Mixture of Regression Experts in fMRI Encoding

Subba Reddy Oota*[1], Adithya Avvaru[1], Naresh Mawani[1], Raju S. Bapi[1,2]
[1] IIIT Hyderabad, [2] University of Hyderabad
{oota.subba,adithya.avvaru}@students.iiit.ac.in, {naresh.manwani,raju.bapi}@iiit.ac.in

Abstract

fMRI semantic category understanding using linguistic encoding models attempt to learn a forward mapping that relates stimuli to the corresponding brain activation. Classical encoding models use linear multi-variate methods to predict the brain activation (all voxels) given the stimulus. However, these methods essentially assume multiple regions as one large uniform region or several independent regions, ignoring connections among them. In this paper, we present a mixture of experts-based model where a group of experts captures brain activity patterns related to particular regions of interest (ROI) and also show the discrimination across different experts. The model is trained word stimuli encoded as 25-dimensional feature vectors as input and the corresponding brain responses as output. Given a new word (25-dimensional feature vector), it predicts the entire brain activation as the linear combination of multiple experts' brain activations. We argue that each expert learns a certain region of brain activations corresponding to its category of words, which solves the problem of identifying the regions with a simple encoding model. We showcase that proposed mixture of experts-based model indeed learns region-based experts to predict the brain activations with high spatial accuracy.

1 Introduction

In recent years, the use of both linear and non-linear multivariate encoding/decoding approaches for analyzing fMRI brain activity have become increasingly popular (Mitchell et al., 2008; Naselaris et al., 2011; Mesgarani et al., 2014; Di Liberto et al., 2015; Pereira et al., 2018). An encoding model that predicts brain activity in response to stimuli is important for neuroscientists who can use the model predictions to investigate and test hypotheses about the transformation from stimulus to brain response in patients. Recent approaches for modeling fMRI data use word embedding representations to build encoding systems (Oota et al., 2018; Abnar et al., 2018). However, all these models do not have a principled way of learning regions specializing in particular category of stimuli. Instead, they simply predict the voxel intensity values either for the whole brain or for a pre-selected set of voxels. In particular, interpreting the non-linear multivariate models can be difficult due to capturing of unsuspected and enigmatic representations (Benjamin et al., 2017; Kording et al., 2018).

Mixture of experts (MoE) models, proposed by Jordan's group (Jordan & Xu, 1995) offer an interesting choice for the problem of learning distinct models for different regions of the input space. MoE were applied to MRI data to investigate the complex patterns of brain changes associated with non-pathological and pathological processes, such as the effects of growth, aging, injury or a disease (Kim et al., 2010; Eavani et al., 2016). MoEs have great potential for use in medical diagnostics to diagnose a variety of clinical conditions such as depression, Alzheimer's dementia etc. (Yao et al., 2009) proposed a Hidden Conditional Random Framework (HCRF) inspired from a mixture of experts model to make predictions in all ROIs which are interconnected.

In this paper, we present a mixture of regression experts (MoRE) model that predicts the brain activity associated with concrete nouns as stimuli. The theory underlying this model is that each expert learns
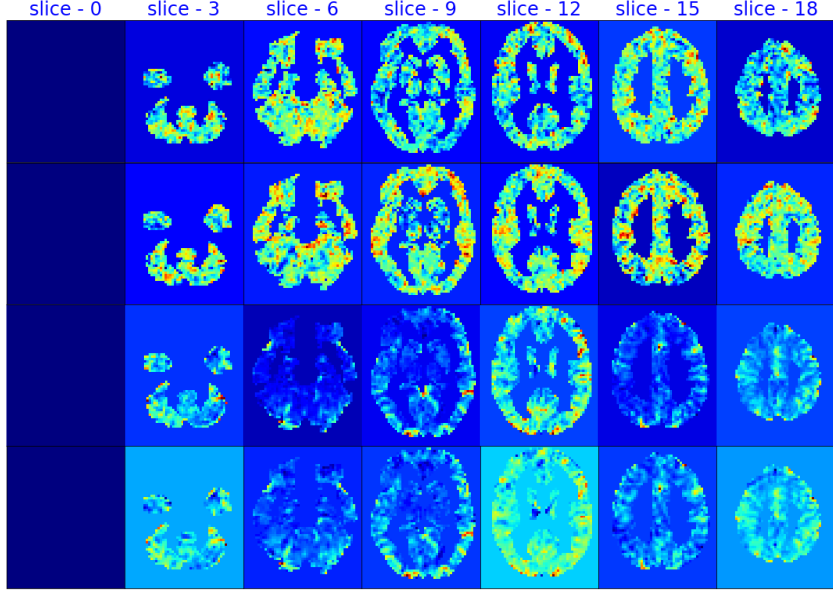
Figure 1: The sequence of slices, for the word "Apartment" after converting voxel activation per subject into 23 slices, (i) in actual brain activation (top row), (ii) for the model trained with mixture of experts model (second row from top), (iii) for the model trained with multilayer perceptron model (third row from top), (iv) for the model trained with ridge regression (bottom row).

to cover a certain brain region of interest (set of voxels that are significantly activated) based on the category of words that are represented by the model. Finally, fMRI signals are modeled as a linear combination of all expert predictions corresponding to the gate probabilities of each expert. We present an experimental evidence showing that the best encoding model is achieved with the mixture of regression experts rather than using a simple linear/non-linear model alone.

2   fMRI Encoding: Mixture of Regression Experts (MoRE) Approach

Current encoding methods attempt to learn either weights in case of linear models (Mitchell et al., 2008; Pereira et al., 2018) or complex representations in non-linear models (Oota et al., 2018). In the next sections, we discuss the proposed MoRE approach and our enhancements.

2.1   Architecture:

We use a mixture of experts-based encoder model, whose architecture is inspired from Jordan & Xu (1995). The mixture of experts architecture is composed of a gating network and several expert networks, each of which solves a function approximation problem over a local region of the input space.

Let $D = \{X, Y\}$ denote the data where $X = \{x^n\}_{n=1}^N$ is the input, $Y = \{y^n\}_{n=1}^N$ is the target, K is number of experts and N is the number of training points. Also, let $\Theta = \{\theta_0, W_e, \Sigma_e\}$ denote the set of parameters where θ_0 is the set of gate parameters, W_e is the set of expert parameters, and Σ_e is the set of expert co-variance matrices.

The mixtures of experts model is based on the following conditional mixture

$$p(\mathbf{y}|\mathbf{x}) = \sum_{j=1}^K P(j|\mathbf{x})p(\mathbf{y}|\mathbf{x}, \theta_j) = \sum_{j=1}^K g_j(\mathbf{x}, \theta_0)p(\mathbf{y}|\mathbf{x}, \theta_j) \tag{1}$$

where $g_j(\mathbf{x}, \theta_0)$ is the gate's rating, i.e., the probability of the jth expert, given $\mathbf{x}$ and $p(\mathbf{y}|\mathbf{x}, \theta_j)$ denotes the density function for the output vector associated with the jth expert, i.e., the probability of jth expert generating $\mathbf{y}$ given $\mathbf{x}$.

Here, we choose $p(\mathbf{y}|\mathbf{x}, \theta_j)$ as multivariate Gaussian probability density for each of the experts, denoted by.

$$p(\mathbf{y}|\mathbf{x}, W_j, \Sigma_j) = \frac{1}{(2\pi)^{m/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - W_j\mathbf{x})^T \Sigma_j^{-1}(\mathbf{y} - W_j\mathbf{x})\right) \qquad (2)$$

where $W_j \in \mathbb{R}^{m \times n}$ is the weight matrix and $\Sigma_j \in \mathbb{R}^{m \times m}$ is the covariance matrix associated with the jth expert. In our problem, we consider two different cases of Σ_j.

1. $\Sigma_j = \sigma_j^2 I, \ \forall j \in [K]$ and
2. $\Sigma_j = diag(\sigma_{j,1}^2, \sigma_{j,2}^2, \ldots, \sigma_{j,m}^2), \ \forall j \in [K]$.

Thus, we assume that the components of the output vector $\mathbf{y} \in \mathbb{R}^m$ are statistically independent of each other. We use this assumption to make the model very simple and reduce the number of overall parameters. This assumption will also make the algorithm computationally less expensive.

Assuming $\Sigma_j = diag(\sigma_{j,1}^2, \sigma_{j,2}^2, \ldots, \sigma_{j,m}^2)$, we rewrite the conditional probability density model for j^{th} expert as follows:

$$P(\mathbf{y}|\mathbf{x}, W_j, \Sigma_j) = \frac{1}{(2\pi)^{m/2}(\sigma_{j,1}\sigma_{j,2}..\sigma_{j,m})} \exp\left(-\sum_{i=1}^{m} \frac{(y_i - \mathbf{w}_{j,i}^T\mathbf{x})^2}{2\sigma_{j,i}^2}\right) \qquad (3)$$

where $\mathbf{w}_{j,i}$ is the ith row of W_j.

We use softmax function for $g(\mathbf{x}, \theta_0)$.

$$g_j(\mathbf{x}, \theta_0) = \frac{\exp\left(\mathbf{v}_j^T\mathbf{x}\right)}{\sum_{i=1}^{K} \exp\left(\mathbf{v}_i^T\mathbf{x}\right)} \qquad (4)$$

where $\mathbf{v}_j \in \mathbb{R}^n, \ \forall j \in [K]$. Thus, $\theta_0 = \{\mathbf{v}_1, \ldots, \mathbf{v}_K\}$.

3 Experiments

Dataset We used data from fMRI experiment 1 (Mitchell et al., 2008), where authors conducted experiments with multiple subjects by showing different categories of words as stimuli. Each category might correspond to activation of distinct brain regions. In the experiment, the target word was presented with a picture that depicted some aspect(s) of the relevant meaning. This fMRI dataset was collected from a total of 9 participants. For each participant in the experiment, a total set of 60 words (12 categories) were used as stimuli in multimodal form (word, picture). The fMRI dataset constitutes 51×61 voxel windows arranged as 23 slices, per subject per stimulus. We use the publicly available Mitchell's 25-feature vector data (Mitchell et al., 2008) of 60 words as input and the corresponding brain response of each participant (containing 21000 voxels) as output to train the model.
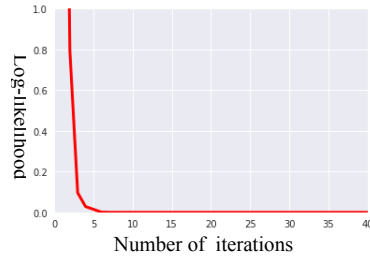

Figure 2: MoRE log-likelihood convergence

Selection of Number of Experts: In this paper, we selected the number of experts based on the number of categories considered in the dataset. The dataset is divided into 12 categories in which each category has 5 concrete nouns. The idea is that each expert should master a particular category and the corresponding activation in relevant brain regions. The detailed selecting number expert models using Bayesian Information Criteria provided in Appendix A.

Description of the Experiment Here, we considered Mitchell's feature vector, $\mathbf{x}$ as input to all the experts and calculated $p(\mathbf{y}|\mathbf{x}, \theta_j)$ where $\mathbf{y}$ is the corresponding brain responses for that $\mathbf{x}$. Gating network trains to pick the expert which gives maximum $g_j(\mathbf{x}, \theta_0)$. The detailed training methodology is provided in Appendix B.
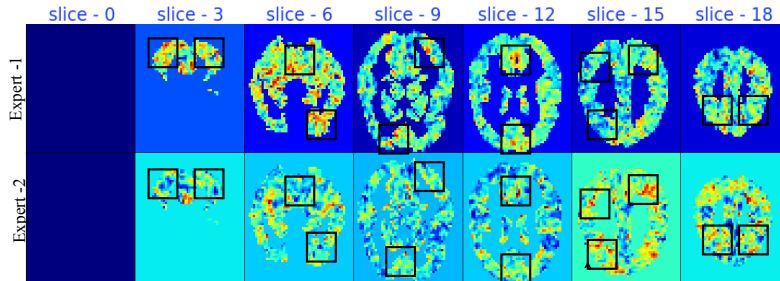
3

Figure 3: Comparison of brain activity (ROIs) in two experts across slices. The squared box indicates the ROIs' activation differences: i) Expert-1 has higher brain activity in "cingulate region" and "Superior and Middle frontal gyrus regions"(top row), ii) ROIs captured in Expert2 are "occipital lobe" from visual region", "Inferior temporal gyrus" and "Middle temporal gyrus" (bottom row)

Method	Subjects (r^2 score)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Mixture of Regression Experts	0.811	0.813	0.806	0.811	0.814	0.815	0.813	0.808	0.813
Multilayer Perceptron	0.755	0.745	0.745	0.772	0.75	0.765	0.756	0.761	0.751
Ridge Regression	0.772	0.769	0.758	0.776	0.778	0.782	0.776	0.770	0.771

Table 1: Subject-wise r^2 scores for different models

Results and Discussion   Using the approach discussed in Section 2 and Appendix B, we trained separate mixture of regression experts model for each subject. The encoding performance was evaluated by training and testing models using different subsets of the 60 concrete words in a 5-fold cross-validation scheme. The encoder models were trained until the model reached convergence as shown in Figure 2. Figure 3 shows the region of activity in two different experts for the word "Church". Across all the slices, there is a considerable number of ROI differences among all the experts. In particular, we depict regions of activation for two experts in Figure 3. Figure 3 shows that Expert 1 has higher brain activity in "cingulate region" and "Superior and Middle frontal gyrus regions". In Expert 2, the ROIs captured are "Occipital lobe" from visual region", "Inferior temporal gyrus" and "Middle temporal gyrus". It appears that while Expert 1 mediates activation in the frontal regions usually involved in decision making, Expert 2 seems to activate regions responsible for visual processing and object recogntion. In addition, we observed some brain activity being common to ROIs across all the experts (results not shown here). In order to assess the similarity between the actual and predicted activation in each brain slice, we compared slice-wise voxel coordinates and their intensity values. We measured the model scores using predictive r^2 score, for voxel v as follows: $r_v^2 = 1 - \frac{||\mathbf{y}_{test}^v - \mathbf{y}_{pred}^v||^2}{||\mathbf{y}_{test}^v - mean(\mathbf{y}_{test}^v)||^2}$, where $\mathbf{y}_{test}^v$ is the testing set brain activations, $\mathbf{y}_{pred}^v$ is the predicted brain activations and $mean(\mathbf{y}_{test}^v)$ is the mean activation on the test set. Table 1 depicts subject-wise (9 subjects) r^2 score comparisons among the mixture of experts model, the model trained with the multilayer perceptron and with the ridge regression model.From Table 1, we observe that r^2 score for mixture of experts model is better than that of the other two models and the score range indicates that MoRE model typically yields better predictions.

4   Conclusion

In this paper, we present a mixture of experts-based model where group of experts capture brain activity patterns related to particular regions of interest (ROIs) and also show the discrimination across different experts. Different from previous work, the underlying model depicts that each expert trains on certain brain regions of interest (set of voxels that are significantly activated) based on the category of words that are represented by the model. The key distinction of our work is the utilization of experts model to capture the ROIs based on various categories of input. In future, we plan to experiment on all paradigms and conduct experiments on other fMRI datasets and identify the number of experts using Bayesian Information Criterion (BIC), with a primary focus on hierarchical mixture of experts at slice-level instead of voxel-level predictions.

References

Samira Abnar, Rasyan Ahmed, Max Mijnheer, and Willem Zuidema. Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pp. 57–66, 2018.

Ari S Benjamin, Hugo L Fernandes, Tucker Tomlinson, Pavan Ramkumar, Chris VerSteeg, Lee Miller, and Konrad Paul Kording. Modern machine learning far outperforms glms at predicting spikes. *BioRxiv*, pp. 111450, 2017.

Giovanni M Di Liberto, James A O'Sullivan, and Edmund C Lalor. Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, 25(19):2457–2465, 2015.

Harini Eavani, Meng Kang Hsieh, Yang An, Guray Erus, Lori Beason-Held, Susan Resnick, and Christos Davatzikos. Capturing heterogeneous group differences using mixture-of-experts: Application to a study of aging. *Neuroimage*, 125:498–514, 2016.

Michael I Jordan and Lei Xu. Convergence results for the em approach to mixtures of experts architectures. *Neural Networks*, 8(9):1409–1431, 1995.

Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

Seyoung Kim, Padhraic Smyth, and Hal Stern. A bayesian mixture approach to modeling spatial activation patterns in multisite fmri data. *IEEE Transactions on Medical Imaging*, 29(6):1260–1274, 2010.

Konrad Paul Kording, A Benjamin, Roozbeh Farhoodi, and Joshua I Glaser. The roles of machine learning in biomedical science. In *Frontiers of Engineering: Reports on Leading-Edge Engineering from the 2017 Symposium*. National Academies Press, 2018.

Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F Chang. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010, 2014.

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195, 2008.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

Subba Reddy Oota, Naresh Manwani, and Raju S. Bapi. fmri semantic category decoding using linguistic encoding of word embeddings. *ArXiv Preprint ArXiv:1806.05177*, 2018.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1):963, 2018.

G Schwarz. Estimating the dimension of a model the annals of statistics 6 (2), 461–464. *URL: http://dx. doi. org/10.1214/aos/1176344136*, 1978.

Bangpeng Yao, Dirk Walther, Diane Beck, and Li Fei-Fei. Hierarchical mixture of classification experts uncovers interactions between brain regions. In *Advances in Neural Information Processing Systems*, pp. 2178–2186, 2009.

A Selecting Number of Experts using BIC

In order to find the number of experts, Bayesian Information Criteria (BIC) is one of the successful measure to approximate the Bayes factor Kass & Raftery (1995) i.e to find a model that has maximum posterior probability or maximum marginal likelihood. The BIC can be formulated as follows.

$$BIC = d \log(n) - 2 \log(p(X|\Theta)) \tag{5}$$

in Schwarz (1978), where d is the dimensionality of the parameter space, n is the number of data points, $p(X|\Theta)$ denotes the parametric model density.

B Training Mixture of Experts using EM Algorithm

Let N be the number of samples, K be the number of experts and $S = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\}$ be the training set where $(\mathbf{x}^n, \mathbf{y}^n) \in \mathbb{R}^k \times \mathbb{R}^m, \ \forall n \in [N]$.

The EM algorithm is an iterative method for finding the maximum likelihood (ML) of a probability model in which some random variables are observed and others are hidden. In training the Mixture of Experts, the indicator variables $Z = z_j^n, \ j \in [K], \ n \in [N]$ are introduced to solve the model with the EM algorithm.

$$z_j^n = \begin{cases} 1, \text{if } \mathbf{y}^n \text{ is generated from the } j^{th} \text{ expert} \\ 0, \text{otherwise} \end{cases}$$

With the indicator variables, the likelihood can be written as

$$L(\Theta; S; Z) = P(S, Z|\Theta) = \prod_{n=1}^{N} P(\mathbf{y}^n|\mathbf{x}^n)$$

$$= \prod_{n=1}^{N} \sum_{j=1}^{K} [g_j(\mathbf{x}^n, \theta_0) P(\mathbf{y}^n|\mathbf{x}^n, W_j, \Sigma_j)]^{z_j^n} \tag{6}$$

The parameter Θ can be estimated by the iterations through the E and M steps.

B.0.1 E-Step : Calculation of Expectation of log-likelihood

The EM algorithm is employed to average out z_j and maximize the expected complete data log-likelihood $E_Z(log P(D, Z|\Theta))$ and $\sum_{j=1}^{K} z_j^n = 1$, for each n. The expectation of the log-likelihood in (5) results in

$$Q(\Theta|\Theta^{(p)}) = E_Z(\log P(D, Z|\Theta))$$

$$= \sum_{n=1}^{N} \sum_{j=1}^{K} h_j^{(p)}(n) \{ log([g_j(\mathbf{x}^n, \theta_0) P(\mathbf{y}^n|\mathbf{x}^n, W_j, \Sigma_j)]) \}$$

$$= \sum_{n=1}^{N} \sum_{j=1}^{K} h_j^{(p)}(n) \log(g_j(\mathbf{x}^n, \theta_0)) + \sum_{n=1}^{N} \sum_{j=1}^{K} h_j^{(p)}(n) \log(P(\mathbf{y}^n|\mathbf{x}^n, W_j, \Sigma_j))) \tag{7}$$

where p is the iteration index and $h_j^{(p)}(n)$ is expectation of indicator variables which is given by

$$h_j^{(p)}(n) = E[z_j^n|D]$$

$$= P(z_j^n = 1|\mathbf{x}^{(n)}, \mathbf{y}^{(n)})$$

$$= \frac{g_j(\mathbf{x}^{(n)}, \theta_0) P(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}, W_j, \Sigma_j)}{\sum_{i=1}^{K} g_i(\mathbf{x}^{(n)}, \theta_0) P(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}, W_i, \Sigma_i)} \tag{8}$$

B.0.2 M-Step : Estimation of new parameters $\Theta^{(p+1)}$

- From E-step, we calculated the expectation of log-likelihood Q, we now investigate the implementation of M-step by estimating the new parameters.

- The M step chooses a parameter Θ that increases the Q function, the expected value of the complete-data log-likelihood i.e.

$$\Theta^{(p+1)} = \underset{\Theta}{\operatorname{argmax}} \, Q(\Theta|\Theta^{(p)}) \tag{9}$$

- An iteration of EM also increases the original log likelihood L. That is $L(\Theta^{(p+1)}; Y) > L(\Theta^{(p)}; Y)$ i.e. the likelihood L increases monotonically along the sequence of parameter estimates generated by an EM algorithm.

Update Gate Parameters

- From the Equations (4) and (6)

$$E_{gate} = -\sum_{t=1}^{N}\sum_{j=1}^{K} h_j^{(p)}(n) log(g_j(\mathbf{x}^{(n)}, \theta_0) \tag{10}$$

- An interpretation of the cost function in Equation (6) is the cross-entropy of distributing a pattern $\mathbf{x}$ amongst the expert networks. This cost is minimal if experts are mutually exclusive and increases when experts share a pattern.

$$
\begin{aligned}
\frac{\partial Q}{\partial \theta_0} &= \sum_{n=1}^{N}\sum_{j=1}^{K} h_j^{(p)}(n)\frac{\partial g_j(\mathbf{x}^{(n)}, \theta_0)}{\partial \theta_0} / g_j(\mathbf{x}^{(n)}, \theta_0) \\
&= \sum_{n=1}^{N}\sum_{j=1}^{K} h_j^{(p)}(n)[\frac{\partial(\mathbf{v}_j^T\mathbf{x}^n)}{\partial \theta_0} - \sum_{i=1}^{K} g_i(\mathbf{x}^{(n)}, \theta_0)\frac{\partial(\mathbf{v}_i^T\mathbf{x}^n)}{\partial \theta_0}] \\
&= \sum_{n=1}^{N}\sum_{j=1}^{K}[h_j^{(p)}(n) - g_j(\mathbf{x}^{(n)}, \theta_0)]\frac{\partial(\mathbf{v}_j^T\mathbf{x}^n)}{\partial \theta_0}
\end{aligned}
\tag{11}
$$

As the gating network is a specific form of a generalized linear model, the general update for θ_0 is written as follows

$$\theta_0^{(p+1)} = \theta_0^{(p)} + \eta_g\frac{\partial Q}{\partial \theta_0} \tag{12}$$

- Calculate the log likelihood using both $\theta_0^{(p)}, \theta_0^{(p+1)}$ in Equation (5), and the difference in log likelihood values is calculated as

$$\nabla L = L(\Theta^p; S; Z) - L(\Theta^{p+1}; S; Z) \tag{13}$$

- If ∇L is below a threshold, then new parameters are discarded and MOE model will use old parameters.

- There is no known closed-form way to find the minimum of cost function, and thus we'll resort to an iterative optimization algorithm such as gradient descent or L-BFGS.

Update Expert Parameters

- From the Equations (3) and (6)

$$E_{expert} = -\sum_{n=1}^{N}\sum_{j=1}^{K} h_j^{(p)}(n)log(P(\mathbf{y}^n|\mathbf{x}^n, W_j, \Sigma_j)) \tag{14}$$

- The update equation for expert parameters W_j as follows

$$\frac{\partial Q}{\partial W_j} = \sum_{n=1}^{N} \sum_{j=1}^{K} h_j^{(p)}(n) \frac{\partial P(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}, W_j)}{\partial W_j} / P(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}, W_j)$$

$$= \sum_{n=1}^{N} \sum_{j=1}^{K} h_j^{(p)}(n) \frac{\partial(W_j^T \mathbf{x}^{(n)})}{\partial W_j} \frac{[\mathbf{y}^{(n)} - W_j^T \mathbf{x}^{(n)}]}{\Sigma_j}$$

$$= \sum_{n=1}^{N} \sum_{j=1}^{K} h_j^{(p)}(n) \mathbf{x}^{(n)} \frac{[\mathbf{y}^{(n)} - W_j^T \mathbf{x}^{(n)}]}{\Sigma_j} \qquad (15)$$

$$W_j^{(p+1)} = W_j^{(p)} + \eta_j \frac{\partial Q}{\partial W_j} \qquad (16)$$

- Calculate the log likelihood using both $W_j^{(p)}, W_j^{(p+1)}$ in Equation (5), and the difference in log likelihood values is calculated as

$$\nabla L = L(\Theta^p; S; Z) - L(\Theta^{p+1}; S; Z) \qquad (17)$$

- If ∇L is below a threshold, then new parameters are discarded and MOE model will use old parameters.

Update Expert Covariance Matrix

- The update equation for covariance matrix as follows.

$$\Sigma_j^{(p+1)} = \frac{1}{\sum_{n=1}^{N} h_j^{(p)}(n)} \sum_{n=1}^{N} h_j^{(p)}(n)[\mathbf{y}^{(n)} - W_j^T \mathbf{x}^{(n)}]^2 \qquad (18)$$

- Calculate the log likelihood using both $\Sigma_j^{(p)}, \Sigma_j^{(p+1)}$ in Equation (5), and the difference in log likelihood values is calculated as

$$\nabla L = L(\Theta^p; S; Z) - L(\Theta^{p+1}; S; Z) \qquad (19)$$

- If ∇L is below a threshold, then new parameters are discarded and MOE model will use old parameters.