\ W_{n1} & \cdots & W_{nn} \end{pmatrix} \begin{pmatrix} \mathbf{X} & & \\ & \ddots & \\ & & \mathbf{X} \end{pmatrix} \\
&= -\tilde{\mathbf{X}}^\intercal \mathbf{W} \tilde{\mathbf{X}}
\end{aligned}
$$

where we have

$$
\begin{aligned}
\tilde{\mathbf{X}} &= \begin{pmatrix} \mathbf{X} & & \\ & \ddots & \\ & & \mathbf{X} \end{pmatrix} \\
\mathbf{W} &= \begin{pmatrix} W_{11} & \cdots & W_{1n} \\ \vdots & \ddots & \vdots \\ W_{n1} & \cdots & W_{nn} \end{pmatrix}
\end{aligned}
$$

As a side note, since (4.7) shows that $\frac{\partial l}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_i^\intercal}$ is a constant, we have

$$
E(\frac{\partial l}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_i^\intercal}) = \frac{\partial l}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_i^\intercal}
$$

From (4.6) and (4.8), we have

$$
\frac{\partial l}{\partial \boldsymbol{\beta}_i} = \sum_t (y_i^{(t)} - \mu_i^{(t)}) \mathbf{x}^{(t)}. \tag{4.11}
$$

where $\mu_i^{(t)} = E(y_i^{(t)})$. It follows that

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial l}{\partial \boldsymbol{\beta}_1} \\ \vdots \\ \frac{\partial l}{\partial \boldsymbol{\beta}_n} \end{pmatrix} = \begin{pmatrix} \sum_t (y_1^{(t)} - \mu_1^{(t)}) \mathbf{x}^{(t)} \\ \vdots \\ \sum_t (y_n^{(t)} - \mu_n^{(t)}) \mathbf{x}^{(t)} \end{pmatrix}$$
$$= \tilde{\mathbf{X}}^\intercal \begin{pmatrix} \mathbf{y}_1 - \boldsymbol{\mu}_1 \\ \vdots \\ \mathbf{y}_n - \boldsymbol{\mu}_n \end{pmatrix} \tag{4.12}$$
$$= \tilde{\mathbf{X}}^\intercal \mathbf{W} \mathbf{W}^{-1} \begin{pmatrix} \mathbf{y}_1 - \boldsymbol{\mu}_1 \\ \vdots \\ \mathbf{y}_n - \boldsymbol{\mu}_n \end{pmatrix}.$$

We define

$$\mathbf{e} = \begin{pmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_n \end{pmatrix} = \mathbf{W}^{-1} \begin{pmatrix} \mathbf{y}_1 - \boldsymbol{\mu}_1 \\ \vdots \\ \mathbf{y}_n - \boldsymbol{\mu}_n \end{pmatrix}.$$

Putting the above results together, we have, from (4.3),

$$\begin{aligned} \boldsymbol{\beta}_{r+1} &= \boldsymbol{\beta}_r - \left(\frac{\partial l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\intercal} \right)^{-1} \frac{\partial l}{\partial \boldsymbol{\beta}} \\ &= \boldsymbol{\beta}_r + (\tilde{\mathbf{X}}^\intercal \mathbf{W} \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\intercal \mathbf{W} \mathbf{e} \\ &= (\tilde{\mathbf{X}}^\intercal \mathbf{W} \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\intercal \mathbf{W} \mathbf{z} \end{aligned}$$

where $\mathbf{z} = \tilde{\mathbf{X}} \boldsymbol{\beta}_r + \mathbf{e}$.

4.3 Observation Weights

Let $c^{(t)}$, $1 \le t \le N$ be constants. We consider the problem of finding the maximum likelihood estimates when the log-likelihood is weighted by the $c^{(t)}$. Namely, we want to estimate $\boldsymbol{\beta}$ that maximize the function

$$l^c(\boldsymbol{\beta}; \mathbf{Y} | \mathbf{X}) = \sum_{t=1}^{N} c^{(t)} \left[\sum_{i=1}^{n} \boldsymbol{\beta}_i^\intercal \mathbf{x}^{(t)} y_i^{(t)} - b(\boldsymbol{\beta}_1^\intercal \mathbf{x}^{(t)}, \cdots, \boldsymbol{\beta}_n^\intercal \mathbf{x}^{(t)}) + c(\mathbf{y}^{(t)}) \right] \tag{4.13}$$

For $1 \le i, j \le n$, let W_{ij}^c denote the diagonal matrix whose (t,t)th element is $\mathrm{Cov}(y_i^{(t)}, y_j^{(t)})$. By a similar derivation to that of (4.10), we have

$$\begin{aligned} E\left(\frac{\partial l}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_j^\intercal}\right) &= -\sum_t c^{(t)} \mathrm{Cov}(y_i^{(t)}, y_j^{(t)}) \mathbf{x}^{(t)} \mathbf{x}^{(t)\intercal} \\ &= -\mathbf{X}^\intercal W_{ij}^t \mathbf{X} \end{aligned}$$

It follows that

$$
\begin{aligned}
E(\frac{\partial l^c}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\intercal}) &= \begin{pmatrix} -\mathbf{X}^\intercal W_{11}^c \mathbf{X} & \cdots & -\mathbf{X}^\intercal W_{1n}^c \mathbf{X} \\ \vdots & \ddots & \vdots \\ -\mathbf{X}^\intercal W_{n1}^c \mathbf{X} & \cdots & -\mathbf{X}^\intercal W_{nn}^c \mathbf{X} \end{pmatrix} \\
&= -\begin{pmatrix} \mathbf{X}^\intercal & & \\ & \ddots & \\ & & \mathbf{X}^\intercal \end{pmatrix} \begin{pmatrix} W_{11}^c & \cdots & W_{1n}^c \\ \vdots & \ddots & \vdots \\ W_{n1}^c & \cdots & W_{nn}^c \end{pmatrix} \begin{pmatrix} \mathbf{X} & & \\ & \ddots & \\ & & \mathbf{X} \end{pmatrix} \\
&= -\tilde{\mathbf{X}}^\intercal \mathbf{W} \tilde{\mathbf{X}}
\end{aligned}
$$

where

$$
\mathbf{W^c} = \begin{pmatrix} W_{11}^c & \cdots & W_{1n}^c \\ \vdots & \ddots & \vdots \\ W_{n1}^c & \cdots & W_{nn}^c \end{pmatrix}
$$

Moreover, similar to (4.12), we have

$$
\begin{aligned}
\frac{\partial l}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial l}{\partial \boldsymbol{\beta}_1} \\ \vdots \\ \frac{\partial l}{\partial \boldsymbol{\beta}_n} \end{pmatrix} &= \begin{pmatrix} \sum_t c^t (y_1^{(t)} - \mu_1^{(t)}) \mathbf{x}^{(t)} \\ \vdots \\ \sum_t c^t (y_n^{(t)} - \mu_n^{(t)}) \mathbf{x}^{(t)} \end{pmatrix} \\
&= \tilde{\mathbf{X}}^\intercal \begin{pmatrix} \left(c^{(t)} (y_1^{(t)} - \mu_1^{(t)}) \right) \\ \vdots \\ \left(c^{(t)} (y_n^{(t)} - \mu_n^{(t)}) \right) \end{pmatrix} \\
&= \tilde{\mathbf{X}}^\intercal \mathbf{W}^c (\mathbf{W}^c)^{-1} \begin{pmatrix} \left(c^{(t)} (y_1^{(t)} - \mu_1^{(t)}) \right) \\ \vdots \\ \left(c^{(t)} (y_n^{(t)} - \mu_n^{(t)}) \right) \end{pmatrix} \\
&= \tilde{\mathbf{X}}^\intercal \mathbf{W}^c \mathbf{W}^{-1} \begin{pmatrix} \mathbf{I}_c^{-1} & & \\ & \ddots & \\ & & \mathbf{I}_c^{-1} \end{pmatrix} \begin{pmatrix} \left(c^{(t)} (y_1^{(t)} - \mu_1^{(t)}) \right) \\ \vdots \\ \left(c^{(t)} (y_n^{(t)} - \mu_n^{(t)}) \right) \end{pmatrix} \\
&= \tilde{\mathbf{X}}^\intercal \mathbf{W}^c \mathbf{W}^{-1} \begin{pmatrix} \mathbf{y}_1 - \boldsymbol{\mu}_1 \\ \vdots \\ \mathbf{y}_n - \boldsymbol{\mu}_n \end{pmatrix} \\
&= \tilde{\mathbf{X}}^\intercal \mathbf{W}^c \mathbf{W}^{-1} \mathbf{e}
\end{aligned}
$$

where $\mathbf{I}_c$ is the diagonal matrix with $\mathbf{c}$ on the diagonal.

$$\mathbf{I}_c = \begin{pmatrix} c^{(1)} & & \\ & \ddots & \\ & & c^{(n)} \end{pmatrix}.$$

It follows that

$$\begin{aligned} \boldsymbol{\beta}^c_{r+1} &= \boldsymbol{\beta}^c_r - \left(\frac{\partial l^c}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\intercal}\right)^{-1} \frac{\partial l}{\partial \boldsymbol{\beta}} \\ &= \boldsymbol{\beta}^c_r + (\tilde{\mathbf{X}}^\intercal \mathbf{W}^c \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\intercal \mathbf{W}^c \mathbf{e} \\ &= (\tilde{\mathbf{X}}^\intercal \mathbf{W}^c \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\intercal \mathbf{W}^c \mathbf{z}. \end{aligned}$$

5 IRLS for Multinomial Distribution

In this section, we will see how IRLS discussed above can be applied in maximum likelihood problems for the multinomial logit model with observation weights. This will be used in solving the problems (3.12) and (3.13).

The multinomial logit model is a generalized linear model in which the underlying distribution is the multinomial distribution whose probability function is given by

$$p(z_1, \cdots, z_n) = \frac{m!}{z_1! \cdots z_n!} p_1^{z_1} \cdots p_n^{z_n}$$

with $\sum z_i = m$. We view it as a function of $z_1, \cdots, z_{n-1}$ and rewrite it as follows

$$p(z_1, \cdots, z_{n-1}) = \exp\left(\log \frac{m!}{z_1! \cdots z_n!} + \sum_{i=1}^{n-1} z_i \log \frac{p_i}{p_n} + m \log_{p_n}\right) \tag{5.1}$$

Comparing this with the general form of an exponential family probability function (4.2), we see that the natural parameters are given by

$$\eta_i = \log \frac{p_i}{p_n},\ i = 1, \cdots, n-1.$$

From this, we obtain

$$\begin{aligned} p_i &= \frac{\exp(\eta_i)}{1 + \sum \exp(\eta_j)},\ i = 1, \cdots, n-1 \\ p_n &= \frac{1}{1 + \sum_{j=1}^{n-1} \exp(\eta_j)}. \end{aligned}$$

Using the expression for p_n and comparing (5.1) with (4.2), we see that

$$
\begin{aligned}
b(\eta_1, \cdots, \eta_{n-1}) &= -m \log p_n \\
&= m \log(1 + \sum \exp_{j=1}^{n-1}(\eta_j)).
\end{aligned}
$$

6 Implementation

In this section, we present implementation details on the EM algorithm applied to the HME architecture.

Problem (3.12) is a maximum likelihood problem for a multinomial distribution. The IRLS discussed in Section 5 can be applied with the following setup

$$
\begin{aligned}
\eta_i^{(t)} &= \sum \boldsymbol{\beta}_i^{(1)\mathsf{T}} \mathbf{x}^{(t)}, \ i = 1, \cdots, n-1 \\
b(\eta_1, \cdots, \eta_{n-1}) &= m \log(1 + \sum_{j=1}^{n-1} \exp(\eta_j)) \\
z_i^{(t)} &= h_i^{(t)}, \ i = 1, \cdots, n-1
\end{aligned}
$$

where $\boldsymbol{\beta}^{(1)} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_{n-1})$.

Similarly, IRLS can be applied to problem (3.13) with the following setup. For each $1 \leq i \leq n$

$$
\begin{aligned}
\eta_{j|i}^{(t)} &= \sum \boldsymbol{\beta}_{j|i}^{\mathsf{T}} \mathbf{x}^{(t)}, \ j = 1, \cdots, m-1 \\
b(\eta_{1|i}, \cdots, \eta_{m-1|i}) &= m \log(1 + \sum_{j=1}^{m-1} \exp(\eta_{j|i})) \\
z_j^{(t)} &= h_{j|i}^{(t)}, \ j = 1, \cdots, m-1 \\
c^{(t)} &= h_i^{(t)}
\end{aligned}
$$

where $\boldsymbol{\beta}^{(2,i)} = (\boldsymbol{\beta}_{1|i}, \cdots, \boldsymbol{\beta}_{m-1|i})$.

Problem (3.14) is also a weighted maximum likelihood problem and thus IRLS also applies. In this section, however, we look into a special case in which it is equivalent to a weighted least squares problem. Namely, if the expert distributions f_{ij} are normal, (3.14) is equivalent to

$$\hat{\boldsymbol{\beta}}_{ij} = \underset{\boldsymbol{\beta}_{ij}}{\operatorname{argmax}} \sum_t h_{ij}^{(t)} (y^{(t)} - \boldsymbol{\beta}_{ij}^\intercal x^{(t)})^2$$

$$\hat{\sigma}_{ij}^2 = \frac{\sum_t h_{ij}^{(t)} (y^{(t)} - \hat{\boldsymbol{\beta}_{ij}}^\intercal x^{(t)})^2}{\sum_t h_{ij}^{(t)}}$$

which can be solved with any least squares algorithm.

An implementation with normal experts is given in

`https://github.com/yuki172/hme`

References

[1] M. Jordan and R. Jacobs, "Hierarchical mixtures of experts and the em algorithm," *Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan)*, vol. 2, pp. 1339–1344 vol.2, 1993.

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York, NY, USA: Springer New York Inc., 2001.