Citation for published version:
Drugowitsch, J & Barry, A 2007, *Generalised mixtures of experts, independent expert training, and learning classifier systems*. Computer Science Technical Reports, no. CSBU-2007-02, University of Bath, Department of Computer Science.

Publication date:
2007

Link to publication

©The Author April 2007

University of Bath

Alternative formats
If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 14. Apr. 2021

Department of Computer Science

Technical Report

Generalised Mixtures of Experts, Independent Expert Training, and Learning Classifier Systems

Jan Drugowitsch and Alwyn Barry

Technical Report 2007-02 **April 2007**

ISSN 1740-9497

Copyright ©April 2007 by the authors.

Contact Address:
Department of Computer Science
University of Bath
Bath, BA2 7AY
United Kingdom
URL: http://www.cs.bath.ac.uk

ISSN 1740-9497

Generalised Mixtures of Experts, Independent Expert Training, and Learning Classifier Systems

Jan Drugowitsch
J.Drugowitsch@bath.ac.uk

Alwyn M. Barry
A.M.Barry@bath.ac.uk

April 2007

Abstract

We present a generalisation to the Mixtures of Experts model that introduces prior localisation of the experts as part of the model structure, and as such relates them strongly to the evolutionary computation ML method known as Learning Classifier Systems. While the introduced generalisation allows specification of more complex localisation patterns, identifying good models becomes more difficult. We approach this tradeoff by introducing a new training schema that makes fitting a single model computationally less demanding and shifts the importance to searching the space of possible model structures, guided by approximate variational Bayesian inference to fit the model and find the model evidence. We demonstrate model search for simple non-linear curve fitting tasks by sampling from the model posterior, as a proof-of-concept alternative to the genetic algorithm used in Learning Classifier Systems for that purpose.

1 Introduction

While Learning Classifier Systems (LCS) are well established in the field of evolutionary computation, and perform competitively in regression and classification tasks when compared to other ML methods (Bernadó-Mansilla et al., 2002; Butz et al., to appear), their performance lacks explanation due to their purely algorithmic description that does not identify the underlying data model. This makes them hard to access and hinders their further improvement.

In this work we for the first time clearly identify the model underlying LCS by linking them to a generalisation of the Mixtures of Experts (MoE) model (Jacobs et al., 1991; Jordan & Jacobs, 1994). The latter method is well established in the machine learning literature, and trains a set of local models, called *experts*, by localising them in the input space with a gating network. This localisation is achieved by the interdependent training of gating network and experts, resulting in a smoothed linear partitioning of the input space. A variant on MoE replaces the linear partitioning by normalised Gaussian localisation, but with the same underlying principle (Xu et al., 1995).

In contrast, LCS add an additional layer of "forced" localisation (determined by the model structure) to each expert[1], where the shape of localisation is only limited by the choice of representation, allowing the capture of a richer data dependency structure and natural integration of nominal, ordinal and metric data attributes. This freedom comes at the cost of making them harder to train, as one needs to search the potentially complex space of possible expert localisations (that is, the space of possible model structures) in addition to training the model for a fixed set of localisations. While LCS acquire a more efficient model fitting scheme for this task, and search the space of model structures with a genetic algorithm, we describe this fitting scheme and its implications, but leave the adaptation of the genetic algorithm to the presented model as future work. Rather, we perform a less powerful model search by sampling from the model posterior as a proof-of-concept[2].

To assess the quality of a certain model structure in explaining the given data, we use the model posterior approximated by variational inference on a full Bayesian LCS model, closely related to similar approaches for MoE by Waterhouse et al. (1996), Ueda and Ghahramani (2002) and Bishop and Svensén (2003). Such a principled approach is new to LCS as their lack of model description requires LCS to fall back on a set of heuristics for model assessment.

In the next section we introduce the MoE and our generalisations to it, and discuss in Section 3 the requirement of a model selection mechanism that is realised by Bayesian means with approximate

[1] In LCS, the experts are called *classifiers*, despite them usually being regression models.
[2] For more information on LCS, the interested reader is referred to Butz (2006).

variational inference. Our method of searching the space of possible models and the required modifications to the fitting procedure are described in Section 4, and demonstrated by sampling from the model posterior in Section 5, after which we offer conclusions about our achievements.

2 Generalised Mixtures of Experts

Let us for now assume a fixed model $\mathcal{M}$ with K experts. After Jordan and Jacobs (1994), each expert provides a conditional probability distribution $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_k)$ over an output vector $\boldsymbol{y}$ given an input vector $\boldsymbol{x}$ and a set of parameters $\boldsymbol{\theta}_k$ of the kth expert. In MoE these experts are combined by a gating network to form the mixture distribution $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$.

The gating network is best explained from the generative point-of-view. The random gating vector $\boldsymbol{z} = (z_1, \ldots, z_K)^T$ of binary latent gating variables z_k determines which expert generates the observation $\{\boldsymbol{x}, \boldsymbol{y}\}$. An observation is always assigned to a single expert, and so $\boldsymbol{z}$ has a 1-of-K structure, where $g_k(\boldsymbol{x}) \equiv p(z_k = 1|\boldsymbol{x}, \boldsymbol{v}_k)$ denotes the probability of expert k generating $\{\boldsymbol{x}, \boldsymbol{y}\}$, and $\boldsymbol{v}_k$ is the parameter vector for g_k.

2.1 Localisation through Matching

While in the original MoE model there are no prior restrictions on the association of observations to experts, we introduce such a restriction by the concept of matching: an additional binary random variable m_k per expert restricts the set of observations that this expert can have generated to the set that it *matches*, that is, for which $m_k = 1$. This corresponds to each expert in the LCS being restricted to modelling a subset of the input space, and so matching is a property of the model $\mathcal{M}$ that is specified by the matching function $m_k(\boldsymbol{x}) \equiv p(m_k = 1|\boldsymbol{x})$ and remains unchanged during the model fitting process. Hence, the model structure is fully specified by the number of experts K and their matching functions $\boldsymbol{M} = \{m_k\}$, that is, $\mathcal{M} = \{K, \boldsymbol{M}\}$.

We enforce matching by defining z_k conditioned on m_k to be

$$p(z_k = 1|m_k, \boldsymbol{x}, \boldsymbol{v}_k) \propto \begin{cases} \exp(\boldsymbol{v}_k^T \vartheta(\boldsymbol{x})) & \text{if } m_k = 1, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

effectively modelling the generative probability for expert k by a generalised linear model of $\vartheta(\boldsymbol{x})$ if it matches, and locking it to 0 if it does not. The transfer function ϑ over the input vectors is an additional generalisation over the original MoE model, which currently uses $\vartheta(\boldsymbol{x}) = \boldsymbol{x}$.

To get the generating probability g_k we marginalise over all m_k and add the normalising term, resulting in

$$\begin{aligned} g_k(\boldsymbol{x}) &= \sum_{m \in \{0,1\}} p(z_k = 1, m_k = m|\boldsymbol{x}, \boldsymbol{v}_k) \\ &= \frac{m_k(\boldsymbol{x}) \exp(\boldsymbol{v}_k^T \vartheta(\boldsymbol{x}))}{\sum_j m_j(\boldsymbol{x}) \exp(\boldsymbol{v}_j^T \vartheta(\boldsymbol{x}))}. \end{aligned} \tag{2}$$

Note that this is the well-known softmax function with augmenting matching functions m_k, which weights the output of the generalised linear model by the degree of matching of the corresponding expert. While we have generalised the MoE model by introducing expert localisation through matching and the additional moderating function ϑ, its original formulation can be recovered by simply setting $m_k(\boldsymbol{x}) = 1$ and $\vartheta(\boldsymbol{x}) = \boldsymbol{x}$ for all $\boldsymbol{x}$ and k, which causes each expert to match all observations.

2.2 The Data Likelihood

Given an input $\boldsymbol{x}$, we stochastically choose among the experts that match this input to generate the observed output $\boldsymbol{y}$. This becomes more obvious when using the 1-of-K structure of $\boldsymbol{z}$ to write

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\theta}) = \prod_{k=1}^{K} p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_k)^{z_k}. \tag{3}$$

Here we do not need to explicitly consider matching, as by Eq. (1) we can only have $z_k = 1$ if $m_k = 1$.

To get the likelihood of a single observation $\{\boldsymbol{x}, \boldsymbol{y}\}$ we marginalise over the gating variables $\boldsymbol{z}$ to obtain

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) &= \sum_{\boldsymbol{z}} \prod_{k=1}^{K} p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_k)^{z_k} \prod_{j=1}^{K} p(z_j = 1|\boldsymbol{x}, \boldsymbol{v}_j)^{z_j} \\ &= \sum_{k=1}^{K} g_k(\boldsymbol{x}) p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_k). \end{aligned} \tag{4}$$

This shows that the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$ is a mixture distribution of the expert models weighted by the gating network.

For a set of N i.i.d. inputs $\boldsymbol{X} = \{\boldsymbol{x}_n\}$ and the corresponding outputs $\boldsymbol{Y} = \{\boldsymbol{y}_n\}$, the likelihood is given by

$$p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\theta}) = \prod_{n} p(\boldsymbol{y}_n|\boldsymbol{x}_n, \boldsymbol{\theta}). \tag{5}$$

where every input/output pair has a set of latent gating variables $\boldsymbol{z}_n = \{z_{nk}\}$ associated with it.

2.3 Model Training

In the original MoE model, the parameters of the experts and the gating network are found by maximising the likelihood by use of the EM-algorithm, where in the E-step one finds the posterior over the gating variables $\{z_n\}$ and in the M-step the complete-data likelihood is maximised with respect to the expert and gating parameters. Independent of the type of experts, the gating parameters are trained by the iteratively re-weighted least-squares (IRLS) algorithm, which can be easily adjusted to handle Eq. (2) of the generalised model.

However, the power of the generalisation lies in the localisation of the experts expressed by the matching function. Thus, in addition to fitting the model to the data, we also want to identify adequate localisation of the experts. As this localisation is fixed for a single model $\mathcal{M}$, we need to be able to i) compare different models and to ii) efficiently search for better ones.

The problem with the maximum likelihood approach is its tendency to over-fit the data, which causes it to always prefer more complex models over simpler ones, making it inappropriate for model selection. Embedding the model into a Bayesian framework, on the other hand, avoids over-fitting and allows us to derive an expression for the probability of a model given the data, which we can use to compare different models. Even though a fully Bayesian treatment of the generalised MoE is not analytically tractable, we will describe in the next section how to use an approximate variational inference similar to Waterhouse et al. (1996) in order to get a quick and sufficiently accurate estimate of model fit and evidence.

While this allows us to evaluate the quality of a single model structure with respect to the data, we also need to find a good model structure by searching the potentially complex space $\{\mathcal{M}\}$. LCS approach this by evaluating the current model structure and stochastically adding or removing experts or changing their matching function, thus traversing the space of possible models. This requires a frequent re-evaluation of the model evidence for a changed model structure, and how to achieve this efficiently is discussed in Section 4.

3 A Bayesian Approach

The Bayesian approach tries not to estimate the parameters by maximum likelihood, but rather by estimating a posterior distribution by conditioning a prior distribution on the available data. In that way it also mitigates over-fitting of the data because the model parameters are integrated out.

Additionally, if we consider the model structure $\mathcal{M}$ as a random variable, we can find the posterior $p(\mathcal{M}|\boldsymbol{X}, \boldsymbol{Y})$ and improve the model structure by maximising this posterior. In order to do so, we can observe that

$$p(\mathcal{M}|\boldsymbol{X}, \boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{X}, \mathcal{M}) p(\mathcal{M}), \tag{6}$$

and hence are only required to find the model evidence $p(\boldsymbol{Y}|\boldsymbol{X}, \mathcal{M})$ and specify some prior distribution $p(\mathcal{M})$ over the possible model structures. The latter can be used to deal with symmetries in the model parameterisation, as we will see in Section 5.

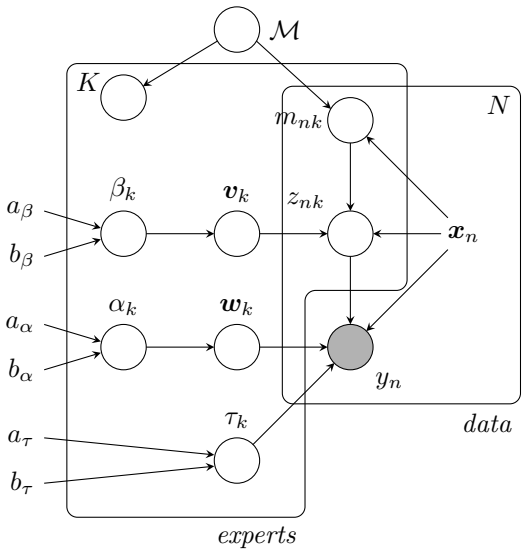


Figure 1: Directed Acyclic Graph showing the Bayesian generalised MoE. The circular nodes are random variables, which are observed when shaded. Labels without nodes are constants. The boxes are "plates", comprising replicas of the entities inside them. Note that K is a random variable that is conditional on the model structure $\mathcal{M}$.

3.1 Expert Structure and Priors

For the rest of the paper we will assume univariate linear regression experts, but the method can be easily adapted to multivariate regression or binomial/multinomial classification experts. Additionally, we assume that the input vector $\boldsymbol{x}$ is augmented by the constant term 1, and therefore automatically introduces the bias term in the regression model. The conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$ for expert k is given by

$$p(y|\boldsymbol{x}, \boldsymbol{\theta}_k) = \mathcal{N}(y|\boldsymbol{w}_k^T \boldsymbol{x}, \tau_k^{-1}), \tag{7}$$

where $\boldsymbol{\theta}_k = \{\boldsymbol{w}_k, \tau_k\}$, $\boldsymbol{w}_k$ is the weight vector, and τ_k is the noise precision (inverse variance) of the observations.

Adopting priors similar to Bishop and Svensén (2003), the prior on $\boldsymbol{w}_k$ is conjugate Gaussian, and given by

$$p(\boldsymbol{w}_k|\alpha_k) = \mathcal{N}(\boldsymbol{w}_k|0, \alpha_k^{-1}\boldsymbol{I}) \tag{8}$$

for each expert separately, where the precision hyper-parameter α_k determines the shrinkage on $\boldsymbol{w}_k$ and $\boldsymbol{I}$ is the identity matrix. Similarly, for parameters $\{\boldsymbol{v}_k\}$ of the gating network we define the Gaussian priors

$$p(\boldsymbol{v}_k|\beta_k) = \mathcal{N}(\boldsymbol{v}_k|0, \beta_k^{-1}\boldsymbol{I}), \tag{9}$$

with precision hyper-parameter β_k. The noise precisions $\boldsymbol{\tau} = \{\tau_k\}$ and hyper-parameters $\boldsymbol{\alpha} = \{\alpha_k\}$ and $\boldsymbol{\beta} = \{\beta_k\}$ get assigned conjugate Gamma priors, given by

$$p(\tau_k) = \text{Gam}(\tau_k|a_\tau, b_\tau), \tag{10}$$
$$p(\alpha_k) = \text{Gam}(\alpha_k|a_\alpha, b_\alpha), \tag{11}$$
$$p(\beta_k) = \text{Gam}(\beta_k|a_\beta, b_\beta). \tag{12}$$

We have used $a_\alpha = a_\beta = a_\tau = 10^{-2}$ and $b_\alpha = b_\beta = b_\tau = 10^{-4}$ to get sufficiently broad priors and hyper-priors. As the matching variables $\boldsymbol{M} = \{m_k\}$ are fixed for each model $\mathcal{M}$, we do not need to specify priors on them. The directed probabilistic graphical model that describes this Bayesian structure is shown in Figure 1.

While the model is not directly analytically tractable, we will follow the same approach as in Waterhouse et al. (1996) or Bishop and Svensén (2003), using variational Bayesian inference to find an approximate posterior and model evidence.

3.2 Variational Bayesian Inference

Our goal is on one hand to find the variational distribution $q(\boldsymbol{U})$ that approximates the true posterior $p(\boldsymbol{U}|\boldsymbol{Y})$ and on the other hand to get the model evidence $p(\boldsymbol{Y})$, where $\boldsymbol{U} = \{\boldsymbol{W}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{Z}, \boldsymbol{V}, \boldsymbol{\beta}\}$ is the set of hidden variables, and all distributions are implicitly conditional on $\boldsymbol{X}$ and $\mathcal{M}$. Variational Bayesian inference is based on the decomposition (Bishop, 2006)

$$\ln p(\boldsymbol{Y}) = \mathcal{L}(q) + \mathrm{KL}(q\|p), \tag{13}$$

$$\mathcal{L}(q) = \int q(\boldsymbol{U}) \ln \frac{p(\boldsymbol{U}, \boldsymbol{Y})}{q(\boldsymbol{U})} \mathrm{d}\boldsymbol{U}, \tag{14}$$

$$\mathrm{KL}(q\|p) = -\int q(\boldsymbol{U}) \ln \frac{p(\boldsymbol{U}|\boldsymbol{Y})}{q(\boldsymbol{U})} \mathrm{d}\boldsymbol{U}, \tag{15}$$

which holds for any choice of q. As the Kullback-Leibler divergence $\mathrm{KL}(q\|p)$ is always non-negative, and 0 if and only if $p(\boldsymbol{U}|\boldsymbol{Y}) = q(\boldsymbol{U})$, $\mathcal{L}(q)$ is a lower bound on $\ln p(\boldsymbol{Y})$ and only equivalent to the latter if $q(\boldsymbol{U})$ is the true posterior $p(\boldsymbol{U}|\boldsymbol{Y})$. Hence, we can approximate the posterior by maximising the lower bound $\mathcal{L}(q)$, which brings the variational distribution closer to the true posterior and at the same time gives us an approximation of the model evidence.

3.3 Factorised Distributions

To make this approach tractable, we need to choose a family of distributions $q(\boldsymbol{U})$ that gives an analytical solution. A frequently used approach (for example, (Bishop & Svensén, 2003; Waterhouse et al., 1996)) that is still sufficiently flexible to give a good approximation to the true posterior is to use the set of distributions that factorises with respect to disjoint groups $\boldsymbol{U}_i$ of variables

$$q(\boldsymbol{U}) = \prod_i q_i(\boldsymbol{U}_i), \tag{16}$$

which allows us to maximise $\mathcal{L}(q)$ with respect to each hidden variable separately while keeping the other ones fixed. This results in

$$\ln q_i^*(\boldsymbol{U}_i) = \mathbb{E}_{i \neq j}(\ln p(\boldsymbol{U}, \boldsymbol{Y})) + \text{const.}, \tag{17}$$

when maximising with respect to $\boldsymbol{U}_i$, where the expectation is taken with respect to all hidden variables except for $\boldsymbol{U}_i$ and the constant is the logarithm of the normalisation constant of q_i^*.

3.4 Handling the Softmax Function

If the model has a conjugate-exponential structure, Eq. (17) gives an analytical solution with a distribution form equal to the prior on the corresponding hidden variable. However, in our case the softmax function in Eq. (2) does not conform to the conjugate-exponential structure, and we need to deal with it separately. A possible approach is to replace the softmax function by an exponential lower bound on it, which consequently introduces additional variational variables with respect to which $\mathcal{L}(q)$ also needs to be maximised. This approach was followed in Bishop and Svensén (2003) for the logistic sigmoid function, but currently there is no known exponential lower bound function on the softmax besides a conjected one in Gibbs (1997)[3]. Alternatively, we can follow the approach taken in Waterhouse (1997), where $q_V^*(\boldsymbol{V})$ is approximated by a Laplace approximation $\tilde{q}_V^*(\boldsymbol{V})$. Despite such approximation invalidating the lower bound nature of $\mathcal{L}(q)$, we have chosen to use it due to the lack of better alternatives.

3.5 Update Equations and Model Posterior

To get the variational update equations, we need to evaluate Eq. (17) for each hidden variable in U separately, similarly to the derivations in Waterhouse et al. (1996; 1997) and (Ueda & Ghahramani, 2002). All update equations are derived in the appendix.

Similarly, we can find a closed-form expression for $\mathcal{L}(\tilde{q})$ by evaluating Eq. (14), where we can reuse many terms that have already been used for finding the variational update equations. The expression for $\mathcal{L}(\tilde{q})$ is derived similarly to those found in Ueda and Ghahramani (2002) and Bishop (2006) and

[3] A more general bound was recently developed in Wainwright et al. (2005), but its applicability still needs to be evaluated.

is presented in the appendix. As the variational bound $\mathcal{L}_{\mathcal{M}}(\tilde{q})$ for model $\mathcal{M}$ is an approximate lower bound on the logarithm of the model evidence $\ln p(\boldsymbol{Y}|\boldsymbol{X}, \mathcal{M})$, we can find a closed-form approximation to the model posterior Eq. (6) by

$$\ln \tilde{p}(\mathcal{M}|\boldsymbol{X}, \boldsymbol{Y}) \geq \mathcal{L}_{\mathcal{M}}(\tilde{q}) + \ln p(\mathcal{M}) + \text{const.}, \tag{18}$$

which we will use to compare different models with the aim of finding better model structures.

4 Searching the Model Structure Space

A model structure $\mathcal{M}$ comprises the number of experts K and their matching functions $\boldsymbol{M}$. While these functions could potentially be arbitrarily defined, they are usually parametric to keep the search tractable. Nonetheless, even in the light of such restrictions, model structure search is still a complex task. For example, given 10-dimensional input vectors and matching functions that have 2 scalar parameters (like some specification of an interval) per dimension we have a 21-dimensional, possibly multi-modal, model structure space. Hence, model structure search is potentially computationally expensive, which we need to counter-balance by making the evaluation of a single model structure cheaper.

Once a model structure $\mathcal{M} = (K, \boldsymbol{M})$ is specified, we need to find its posterior approximation $\tilde{p}(\mathcal{M}|\boldsymbol{X}, \boldsymbol{Y})$ by Eq. (18), which involves evaluating the variational bound $\mathcal{L}_{\mathcal{M}}(\tilde{q})$ and, hence, fitting the model to the data. In this process it is important to obtain good solutions to the variational equations by avoiding poor local minima, of which there are many in the MoE model (Bishop & Svensén, 2003). Repeated model fitting with randomised initial conditions is not an option in our case, as i) it is computationally too expensive, and ii) it does not guarantee finding the global optimum. Hence, we will use a different training strategy to bypass the issue of local optima in fitting the model.

4.1 Independent Expert Training

The MoE method achieves localisation of the experts by interdependent training of experts and the gating network. This becomes apparent by inspecting the maximised variational distribution after Eq. (17) for the expert weight vectors $\boldsymbol{W}$, which factorises with respect to the experts, and for expert k is given by

$$\begin{aligned} q^*_W(\boldsymbol{w}_k) &= \mathcal{N}(\boldsymbol{w}_k|\boldsymbol{w}^*_k, \boldsymbol{\Sigma}^*_k), \\ \boldsymbol{\Sigma}^*_k &= \left(\mathbb{E}_\alpha(\alpha_k)\boldsymbol{I} + \mathbb{E}_\tau(\tau_k)\sum_n r_{nk}\boldsymbol{x}_n\boldsymbol{x}_n^T\right)^{-1}, \\ \boldsymbol{w}^*_k &= \mathbb{E}_\tau(\tau_k)\boldsymbol{\Sigma}^*_k\sum_n r_{nk}\boldsymbol{x}_n y_n, \end{aligned} \tag{19}$$

where r_{nk} is the *responsibility* of expert k for observation n, given by $r_{nk} = \mathbb{E}_Z(z_{nk})$. Hence, the experts are trained by weighted linear regression with a Gaussian shrinkage prior, where the observations are weighted according to the expert's responsibilities. These are found by evaluating $q^*_Z(\boldsymbol{Z})$, and reveal that the gating network distributes the responsibilities according to the goodness-of-fit of the experts, which depends on the responsibilities of the last iteration. While these interdependencies make localisation of the experts possible, they are also the main cause of the highly multi-modal structure of the variational bound that we want to maximise.

In our generalisation of the MoE model we provide a second layer of localisation that is determined by the model structure rather than applied while fitting it to the data. Hence, we can remove the interdependence of expert and gating network training by replacing the responsibilities r_{nk} in the expert training with the matching functions $m_k(\boldsymbol{x}_n)$. From the generative point-of-view, this causes a shift from the assumption that each observation was generated by one and only one expert to the idea that the observation is generated by some mixture of independent processes, as a single observation can now be fully attributed to several experts at once (Note that matching does not underlie the 1-of-K assumption). As a result, the role of the gating network is now to combine the experts to best explain the set of observations, rather than to assign each observation to one and only one expert. This brings the whole model interpretation closer to ensemble methods.

For fitting a certain model structure the modification has the important consequence that the experts can now be trained independently of the gating network. Hence, if new experts are added in the process of searching for better model structures, only the new experts need to be fitted in addition

to updating the gating network. When removing experts, the only task is to re-fit the gating network. This reduces the computational complexity of evaluating a model structure immensely.

On the downside, the quality of the fit is reduced if many experts are localised in the same area of the input space. While interdependent expert and gating network training leads to at least a local optimum of the variational bound $\mathcal{L}_{\mathcal{M}}(q)$ of that model structure, removing the interdependency by fixing the responsibilities in expert training causes a worse fit due to the lack of feedback from the gating networks to the experts. Therefore, the modified training scheme shifts the emphasis from expert localisation by the gating network to localisation by the matching functions, and adds additional responsibility to the model search mechanism.

While the change in training policy causes a worse fit of the model when several experts are localised in the same regions of the input space, a locally mostly disjoint set of experts behaves similarly to the original MoE, but with higher degrees of freedom in how to choose the localisation.

4.2 Model Search in LCS

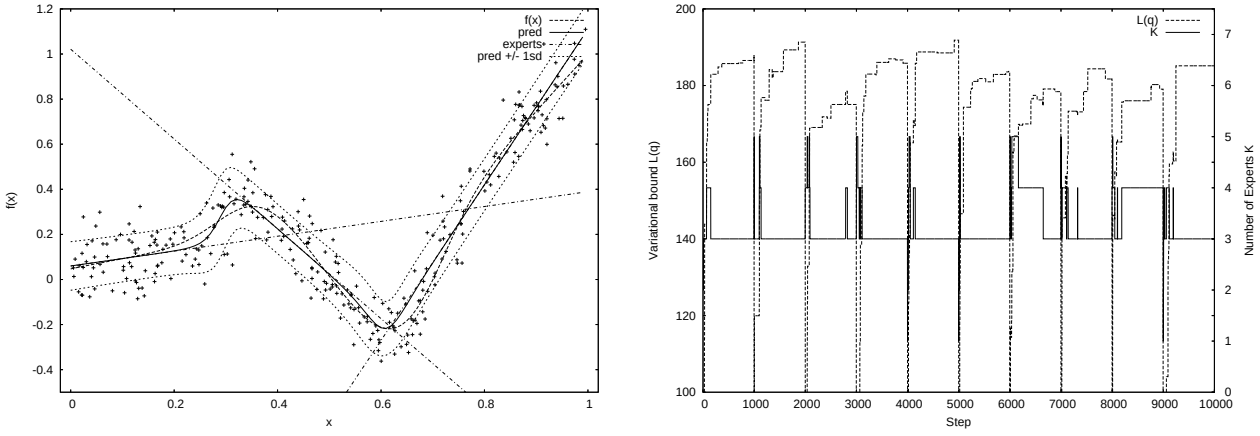
In Learning Classifier Systems, the space of possible models is searched by adding, removing or replacing single experts, based on some quality metric, such as, for example, the noise precision for regression experts. The space is traversed stochastically with a rather complex genetic algorithm that heuristically balances over- and underfitting of the experts and aims at finding an adequate coverage of the input space. The model quality is determined by the interaction between the algorithm and various system parameters, and there is currently no explicit quality metric for a model.

While we provide such a quality metric by the Bayesian model posterior, adjusting the genetic algorithm to use this metric is clearly outside of the scope of this paper, and a topic of further research. Instead, we provide a simple demonstration of how the model space can be searched by means of MCMC.

5 Sampling the Model Posterior

As a proof-of-concept on how to train a model and how to search the space of possible model structures, we have designed a simple Metropolis-Hastings algorithm that samples from the model posterior, similar to the one used for CART model search in Chipman et al. (1998).

5.1 Sampling by Metropolis-Hastings



(a) (b)

Figure 2: (a) shows the original function $f_1(x)$, generated by the mixture of 3 localised experts, and the data points used for training. The search method correctly identifies all 3 experts, which are shown by the dashed lines, with their localisation and mixed prediction. The error bars show one standard deviation from the model prediction. (b) shows how $\mathcal{L}(\tilde{q})$ and K change during the sampling process.

The algorithm starts with an initial model structure $\mathcal{M}_0$ at $t = 0$, and then proceeds by iterating over the following steps:

1. Sample a candidate model structure $\mathcal{M}^*$ from the Markov chain $p(\mathcal{M}^*|\mathcal{M}_t)$.

2. Accept the candidate model structure by setting $\mathcal{M}_{t+1} = \mathcal{M}^*$ with probability

$$\min\left(\frac{p(\mathcal{M}_t|\mathcal{M}^*)}{p(\mathcal{M}^*|\mathcal{M}_t)}\frac{\tilde{p}(\mathcal{M}^*|\boldsymbol{X},\boldsymbol{Y})}{\tilde{p}(\mathcal{M}_t|\boldsymbol{X},\boldsymbol{Y})}, 1\right), \tag{20}$$

where Eq. (18) is used to compute the model posteriors, or otherwise reject it by setting $\mathcal{M}_{t+1} = \mathcal{M}_t$.

Following this algorithm causes the sequence $\mathcal{M}_0, \mathcal{M}_1, \ldots$ to approach the approximated model posterior distribution $\tilde{p}(\mathcal{M}|\boldsymbol{X},\boldsymbol{Y})$, and therefore we will draw more samples from models with a high posterior probability.

We generate the Markov chain $p(\mathcal{M}^*|\mathcal{M}_t)$ by choosing randomly from one of the following actions: **Change:** Pick one expert uniformly at random and randomly re-initialise its matching function parameters; **Add:** Add one expert and randomly initialise its matching function parameters; **Remove:** Pick one expert uniformly at random and remove it. While these actions are rather simplistic and do not exploit the information that is available in the model, they were chosen such that the reverse transition probabilities $p(\mathcal{M}_t|\mathcal{M}^*)$ are easy to evaluate. In fact, for the *Change* action, the reverse transition equals the forward transition probability, and for the *Add* and *Remove* actions there is significant cancellation in the computation of the acceptance probability due to their complementary nature. In all experiments we have chosen *Add* and *Remove* with a probability of 1/4, and *Change* with a probability of 1/2.

The matching function we have used is a Gaussian basis function with a diagonal covariance structure to keep the model search space small. Its matching probability for an input $\boldsymbol{x}$ (not considering its bias term) is given by

$$m_k(\boldsymbol{x}) = \exp\left(-\sum_i \frac{1}{2\sigma_i^2}(x_i - \mu_i)^2\right), \tag{21}$$

where x_i is the ith component of $\boldsymbol{x}$, σ_i is the ith diagonal component of the covariance matrix, and μ_i is the ith component of the mean vector $\boldsymbol{\mu}$. Such a matching function introduces symmetries in the model space which we need to account for in the model prior $p(\mathcal{M})$ to avoid a bias towards models with a higher number of experts. Given, for example, a model with K experts, there are $K!$ possible permutations of the same model by rearranging the order of the experts. We have countered this effect by using the model prior $p(\mathcal{M}) \propto 1/K_{\mathcal{M}}!$ when evaluating Eq. (18).

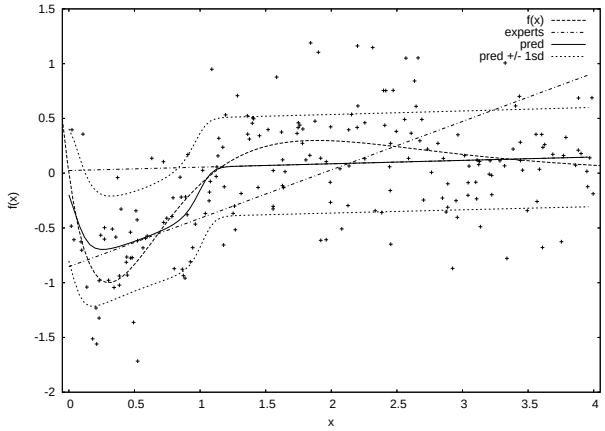


Figure 3: The original function $f_2(x)$, the training data with added noise, the prediction of the separate experts and the mixed prediction. The error bars show a single standard deviation from the model prediction.

5.2 Experiments

To evaluate the method's ability to identify the correct model, we have generated a 1-dimensional function with added Gaussian noise from a set of 3 linear experts, and have used the above procedure to identify the correct number of experts and matching function parameters, with random re-initialisation of the model structure every 1000 steps to escape local peaks in the model posterior. As done in LCS,

we completely rely on the model search for localisation by setting the gating features to $\vartheta(\boldsymbol{x}) = 1$, such that the gating network only performs a non-localised re-weighting of the experts. As can be seen from the results in Figure 2(a), the MCMC method with Bayesian model selection was able to correctly identify both the number of experts and their localisation in the input space. Figure 2(b) shows the change of the variational bound $\mathcal{L}(q)$ and the current number of experts while searching the model space, and demonstrates that the method quickly finds the correct number of experts after almost every random restart.

In an additional experiment, we have tested the performance on an artificial data set used in Waterhouse et al. (1996) by sampling from $f_2(x) = 4.26(e^{-x} - 4e^{-2x} + 3e^{-3x}) + \epsilon$ over the range $[0, 4]$, where ϵ is Gaussian noise with variance 0.44. With such a high variance no pattern was apparent, and the best model contained only a single expert. Upon reducing the noise variance to 0.2, two experts were identified as shown in Figure 3. Surprisingly, the curve was fitted more compactly with one local expert interleaving the prediction of another global expert, where we would have expected the use of 3 experts that perform a piecewise linear fit.

While the MCMC model search was able to find good models for low-dimensional data, we quickly reached its limits when increasing the dimensionality. The commonly observed pattern was the removal of all except for a single expert, and maintaining that expert. This behaviour is explained by the very low probability of finding good matching functions with random sampling in a high-dimensional model space. Additionally, the underlying generative Markov chain causes two successive models to be highly correlated and therefore does not support large changes in the model: Given that the single expert's matching is well tuned to the data, then the probability of adding another expert that in combination with the previous one performs better is highly unlikely.

Clearly, to improve model search we need to either design a sampling algorithm that uses more of the information on the structure of the data that is available in the model, or redesign the genetic algorithm of current LCS to make use of the model posterior approximate. Either approach is left as future work.

6 Discussion and Conclusion

By introducing a generalisation to the MoE model we have for the first time explicitly identified the model that underlies the LCS method. Compared to the original MoE, LCS allow for more complex localisation pattern of the experts, only limited by the choice of representation for the matching functions. This increased flexibility comes at the cost of having to perform a search in the potentially complex model structure space, which is made computationally cheaper by introducing a training scheme that makes the expert training independent of the gating network.

The link between MoE and LCS immediately enables several improvements to LCS: i) while matching is usually binary in LCS, we can now match to a degree by specifying the probability of matching $m_k(\boldsymbol{x})$; ii) the experts were so far restricted to linear regression models, but can now be easily extended to any generalised linear model; iii) we do not need to use $\vartheta(\boldsymbol{x}) = 1$, which allows the gating network to combine equally localised experts better; iv) Bayesian model selection replaces the heuristics in LCS. This clearly opens up a wide range of future work, some of which has been identified throughout the paper.

A Variational Distributions

In order to derive an expression for the variational distributions over the hidden random variables $\boldsymbol{U} = \{\boldsymbol{W}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{Z}, \boldsymbol{V}, \boldsymbol{\beta}\}$, we need to evaluate Eq. (17) for each element of $\boldsymbol{U}$, and the necessary moments of these. In all derivations we drop the terms that are independent of the hidden variable in question, making use of the factorisation

$$p(\boldsymbol{U}, \boldsymbol{Y}) = p(\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{\tau}) p(\boldsymbol{W}|\boldsymbol{\alpha}) p(\boldsymbol{\tau}) p(\boldsymbol{\alpha}) p(\boldsymbol{Z}|\boldsymbol{V}) p(\boldsymbol{V}|\boldsymbol{\beta}) p(\boldsymbol{\beta}). \tag{22}$$

All distributions are implicitly conditional on $\boldsymbol{X}$ and $\mathcal{M}$.

A.1 Expert weights $q_W(\boldsymbol{W})$

We need to evaluate

$$\begin{aligned} \ln q_W^*(\boldsymbol{W}) &= \mathbb{E}_{\tau,\alpha,Z,V,\beta}(\ln p(\boldsymbol{Y}, \boldsymbol{U})) + \text{const.} \\ &= \mathbb{E}_{\tau,Z}(\ln p(\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{\tau})) + \mathbb{E}_{\alpha}(\ln p(\boldsymbol{W}|\boldsymbol{\alpha})) + \text{const.}, \end{aligned}$$

using Eqs. (3), (7), and (8) to get

$$\begin{aligned} \mathbb{E}_{\tau,Z}(\ln p(\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{\tau})) &= \sum_n \sum_k \mathbb{E}_Z(z_{nk}) \mathbb{E}_\tau \left(\ln \mathcal{N}(y_n | \boldsymbol{w}_k^T \boldsymbol{x}_n, \tau_k^{-1}) \right) \\ &= \sum_n \sum_k \mathbb{E}_Z(z_{nk}) \mathbb{E}_\tau \left(\frac{\tau_k}{2} (y_n - \boldsymbol{w}_k^T \boldsymbol{x}_n)^2 \right) + \text{const.} \\ &= \sum_k \left(-\frac{\mathbb{E}_\tau(\tau_k)}{2} \boldsymbol{w}_k^T \sum_n (\mathbb{E}_Z(z_{nk}) \boldsymbol{x}_n \boldsymbol{x}_n^T) \boldsymbol{w}_k + \mathbb{E}_\tau(\tau_k) \boldsymbol{w}_k^T \sum_n \mathbb{E}_Z(z_{nk}) \boldsymbol{x}_n y_n \right) \\ &\quad + \text{const.}, \\ \mathbb{E}_\alpha(\ln p(\boldsymbol{W}|\boldsymbol{\alpha})) &= \sum_k \mathcal{N}(\boldsymbol{w}_k | 0, \alpha_k^{-1} \boldsymbol{I}) \\ &= \sum_k \left(-\frac{\mathbb{E}_\alpha(\alpha_k)}{2} \boldsymbol{w}_k^T \boldsymbol{w}_k \right) + \text{const.} \\ &= \sum_k \left(-\frac{1}{2} \boldsymbol{w}_k^T (\mathbb{E}_\alpha(\alpha_k) \boldsymbol{I}) \boldsymbol{w}_k \right) + \text{const.} \end{aligned}$$

Hence, $q_W^*(\boldsymbol{W})$ factorises w.r.t. k, which allows us to treat each $q_W^*(\boldsymbol{w}_k)$ separately, resulting in

$$\begin{aligned} \ln q_W^*(\boldsymbol{w}_k) &= -\frac{1}{2} \boldsymbol{w}_k^T \left(\mathbb{E}_\alpha(\alpha_k) \boldsymbol{I} + \mathbb{E}_\tau(\tau_k) \sum_n \mathbb{E}_Z(z_{nk}) \boldsymbol{x}_n \boldsymbol{x}_n^T \right) \boldsymbol{w}_k \\ &\quad + \mathbb{E}_\tau(\tau_k) \boldsymbol{w}_k^T \sum_n \mathbb{E}_Z(z_{nk}) \boldsymbol{x}_n y_n + \text{const.} \end{aligned}$$

By completing the square we can see that hence $q_W^*(\boldsymbol{w}_k)$ is a Gaussian distribution given by

$$\begin{aligned} q_W^*(\boldsymbol{w}_k) &= \mathcal{N}(\boldsymbol{w}_k | \boldsymbol{w}_k^*, \boldsymbol{\Sigma}_k^*), \\ \boldsymbol{\Sigma}_k^* &= \left(\mathbb{E}_\alpha(\alpha_k) \boldsymbol{I} + \mathbb{E}_\tau(\tau_k) \sum_n \mathbb{E}_Z(z_{nk}) \boldsymbol{x}_n \boldsymbol{x}_n^T \right)^{-1}, \\ \boldsymbol{w}_k^* &= \mathbb{E}_\tau(\tau_k) \boldsymbol{\Sigma}_k^* \sum_n \mathbb{E}_Z(z_{nk}) \boldsymbol{x}_n y_n. \end{aligned} \tag{23}$$

A.2 Expert precisions $q_\tau(\boldsymbol{\tau})$

We require

$$\begin{aligned} \ln q_\tau^*(\boldsymbol{\tau}) &= \mathbb{E}_{W,\alpha,Z,V,\beta}(\ln p(\boldsymbol{Y}, \boldsymbol{U})) + \text{const.} \\ &= \mathbb{E}_{W,Z}(\ln p(\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{\tau})) + \ln p(\boldsymbol{\tau}) + \text{const.} \end{aligned}$$

Using Eqs. (3), (7), and (10) we get

$$\begin{aligned}
\mathbb{E}_{W,Z}(\ln p(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{W},\boldsymbol{\tau})) &= \sum_n \sum_k \mathbb{E}_Z(z_{nk}) \mathbb{E}_W \left(\ln \mathcal{N}(y_n|\boldsymbol{w}_k^T \boldsymbol{x}_n, \tau_k^{-1})\right) \\
&= \sum_n \sum_k \mathbb{E}_Z(z_{nk}) \left(\frac{1}{2} \ln \tau_k - \frac{\tau_k}{2} \mathbb{E}_W \left((y_n - \boldsymbol{w}_k^T \boldsymbol{x}_n)^2\right)\right) + \text{const.}, \\
\ln p(\boldsymbol{\tau}) &= \sum_k \ln \text{Gam}(\tau_k|a_\tau, b_\tau) \\
&= \sum_k \left((a_\tau - 1) \ln \tau_k - b_\tau \tau_k\right) + \text{const.}
\end{aligned}$$

As for $q_W^*(\boldsymbol{W})$, this distribution factorises w.r.t. k, and therefore we can evaluate each $q_\tau^*(\tau_k)$ separately, resulting in

$$\begin{aligned}
\ln q_\tau^*(\tau_k) &= \left(\frac{1}{2} \sum_n \mathbb{E}_Z(z_{nk}) + a_\tau - 1\right) \ln \tau_k \\
&\quad + \tau_k \left(b_\tau + \frac{1}{2} \sum_n \mathbb{E}_Z(z_{nk}) \mathbb{E}_W \left((y_n - \boldsymbol{w}_k^T \boldsymbol{x}_n)^2\right)\right) + \text{const..},
\end{aligned}$$

which is a Gamma distribution of the form

$$\begin{aligned}
q_\tau(\tau_k) &= \text{Gam}(\tau_k|a_{\tau_k}^*, b_{\tau_k}^*), \\
a_{\tau_k}^* &= a_\tau + \frac{1}{2} \sum_n \mathbb{E}_Z(z_{nk}), \\
b_{\tau_k}^* &= b_\tau + \frac{1}{2} \sum_n \mathbb{E}_Z(z_{nk}) \mathbb{E}_W \left((y_n + \boldsymbol{w}_k^T \boldsymbol{x}_n)^2\right).
\end{aligned} \tag{24}$$

A.3 Expert weight priors $q_\alpha(\boldsymbol{\alpha})$

We need to evaluate

$$\begin{aligned}
\ln q_\alpha^*(\boldsymbol{\alpha}) &= \mathbb{E}_{W,\tau,Z,V,\beta}(\ln p(\boldsymbol{Y},\boldsymbol{U})) + \text{const.} \\
&= \mathbb{E}_W(\ln p(\boldsymbol{W},\boldsymbol{\alpha})) + \ln p(\boldsymbol{\alpha}) + \text{const.}
\end{aligned}$$

Using Eqs. (8) and (11) we get

$$\begin{aligned}
\mathbb{E}_W(\ln p(\boldsymbol{W},\boldsymbol{\alpha})) &= \sum_k \mathbb{E}_W(\ln \mathcal{N}(\boldsymbol{w}_k|0, \alpha_k^{-1}\boldsymbol{I})) \\
&= \sum_k \left(\frac{D_w}{2} \ln \alpha_k - \frac{\alpha_k}{2} \mathbb{E}_W(\boldsymbol{w}_k^T \boldsymbol{w}_k)\right) + \text{const.}, \\
\ln p(\boldsymbol{\alpha}) &= \sum_k \ln \text{Gam}(\alpha_k|a_\alpha, b_\alpha) \\
&= \sum_k \left((a_\alpha - 1) \ln \alpha_k - \alpha_k b_\alpha\right) + \text{const.},
\end{aligned}$$

where D_w is the size of the weight vector $\boldsymbol{w}_k$. Again, this distribution factorises w.r.t. k, and so we get for $q_\alpha^*(\alpha_k)$

$$\ln q_\alpha^*(\alpha_k) = \left(\frac{D_w}{2} + a_\alpha - 1\right) \ln \alpha_k - \alpha_k \left(b_\alpha + \frac{1}{2} \mathbb{E}_W(\boldsymbol{w}_k^T \boldsymbol{w}_k)\right) + \text{const.},$$

which is the Gamma distribution

$$\begin{aligned}
q_\alpha^*(\alpha_k) &= \text{Gam}(\alpha_k|a_{\alpha_k}^*, b_{\alpha_k}^*), \\
a_{\alpha_k}^* &= a_\alpha + \frac{D_w}{2}, \\
b_{\alpha_K}^* &= b_\alpha + \frac{1}{2} \mathbb{E}_W(\boldsymbol{w}_k^T \boldsymbol{w}_k).
\end{aligned} \tag{25}$$

A.4 Gating weights $q_V(\boldsymbol{V})$

To get $q_V^*(\boldsymbol{V})$ we need to evaluate

$$\begin{aligned}\ln q_V^*(\boldsymbol{V}) &= \mathbb{E}_{W,\tau,\alpha,Z,\beta}(\ln p(\boldsymbol{Y},\boldsymbol{U})) + \text{const.} \\ &= \mathbb{E}_Z(\ln p(\boldsymbol{Z}|\boldsymbol{V})) + \mathbb{E}_\beta(\ln p(\boldsymbol{V}|\boldsymbol{\beta})) + \text{const.}\end{aligned}$$

Using Eq. (12) we get

$$\begin{aligned}\mathbb{E}_\beta(\ln p(\boldsymbol{V}|\boldsymbol{\beta})) &= \sum_k \mathbb{E}_\beta\left(\ln \mathcal{N}(\boldsymbol{v}_k|0,\beta_k^{-1}\boldsymbol{I})\right) \\ &= \sum_k \left(-\frac{\mathbb{E}_\beta(\beta_k)}{2}\boldsymbol{v}_k^T\boldsymbol{v}_k\right) + \text{const.}.\end{aligned}$$

To get $\mathbb{E}_Z(\ln p(\boldsymbol{Z}|\boldsymbol{V}))$ we can use the 1-of-K structure of $\boldsymbol{z}$, resulting in

$$p(\boldsymbol{Z}|\boldsymbol{V}) = \prod_n\prod_k p(z_{nk}=1|\boldsymbol{x}_n,\boldsymbol{v}_k)^{z_{nk}} = \prod_n\prod_k g_k(\boldsymbol{x}_n)^{z_{nk}}, \tag{26}$$

which gives

$$\mathbb{E}_Z(\ln p(\boldsymbol{Z}|\boldsymbol{V})) = \sum_n\sum_k \mathbb{E}_Z(z_{nk})\ln g_k(\boldsymbol{x}_n).$$

Therefore, we get for $\ln q_V^*(\boldsymbol{V})$

$$\ln q_V^*(\boldsymbol{V}) = \sum_n\sum_k \mathbb{E}_Z(z_{nk})\ln g_k(\boldsymbol{x}_n) - \sum_k \frac{\mathbb{E}_\beta(\beta_k)}{2}\boldsymbol{v}_k^T\boldsymbol{v}_k + \text{const.},$$

which is not an exponential distribution and therefore spoils our conjugate-exponential structure. Hence, we will proceed as in Waterhouse et al. (1996) and approximate q_V^* by a Laplace approximation, using a Gaussian $\mathcal{N}(\boldsymbol{V}|\tilde{\boldsymbol{V}},\tilde{\boldsymbol{\Lambda}}_V^{-1})$ centred on the mode of q_V^*, and with a similar covariance structure. We get the mode by setting the derivative of $\ln q_V^*(\boldsymbol{V})$ w.r.t. $\boldsymbol{V}$ to zero, that is

$$\sum_n(\mathbb{E}_Z(z_{nk}) - g_k(\boldsymbol{x}_n))\vartheta(\boldsymbol{x}_n) - \mathbb{E}_\beta(\beta_k)\boldsymbol{v}_k = 0, \qquad \forall k,$$

for which the solution can be obtained by the IRLS algorithm, as, for example, described in Bishop (2006).

The precision matrix $\tilde{\boldsymbol{\Lambda}}_V$ is block-symmetrical and is obtained in these blocks $(\tilde{\boldsymbol{\Lambda}}_V)_{jk}$ by

$$(\tilde{\boldsymbol{\Lambda}}_V)_{jk} = \frac{\partial \ln q_V^*(\boldsymbol{V})}{\partial\boldsymbol{v}_j\partial\boldsymbol{v}_k} = \left(-\sum_n g_k(\boldsymbol{x}_n)(\boldsymbol{I}_{jk} - g_j(\boldsymbol{x}_n))\vartheta(\boldsymbol{x}_n)\vartheta(\boldsymbol{x}_n)^T\right) - \boldsymbol{I}_{jk}\mathbb{E}_\beta(\beta_k),$$

where $\boldsymbol{I}_{jk}$ is the jkth element of the identity matrix, and $\tilde{\boldsymbol{V}}$ is used to evaluate $g_k(\boldsymbol{x}_n)$.

This results in the Gaussian approximation to q_V^*, given by

$$\tilde{q}_V^*(\boldsymbol{V}) = \mathcal{N}(\boldsymbol{V}|\tilde{\boldsymbol{V}},\tilde{\boldsymbol{\Lambda}}_V^{-1}). \tag{27}$$

A.5 Gating weight priors $q_\beta(\boldsymbol{\beta})$

We require

$$\begin{aligned}\ln q_\beta^*(\boldsymbol{\beta}) &= \mathbb{E}_{W,\tau,\alpha,Z,V}(\ln p(\boldsymbol{Y},\boldsymbol{U})) + \text{const.} \\ &= \mathbb{E}_V(\ln p(\boldsymbol{V}|\boldsymbol{\beta})) + \ln p(\boldsymbol{\beta}) + \text{const.}\end{aligned}$$

Using Eqs. (9) and (12) we get

$$\begin{aligned}\mathbb{E}_V(\ln p(\boldsymbol{V}|\boldsymbol{\beta})) &= \sum_k \ln \mathcal{N}(\boldsymbol{v}_k|0,\beta_k^{-1}\boldsymbol{I}) \\ &= \sum_k \left(\frac{D_v}{2}\ln\beta_k - \frac{\beta_k}{2}\mathbb{E}_V(\boldsymbol{v}_k^T\boldsymbol{v}_k)\right) + \text{const.}, \\ \ln p(\boldsymbol{\beta}) &= \sum_k \ln \text{Gam}(\beta_k|a_\beta,b_\beta) \\ &= \sum_k \left((a_\beta-1)\ln\beta_k - \beta_k b_\beta\right) + \text{const.},\end{aligned}$$

where D_v is the size of $\boldsymbol{v}_k$, and which factorise w.r.t. k. Hence, we have

$$\ln q^*_\beta(\beta_k) = \left(\frac{D_v}{2} + a_\beta - 1\right)\ln\beta_k - \beta_k\left(b_\beta + \frac{1}{2}\mathbb{E}_V(\boldsymbol{v}_k^T\boldsymbol{v}_k)\right) + \text{const.},$$

which is the Gamma distribution

$$\begin{aligned} q^*_\beta(\beta_k) &= \text{Gam}(\beta_k|a^*_{\beta_k}, b^*_{\beta_k}), \\ a^*_{\beta_k} &= a_\beta + \frac{D_v}{2}, \\ b^*_{\beta_k} &= b_\beta + \frac{1}{2}\mathbb{E}_V(\boldsymbol{v}_k^T\boldsymbol{v}_k). \end{aligned} \tag{28}$$

A.6 Gating $q_Z(\boldsymbol{Z})$

We need to evaluate

$$\begin{aligned} \ln q^*_Z(\boldsymbol{Z}) &= \mathbb{E}_{W,\tau,\alpha,V,\beta}(\ln p(\boldsymbol{Y},\boldsymbol{U})) + \text{const.} \\ &= \mathbb{E}_{W,\tau}(\ln p(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{W},\boldsymbol{\tau})) + \mathbb{E}_V(\ln p(\boldsymbol{Z}|\boldsymbol{V})) + \text{const.} \end{aligned}$$

With the use of Eqs. (3), (7), and (26), we get

$$\begin{aligned} \mathbb{E}_{W,\tau}(\ln p(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{W},\boldsymbol{\tau})) &= \sum_n\sum_k z_{nk}\mathbb{E}_{W,\tau}(\ln\mathcal{N}(y_n|\boldsymbol{w}_k^T\boldsymbol{x}_n,\tau_k^{-1})) \\ &= \sum_n\sum_k z_{nk}\left(\frac{1}{2}\mathbb{E}_\tau(\ln\tau_k) - \frac{\mathbb{E}_\tau(\tau_k)}{2}\mathbb{E}_W\left((y_n - \boldsymbol{w}_k^T\boldsymbol{x}_n)^2\right)\right) + \text{const.}, \\ \mathbb{E}_V(\ln p(\boldsymbol{Z}|\boldsymbol{V})) &= \sum_n\sum_k z_{nk}\mathbb{E}_V(\ln g_k(\boldsymbol{x}_n)). \end{aligned}$$

Hence, we have

$$\ln q^*_Z(\boldsymbol{Z}) = \sum_n\sum_k z_{nk}\ln\rho_{nk} + \text{const.},$$

with

$$\ln\rho_{nk} = \ln\mathbb{E}_V(\ln g_k(\boldsymbol{x}_n)) + \frac{1}{2}\mathbb{E}_\tau(\ln\tau_k) - \frac{1}{2}\mathbb{E}_\tau(\tau_k)\mathbb{E}_W\left((y_n - \boldsymbol{w}_k^T\boldsymbol{x}_n)^2\right).$$

It follows that

$$q^*_Z(\boldsymbol{Z}) \propto \prod_n\prod_k \rho_{nk}^{z_{nk}},$$

which, when normalised by $\sum_k z_{nk} = 1$, gives

$$q^*_Z(\boldsymbol{Z}) = \prod_n\prod_k r_{nk}^{z_{nk}}, \tag{29}$$

$$r_{nk} = \frac{\rho_{nk}}{\sum_j \rho_{nj}}. \tag{30}$$

A.7 Moments

To evaluate the parameters of the variational distributions we require several moments of other distributions that we evaluate below.

The required moments of the expert weights are based on Eq. (23) and are given by

$$\begin{aligned} \mathbb{E}_W(\boldsymbol{w}_k^T\boldsymbol{w}_k) &= \boldsymbol{w}_k^{*T}\boldsymbol{w}_k^* + \text{Tr}(\boldsymbol{\Sigma}_k^*), \\ \mathbb{E}_W\left((y_n - \boldsymbol{w}_k^T\boldsymbol{x}_n)^2\right) &= y_n^2 - 2\mathbb{E}_W(\boldsymbol{w}_k)^T\boldsymbol{x}_n y_n + \boldsymbol{x}_n^T\mathbb{E}_W(\boldsymbol{w}_k\boldsymbol{w}_k^T)\boldsymbol{x}_n \\ &= y_n^2 - 2\boldsymbol{w}_k^{*T}\boldsymbol{x}_n y_n + \boldsymbol{x}_n^T(\boldsymbol{w}_k^*\boldsymbol{w}_k^{*T} + \boldsymbol{\Sigma}_k^*)\boldsymbol{x}_n \\ &= \|y_n - \boldsymbol{w}_k^T\boldsymbol{x}_n\|^2 + \boldsymbol{x}_n^T\boldsymbol{\Sigma}_k^*\boldsymbol{x}_n. \end{aligned}$$

The moments of the expert precision Eq. (24) are

$$\begin{aligned} \mathbb{E}_\tau(\tau_k) &= \frac{a^*_{\tau_k}}{b^*_{\tau_k}}, \\ \mathbb{E}_\tau(\ln\tau_k) &= \psi(a^*_{\tau_k}) - \ln b^*_{\tau_k}, \end{aligned}$$

where $\psi(\cdot)$ is the digamma function. For the expert weight priors Eq. (25) we have the moments

$$\begin{aligned}\mathbb{E}_\alpha(\alpha_k) &= \frac{a^*_{\alpha_k}}{b^*_{\alpha_k}},\\ \mathbb{E}_\alpha(\ln \alpha_k) &= \psi(a^*_{\alpha_k}) - \ln b^*_{\alpha_k}.\end{aligned}$$

Even though $\mathbb{E}_\alpha(\ln \alpha_k)$ is not required to find the variational distribution, we have evaluated it for use in a later section.

We get the expectation of the gating variables z_{nk} by inspection of Eq. (29) ,

$$\mathbb{E}_Z(z_{nk}) = r_{nk},$$

where r_{nk} is called the *responsibility* of expert k for observation n. The moments of the gating weights are based on the approximation Eq. (27), and are

$$\begin{aligned}\mathbb{E}_V(\ln g_k(\boldsymbol{x}_n)) &\approx \ln g_k(\boldsymbol{x}_n)|_{\boldsymbol{V}=\tilde{\boldsymbol{V}}},\\ \mathbb{E}_V(\boldsymbol{v}_k^T \boldsymbol{v}_k) &= \tilde{\boldsymbol{v}}_k^T \tilde{\boldsymbol{v}}_k + \mathrm{Tr}\left((\tilde{\boldsymbol{\Lambda}}_V^{-1})_{kk}\right),\end{aligned}$$

where the moment $\mathbb{E}_V(\ln g_k(\boldsymbol{x}_n))$ cannot be directly evaluated and is therefore crudely approximated by the logarithm of the expectation, as also done in Waterhouse et al. (1996). In $\mathbb{E}_V(\boldsymbol{v}_k^T \boldsymbol{v}_k)$, $(\tilde{\boldsymbol{\Lambda}}_V^{-1})_{kk}$ stands for the kth diagonal block matrix in $\tilde{\boldsymbol{\Lambda}}_V^{-1}$. The moments of the gating weight prior is after Eq. (28) given by

$$\begin{aligned}\mathbb{E}_\beta(\beta_k) &= \frac{a^*_{\beta_k}}{b^*_{\beta_k}},\\ \mathbb{E}_\beta(\ln \beta_k) &= \psi(a^*_{\beta_k}) - \ln b^*_{\beta_k}.\end{aligned}$$

Again, $\mathbb{E}_\beta(\ln \beta_k)$ is evaluated for use in a later section.

A.8 Independent Expert Training

When the experts are trained independently of the gating network, then the update equations for $q^*_W(\boldsymbol{w}_k)$ and $q^*_\tau(\tau_k)$ are changed to

$$\begin{aligned}\boldsymbol{\Sigma}^*_k &= \left(\mathbb{E}_\alpha(\alpha_k)\boldsymbol{I} + \mathbb{E}_\tau(\tau_k)\sum_n m_k(\boldsymbol{x}_n)\boldsymbol{x}_n\boldsymbol{x}_n^T\right)^{-1},\\ \boldsymbol{w}^*_k &= \mathbb{E}_\tau(\tau_k)\boldsymbol{\Sigma}^*_k \sum_n m_k(\boldsymbol{x})_n \boldsymbol{x}_n y_n,\\ a^*_{\tau_k} &= a_\tau + \frac{1}{2}\sum_n m_k(\boldsymbol{x}_n),\\ b^*_{\tau_k} &= b_\tau + \frac{1}{2}\sum_n m_k(\boldsymbol{x}_n)\mathbb{E}_W\left((y_n + \boldsymbol{w}_k^T\boldsymbol{x}_n)^2\right),\end{aligned}$$

by replacing $\mathbb{E}_Z(z_{nk})$ with $m_k(\boldsymbol{x}_n)$. All other update equations stay unchanged.

B Variational Bound $\mathcal{L}(\tilde{q})$

The variational bound $\mathcal{L}(\tilde{q})$ is evaluated after Eq. (14), which is given by

$$\begin{aligned}\mathcal{L}(\tilde{q}) &= \sum_Z \int\cdots\int \tilde{q}(\boldsymbol{W},\boldsymbol{\tau},\boldsymbol{\alpha},\boldsymbol{Z},\boldsymbol{V},\boldsymbol{\beta}) \ln\left(\frac{p(\boldsymbol{Y},\boldsymbol{W},\boldsymbol{\tau},\boldsymbol{\alpha},\boldsymbol{Z},\boldsymbol{V},\boldsymbol{\beta})}{\tilde{q}(\boldsymbol{W},\boldsymbol{\tau},\boldsymbol{\alpha},\boldsymbol{Z},\boldsymbol{V},\boldsymbol{\beta})}\right)\mathrm{d}\boldsymbol{W}\mathrm{d}\boldsymbol{\tau}\mathrm{d}\boldsymbol{\alpha}\mathrm{d}\boldsymbol{V}\mathrm{d}\boldsymbol{\beta}\\ &= \mathbb{E}_{W,\tau,\alpha,Z,V,\beta}(\ln p(\boldsymbol{Y},\boldsymbol{W},\boldsymbol{\tau},\boldsymbol{\alpha},\boldsymbol{Z},\boldsymbol{V},\boldsymbol{\beta})) - \mathbb{E}_{W,\tau,\alpha,Z,V,\beta}(\ln q(\boldsymbol{W},\boldsymbol{\tau},\boldsymbol{\alpha},\boldsymbol{Z},\boldsymbol{V},\boldsymbol{\beta}))\\ &= \mathbb{E}_{W,\tau,Z}(\ln p(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{W},\boldsymbol{\tau})) + \mathbb{E}_{W,\alpha}(\ln p(\boldsymbol{W}|\boldsymbol{\alpha})) + \mathbb{E}_\tau(\ln p(\boldsymbol{\tau})) + \mathbb{E}_\alpha(\ln p(\boldsymbol{\alpha}))\\ &\quad + \mathbb{E}_{Z,V}(\ln p(\boldsymbol{Z}|\boldsymbol{V})) + \mathbb{E}_{V,\beta}(\ln p(\boldsymbol{V}|\boldsymbol{\beta})) + \mathbb{E}_\beta(\ln p(\boldsymbol{\beta}))\\ &\quad - \mathbb{E}_W(\ln q(\boldsymbol{W})) - \mathbb{E}_\tau(\ln q(\boldsymbol{\tau})) - \mathbb{E}_\alpha(\ln q(\boldsymbol{\alpha}))\\ &\quad - \mathbb{E}_Z(\ln q(\boldsymbol{Z})) - \mathbb{E}_V(\ln \tilde{q}(\boldsymbol{V})) - \mathbb{E}_\beta(\ln q(\boldsymbol{\beta})).\end{aligned}$$

All distributions are implicitly conditional on $\boldsymbol{X}$ and $\mathcal{M}$. The next sections are dedicated to deriving the required moments, after which we will return to deriving the close-form expression for $\mathcal{L}(\tilde{q})$.

B.1 $\mathbb{E}_{W,\tau,Z}(\ln p(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{W},\boldsymbol{\tau})$

Using Eqs. (3) and (7), and the moments from Section A.7, we get

$$\begin{aligned}
\mathbb{E}_{W,\tau,Z}(\ln p(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{W},\boldsymbol{\tau}) &= \iiint q(\boldsymbol{W},\boldsymbol{W},\boldsymbol{\tau}) \ln p(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{W},\boldsymbol{\tau}) \mathrm{d}\boldsymbol{Z}\mathrm{d}\boldsymbol{W}\mathrm{d}\boldsymbol{\tau} \\
&= \sum_n \sum_k \int q(\boldsymbol{Z}) z_{nk} \mathrm{d}\boldsymbol{Z} \iint q(\boldsymbol{W},\boldsymbol{\tau}) \ln \mathcal{N}(y_n|\boldsymbol{w}_k^T\boldsymbol{x}_n, \tau_k^{-1}) \mathrm{d}\boldsymbol{W}\mathrm{d}\boldsymbol{\tau} \\
&= \sum_n \sum_k \mathbb{E}_Z(z_{nk}) \left(\frac{1}{2}\mathbb{E}_\tau(\ln\tau_k) - \frac{1}{2}\ln 2\pi - \frac{1}{2}\mathbb{E}_\tau(\tau_k)\mathbb{E}_W\left((y_n - \boldsymbol{w}_k^T\boldsymbol{x}_n)^2\right)\right) \\
&= \sum_k \left((\psi(a^*_{\tau_k}) - \ln b^*_{\tau_k} - \ln 2\pi)\frac{1}{2}\sum_k r_{nk} - \frac{1}{2}\frac{a^*_{\tau_k}}{b^*_{\tau_k}} \sum_n r_{nk}\mathbb{E}_W\left((y_n - \boldsymbol{w}_k^T\boldsymbol{x}_n)^2\right)\right).
\end{aligned}$$

B.2 $\mathbb{E}_{W,\alpha}(\ln p(\boldsymbol{W}|\boldsymbol{\alpha}))$ and $\mathbb{E}_W(\ln q(\boldsymbol{W}))$

Using Eq. (8) and the moments from Section A.7, we get

$$\begin{aligned}
\mathbb{E}_{W,\alpha}(\ln p(\boldsymbol{W}|\boldsymbol{\alpha})) &= \iint q(\boldsymbol{W},\boldsymbol{\alpha}) \ln p(\boldsymbol{W}|\boldsymbol{\alpha}) \mathrm{d}\boldsymbol{W}\mathrm{d}\boldsymbol{\alpha} \\
&= \sum_k \iint q(\boldsymbol{w_k},\alpha_k) \ln \mathcal{N}(\boldsymbol{w_k}|0,\alpha_k^{-1}\boldsymbol{I}) \mathrm{d}\boldsymbol{w}_k\mathrm{d}\alpha_k \\
&= \sum_k \left(-\frac{D_w}{2}\ln 2\pi + \frac{D_w}{2}\mathbb{E}_\alpha(\ln\alpha_k) - \frac{1}{2}\mathbb{E}_\alpha(\alpha_k)\mathbb{E}_W(\boldsymbol{w}_k^T\boldsymbol{w}_k)\right) \\
&= \sum_k \left(-\frac{D_w}{2}\ln 2\pi + \frac{D_w}{2}(\psi(a^*_{\alpha_k}) - \ln b^*_{\alpha_k}) - \frac{1}{2}\frac{a^*_{\alpha_k}}{b^*_{\alpha_k}}(\boldsymbol{w}_k^{*T}\boldsymbol{w}_k^* + \mathrm{Tr}(\boldsymbol{\Sigma}_k^*))\right).
\end{aligned}$$

We can evaluate $\mathbb{E}_W(\ln q(\boldsymbol{W}))$ by using Eq. (23) and observing that

$$-\mathbb{E}_W(\ln q(\boldsymbol{W})) = -\int q(\boldsymbol{W})\ln q(\boldsymbol{W})\mathrm{d}\boldsymbol{W} = \sum_k \left(-\int q(\boldsymbol{w}_k)\ln q(\boldsymbol{w}_k)\mathrm{d}\boldsymbol{w}_k\right)$$

is the sum of entropies of $q(\boldsymbol{w}_k)$, and therefore $\mathbb{E}_W(\ln q(\boldsymbol{W}))$ is given by

$$\mathbb{E}_W(\ln q(\boldsymbol{W})) = -\sum_k \left(\frac{1}{2}\ln|\boldsymbol{\Sigma}_k^*| + \frac{D_w}{2}(1 + \ln 2\pi)\right).$$

B.3 $\mathbb{E}_\tau(\ln p(\boldsymbol{\tau}))$ and $\mathbb{E}_\tau(\ln q(\boldsymbol{\tau}))$

From Eq. (10) and the moments evaluated in Section A.7 follows

$$\begin{aligned}
\mathbb{E}_\tau(\ln p(\boldsymbol{\tau})) &= \int q(\boldsymbol{\tau})\ln p(\boldsymbol{\tau})\mathrm{d}\boldsymbol{\tau} \\
&= \sum_k \int q(\tau_k)\ln \mathrm{Gam}(\tau_k|a_\tau,b_\tau)\mathrm{d}\tau_k \\
&= \sum_k (-\ln\Gamma(a_\tau) + a_\tau\ln b_\tau + (a_\tau - 1)\mathbb{E}_\tau(\ln\tau_k) - b_\tau\mathbb{E}_\tau(\tau_k)) \\
&= \sum_k \left(-\ln\Gamma(a_\tau) + a_\tau\ln b_\tau + (a_\tau - 1)(\psi(a^*_{\tau_k}) - \ln b^*_{\tau_k}) - b_\tau\frac{a^*_{\tau_k}}{b^*_{\tau_k}}\right),
\end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function. We can again observe that $-\mathbb{E}_\tau(\ln q(\boldsymbol{\tau}))$ is the sum of entropies of $q^*_\tau(\tau_k)$ after Eq. (24), and therefore we get

$$\begin{aligned}
\mathbb{E}_\tau(\ln q(\boldsymbol{\tau})) &= \int q(\boldsymbol{\tau})\ln q(\boldsymbol{\tau})\mathrm{d}\boldsymbol{\tau} \\
&= -\sum_k \left(-\int q(\tau_k)\ln q(\tau_k)\mathrm{d}\tau_k\right) \\
&= -\sum_k \left(\ln\Gamma(a^*_{\tau_k}) - (a^*_{\tau_k} - 1)\psi(a^*_{\tau_k}) - \ln b^*_{\tau_k} + a^*_{\tau_k}\right)
\end{aligned}$$

B.4 $\mathbb{E}_\alpha(\ln p(\boldsymbol{\alpha}))$ and $\mathbb{E}_\alpha(\ln q(\boldsymbol{\alpha}))$

The derivations are the same as for $\mathbb{E}_\tau(\ln p(\boldsymbol{\tau}))$ and $\mathbb{E}_\tau(\ln q(\boldsymbol{\tau}))$ and result in

$$\begin{aligned}\mathbb{E}_\alpha(\ln p(\boldsymbol{\alpha})) &= \sum_k \left(-\ln\Gamma(a_\alpha) + a_\alpha \ln b_\alpha + (a_\alpha - 1)(\psi(a^*_{\alpha_k}) - \ln b^*_{\alpha_k}) - b_\alpha \frac{a^*_{\alpha_k}}{b^*_{\alpha_k}} \right), \\ \mathbb{E}_\alpha(\ln q(\boldsymbol{\alpha})) &= -\sum_k \left(\ln\Gamma(a^*_{\alpha_k}) - (a^*_{\alpha_k} - 1)\psi(a^*_{\alpha_k}) - \ln b^*_{\alpha_k} + a^*_{\alpha_k} \right).\end{aligned}$$

B.5 $\mathbb{E}_{Z,V}(\ln p(\boldsymbol{Z}|\boldsymbol{V}))$ and $\mathbb{E}_Z(\ln q(\boldsymbol{Z}))$

From Eq. (26) and Section A.7 we get

$$\begin{aligned}\mathbb{E}_{Z,V}(\ln p(\boldsymbol{Z}|\boldsymbol{V})) &= \iint q(\boldsymbol{Z},\boldsymbol{V}) \ln p(\boldsymbol{Z}|\boldsymbol{V}) \mathrm{d}\boldsymbol{Z}\mathrm{d}\boldsymbol{V} \\ &= \sum_n \sum_k \int q(z_{nk}) z_{nk} \mathrm{d}z_{nk} \int q(\boldsymbol{v}_k) \ln g_k(\boldsymbol{x}_n) \mathrm{d}\boldsymbol{v}_k \\ &= \sum_n \sum_k \mathbb{E}_Z(z_{nk}) \mathbb{E}_V(g_k(\boldsymbol{x}_n)) \\ &= \sum_n \sum_k r_{nk} \ln g_k(\boldsymbol{x}_n)|_{\boldsymbol{v}_k = \tilde{\boldsymbol{v}}_k}.\end{aligned}$$

Additionally, using Eq. (29),

$$\begin{aligned}\mathbb{E}_Z(\ln q(\boldsymbol{Z})) &= \int q(\boldsymbol{Z}) \ln q(\boldsymbol{Z}) \mathrm{d}\boldsymbol{Z} \\ &= \sum_n \sum_k \mathbb{E}_Z(z_{nk}) \ln r_{nk} \\ &= \sum_n \sum_k r_{nk} \ln r_{nk},\end{aligned}$$

which is the negative entropy of $q(\boldsymbol{Z})$.

B.6 $\mathbb{E}_{V,\beta}(\ln p(\boldsymbol{V}|\boldsymbol{\beta}))$ and $\mathbb{E}_V(\ln \tilde{q}(\boldsymbol{V}))$

Deriving $\mathbb{E}_{V,\beta}(\ln p(\boldsymbol{V}|\boldsymbol{\beta}))$ is similar to the derivation of $\mathbb{E}_{W,\alpha}(\ln p(\boldsymbol{W}|\boldsymbol{\alpha}))$ and results in

$$\mathbb{E}_{V,\beta}(\ln p(\boldsymbol{V}|\boldsymbol{\beta})) = \sum_k \left(-\frac{D_v}{2} \ln 2\pi + \frac{D_v}{2}(\psi(a^*_{\beta_k}) - \ln b^*_{\beta_k}) - \frac{1}{2} \frac{a^*_{\beta_k}}{b^*_{\beta_k}} \left(\tilde{\boldsymbol{v}}_k^T \tilde{\boldsymbol{v}}_k + \mathrm{Tr}((\tilde{\boldsymbol{\Lambda}}_V^{-1})_{kk}) \right) \right).$$

To evaluate $\mathbb{E}_V(\ln \tilde{q}(\boldsymbol{V}))$ we note that $-\mathbb{E}_V(\ln \tilde{q}(\boldsymbol{V}))$ is the entropy of $\tilde{q}(\boldsymbol{V})$ after Eq. (27) to get

$$\begin{aligned}\mathbb{E}_V(\ln \tilde{q}(\boldsymbol{V})) &= -\left(-\int \tilde{q}(\boldsymbol{V}) \ln \tilde{q}(\boldsymbol{V}) \mathrm{d}\boldsymbol{V} \right) \\ &= -\left(\frac{1}{2} \ln |\tilde{\boldsymbol{\Lambda}}_V^{-1}| + \frac{K D_v}{2}(1 + \ln 2\pi) \right).\end{aligned}$$

B.7 $\mathbb{E}_\beta(\ln p(\boldsymbol{\beta}))$ and $\mathbb{E}_\beta(\ln q(\boldsymbol{\beta}))$

The derivations are the same as for $\mathbb{E}_\tau(\ln p(\boldsymbol{\tau}))$ and $\mathbb{E}_\tau(\ln q(\boldsymbol{\tau}))$ and result in

$$\begin{aligned}\mathbb{E}_\beta(\ln p(\boldsymbol{\beta})) &= \sum_k \left(-\ln\Gamma(a_\beta) + a_\beta \ln b_\beta + (a_\beta - 1)(\psi(a^*_{\beta_k}) - \ln b^*_{\beta_k}) - b_\beta \frac{a^*_{\beta_k}}{b^*_{\beta_k}} \right), \\ \mathbb{E}_\beta(\ln q(\boldsymbol{\beta})) &= -\sum_k \left(\ln\Gamma(a^*_{\beta_k}) - (a^*_{\beta_k} - 1)\psi(a^*_{\beta_k}) - \ln b^*_{\beta_k} + a^*_{\beta_k} \right).\end{aligned}$$

B.8 Closed-Form Equation for $\mathcal{L}(\tilde{q})$

Let $\mathcal{L}_k(q)$ denote the contribution of expert k independent of the gating network to the variational bound $\mathcal{L}(q)$. Using the contribution of expert k to $\mathbb{E}_{W,\tau,Z}(\ln p(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{W},\boldsymbol{\tau}))$, $\mathbb{E}_{W,\alpha}(\ln p(\boldsymbol{W}|\boldsymbol{\alpha}))$, $\mathbb{E}_\tau(\ln p(\boldsymbol{\tau}))$, $\mathbb{E}_\alpha(\ln p(\boldsymbol{\alpha}))$, $\mathbb{E}_W(\ln q(\boldsymbol{W}))$, $\mathbb{E}_\tau(\ln q(\tau))$ and $\mathbb{E}_\alpha(\ln q(\boldsymbol{\alpha}))$, this contribution is given by

$$
\begin{aligned}
\mathcal{L}_k(q) &= (\psi(a^*_{\tau_k}) - \ln b^*_{\tau_k} - \ln 2\pi)\frac{1}{2}\sum_n r_{nk} - \frac{1}{2}\frac{a^*_{\tau_k}}{b^*_{\tau_k}}\sum_n r_{nk}\mathbb{E}_W\left((y_n - \boldsymbol{w}_k^T\boldsymbol{x}_n)^2\right) \\
&\quad - \frac{D_w}{2}\ln 2\pi + \frac{D_w}{2}(\psi(a^*_{\alpha_k}) - \ln b^*_{\alpha_k}) - \frac{1}{2}\frac{a^*_{\alpha_k}}{b^*_{\alpha_k}}(\boldsymbol{w}_k^{*T}\boldsymbol{w}_k^* + \mathrm{Tr}(\boldsymbol{\Sigma}_k^*)) \\
&\quad - \ln\Gamma(a_\tau) + a_\tau \ln b_\tau + (a_\tau - 1)(\psi(a^*_{\tau_k}) - \ln b^*_{\tau_k}) - b_\tau\frac{a^*_{\tau_k}}{b^*_{\tau_k}} \\
&\quad - \ln\Gamma(a_\alpha) + a_\alpha \ln b_\alpha + (a_\alpha - 1)(\psi(a^*_{\alpha_k}) - \ln b^*_{\alpha_k}) - b_\alpha\frac{a^*_{\alpha_k}}{b^*_{\alpha_k}} \\
&\quad + \frac{1}{2}\ln|\boldsymbol{\Sigma}_k^*| + \frac{D_w}{2}(1 + \ln 2\pi) \\
&\quad + \ln\Gamma(a^*_{\tau_k}) - (a^*_{\tau_k} - 1)\psi(a^*_{\tau_k}) - \ln b^*_{\tau_k} + a^*_{\tau_k} \\
&\quad + \ln\Gamma(a^*_{\alpha_k}) - (a^*_{\alpha_k} - 1)\psi(a^*_{\alpha_k}) - \ln b^*_{\alpha_k} + a^*_{\alpha_k}.
\end{aligned}
$$

Observing that

$$
\begin{aligned}
a^*_{\tau_k} - a_\tau &= \frac{1}{2}\sum_n r_{nk}, \\
b^*_{\tau_k} - b_\tau &= \frac{1}{2}\sum_n r_{nk}\mathbb{E}_W\left((y_n + \boldsymbol{w}_k^T\boldsymbol{x}_n)^2\right), \\
a^*_{\alpha_k} - a_\alpha &= \frac{D_w}{2}, \\
b^*_{\alpha_k} - b_\alpha &= \frac{1}{2}\left(\boldsymbol{w}_k^{*T}\boldsymbol{w}_k^* + \mathrm{Tr}(\boldsymbol{\Sigma}_k^*)\right),
\end{aligned}
$$

we can simplify $\mathcal{L}_k(q)$ to

$$
\begin{aligned}
\mathcal{L}_k(q) &= -\ln\Gamma(a_\tau) + a_\tau\ln b_\tau + \ln\Gamma(a^*_{\tau_k}) - a^*_{\tau_k}\ln b^*_{\tau_k} + (a_\tau - a^*_{\tau_k})\ln 2\pi \\
&\quad - \ln\Gamma(a_\alpha) + a_\alpha\ln b_\alpha + \ln\Gamma(a^*_{\alpha_k}) - a^*_{\alpha_k}\ln b^*_{\alpha_k} \\
&\quad + \frac{1}{2}\ln|\boldsymbol{\Sigma}_k^*| + \frac{D_w}{2}.
\end{aligned}
$$

To get the full variational bound, we can use

$$
\begin{aligned}
a^*_{\beta_k} - a_\beta &= \frac{D_v}{2}, \\
b^*_{\beta_k} - b_\beta &= \frac{1}{2}\left(\tilde{\boldsymbol{v}}_k^T\tilde{\boldsymbol{v}}_k + \mathrm{Tr}\left((\tilde{\boldsymbol{\Lambda}}_V^{-1})_{kk}\right)\right),
\end{aligned}
$$

to get

$$
\begin{aligned}
\mathcal{L}(\tilde{q}) &= \sum_k \mathcal{L}_k(q) + \sum_k\left(-\ln\Gamma(a_\beta) + a_\beta\ln b_\beta + \ln\Gamma(a^*_{\beta_k}) - a^*_{\beta_k}\ln b^*_{\beta_k}\right) \\
&\quad + \frac{1}{2}\ln|\tilde{\boldsymbol{\Lambda}}_V^{-1}| + \frac{KD_v}{2} + \sum_n\sum_k r_{nk}\ln\frac{g_k(\boldsymbol{x}_n)|_{\boldsymbol{v}_k=\tilde{\boldsymbol{v}}_k}}{r_{nk}}.
\end{aligned}
$$

B.9 $\mathcal{L}(\tilde{q})$ for Independent Expert Training

When the experts are trained independently, their update equations are modified as described in Section A.8. Hence, our simplification based on the expression for $a^*_{\tau_k} - a_\tau$ and $b^*_{\tau_k} - b_\tau$ is not valid

anymore. Therefore, the expression of the contribution of expert k to $\mathcal{L}(\tilde{q})$ becomes

$$\begin{aligned}\mathcal{L}_k(q) &= (\psi(a^*_{\tau_k}) - \ln b^*_{\tau_k} - \ln 2\pi)\frac{1}{2}\sum_n r_{nk} - \frac{1}{2}\frac{a^*_{\tau_k}}{b^*_{\tau_k}}\sum_n r_{nk}\left(\|y_n - \boldsymbol{w}_k^T\boldsymbol{x}_n\|^2 + \boldsymbol{x}_n^T\boldsymbol{\Sigma}_k^*\boldsymbol{x}_n\right)\\ &\quad - \ln\Gamma(a_\tau) + a_\tau(\ln b_\tau - \ln b^*_{\tau_k}) + \ln\Gamma(a^*_{\tau_k}) - b_\tau\frac{a^*_{\tau_k}}{b^*_{\tau_k}}a^*_{\tau_k}\psi(a^*_{\tau_k}) + a^*_{\tau_k}\\ &\quad + \frac{1}{2}\ln|\boldsymbol{\Sigma}_k^*| + \frac{D_w}{2}\\ &\quad - \ln\Gamma(a_\alpha) + a_\alpha\ln b_\alpha + \ln\Gamma(a^*_{\alpha_k}) - a^*_{\alpha_k}\ln b^*_{\alpha_k}.\end{aligned}$$

The expression for $\mathcal{L}(\tilde{q})$ remains unchanged.

C Predictive Distribution

Given a new input $\hat{\boldsymbol{x}}$, we assume that $\hat{\boldsymbol{z}}$ is the associated latent variable, and want to find $p(\hat{y}|\hat{\boldsymbol{x}}, \boldsymbol{X}, \boldsymbol{Y})$ by evaluating

$$p(\hat{y}|\hat{\boldsymbol{x}}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_{\hat{\boldsymbol{z}}}\iiint p(\hat{y}|\hat{\boldsymbol{x}}, \hat{\boldsymbol{z}}, \boldsymbol{W}, \boldsymbol{\tau})p(\hat{\boldsymbol{z}}|\hat{\boldsymbol{x}}, \boldsymbol{V})p(\boldsymbol{W}, \boldsymbol{\tau}, \boldsymbol{V}|\boldsymbol{X}, \boldsymbol{Y})\mathrm{d}\boldsymbol{W}\mathrm{d}\boldsymbol{\tau}\mathrm{d}\boldsymbol{V}.$$

Using the definition of the various distributions and summing over $\hat{\boldsymbol{z}}$ we get

$$p(\hat{y}|\hat{\boldsymbol{x}}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_k\iiint g_k(\hat{\boldsymbol{x}})\mathcal{N}(\hat{y}|\boldsymbol{x}_k^T\hat{\boldsymbol{x}}, \tau_k^{-1})p(\boldsymbol{W}, \boldsymbol{\tau}, \boldsymbol{V}|\boldsymbol{X}, \boldsymbol{Y})\mathrm{d}\boldsymbol{W}\mathrm{d}\boldsymbol{\tau}\mathrm{d}\boldsymbol{V},$$

where we will approximate the posterior $p(\boldsymbol{W}, \boldsymbol{\tau}, \boldsymbol{V}|\boldsymbol{X}, \boldsymbol{Y})$ by the variational distribution $q^*_W(\boldsymbol{W})q^*_\tau(\boldsymbol{\tau})q^*_V(\boldsymbol{V})$. Thus, the predictive distribution becomes

$$p(\hat{y}|\hat{\boldsymbol{x}}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_k\int q^*_V(\boldsymbol{v}_k)g_k(\hat{\boldsymbol{x}})\mathrm{d}\boldsymbol{v}_k\iint q^*_W(\boldsymbol{w}_k)q^*_\tau(\tau_k)\mathcal{N}(\hat{y}|\boldsymbol{w}_k^T\hat{\boldsymbol{x}}, \tau_k^{-1})\mathrm{d}\boldsymbol{w}_k\mathrm{d}\boldsymbol{\tau}_k.$$

The first integral is $\mathbb{E}_V(g_k(\hat{\boldsymbol{x}}))$ which we approximate as in Ueda and Ghahramani (2002) by its MAP estimate

$$\int q^*_V(\boldsymbol{v}_k)g_k(\hat{\boldsymbol{x}})\mathrm{d}\boldsymbol{v}_k \approx g_k(\hat{\boldsymbol{x}})|_{\boldsymbol{v}_k=\tilde{\boldsymbol{v}}_k}.$$

The second integral is also solved as in Ueda and Ghahramani (2002) and results in the Student-t distribution

$$\iint q^*_W(\boldsymbol{w}_k)q^*_\tau(\tau_k)\mathcal{N}(\hat{y}|\boldsymbol{w}_k^T\hat{\boldsymbol{x}}, \tau_k^{-1})\mathrm{d}\boldsymbol{w}_k\mathrm{d}\boldsymbol{\tau}_k = \mathrm{St}\left(\hat{y}|\boldsymbol{w}_k^T\hat{\boldsymbol{x}}, \frac{a^*_{\tau_k}}{b^*_{\tau_k}}(1 + \hat{\boldsymbol{x}}^T\boldsymbol{\Sigma}_k^*\hat{\boldsymbol{x}})^{-1}, 2a^*_{\tau_k}\right).$$

Therefore, our predictive distribution is the mixture of Student-t distributions

$$p(\hat{y}|\hat{\boldsymbol{x}}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_k g_k(\hat{\boldsymbol{x}})|_{\boldsymbol{v}_k=\tilde{\boldsymbol{v}}_k}\mathrm{St}\left(\hat{y}|\boldsymbol{w}_k^T\hat{\boldsymbol{x}}, \frac{a^*_{\tau_k}}{b^*_{\tau_k}}(1 + \hat{\boldsymbol{x}}^T\boldsymbol{\Sigma}_k^*\hat{\boldsymbol{x}})^{-1}, 2a^*_{\tau_k}\right),$$

with mean

$$\mathbb{E}(\hat{y}|\hat{\boldsymbol{x}}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_k g_k(\hat{\boldsymbol{x}})|_{\boldsymbol{v}_k=\tilde{\boldsymbol{v}}_k}\boldsymbol{w}_k^T\hat{\boldsymbol{x}}.$$

References

Bernadó-Mansilla, E., Llorá, X., & Garrell-Guiu, J. M. (2002). XCS and GALE: A Comparative Study of Two Learning Classifier Systems on Data Mining. *IWLCS '01: Revised Papers from the 4th International Workshop on Advances in Learning Classifier Systems* (pp. 115–132). London, UK: Springer.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer.

Bishop, C. M., & Svensén, M. (2003). Bayesian Hierarchical Mixtures of Experts. *Proceedings of the 19th Annual Conference on Uncertainty in Artificial Intelligence (UAI-03)* (pp. 57–64). San Francisco, CA: Morgan Kaufmann.

Butz, M. V. (2006). *Rule-Based Evolutionary Online Learning Systems: A Principled Approach to LCS Analysis and Design*, vol. 191 of *Studies in Fuzziness and Soft Computing*. Springer.

Butz, M. V., Lanzi, P. L., & Wilson, S. W. (to appear). Function Approximation with XCS: Hyperellipsoidal Conditions, Recursive Least Squares, and Compaction. *IEEE Transactions on Evolutionary Computations*.

Chipman, H. A., George, E. I., & McCulloch, R. E. (1998). Bayesian cart model search. *Journal of the American Statistical Association*, *93*, 935–948.

Gibbs, M. N. (1997). *Bayesian gaussian processes for regression and classification*. Doctoral dissertation, University of Cambridge.

Jacobs, R. A., Jordan, M. I., Nowlan, S., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, *3*, 1–12.

Jordan, M. I., & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, *6*, 181–214.

Ueda, N., & Ghahramani, Z. (2002). Bayesian model search for mixture models based on optimizing variational bounds. *Neural Networks*, *15*, 1223–1241.

Wainwright, M., Jaakkola, T., , & Willsky, A. (2005). A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory*, *51*, 2313–2335.

Waterhouse, S. (1997). *Classification and Regression using Mixtures of Experts*. Doctoral dissertation, Department of Engineering, University of Cambridge.

Waterhouse, S., MacKay, D., & Robinson, T. (1996). Bayesian Methods for Mixtures of Experts. *Advances in Neural Information Processing Systems 8* (pp. 351–357). Cambridge, MA: MIT Press.

Xu, L., I.Jordan, M., & Hinton, G. E. (1995). An Alternative Model for Mixtures of Experts. *Advances in Neural Information Processing Systems 7* (pp. 633–640). Cambridge, MA: MIT Press.