Ensemble Learning and Evidence Maximization

David J.C. MacKay
Cavendish Laboratory
Madingley Road, Cambridge CB3 0HE
United Kingdom
mackay@mrao.cam.ac.uk

May 1995—Submitted to NIPS 1995

Abstract

Ensemble learning by variational free energy minimization is a tool introduced to neural networks by Hinton and van Camp in which learning is described in terms of the optimization of an ensemble of parameter vectors. The optimized ensemble is an approximation to the posterior probability distribution of the parameters. This tool has now been applied to a variety of statistical inference problems.

In this paper I study a linear regression model with both parameters and hyperparameters. I demonstrate that the evidence approximation for the optimization of regularization constants can be derived in detail from a free energy minimization viewpoint.

1 Ensemble Learning by Free Energy Minimization

A new tool has recently been introduced into the field of neural networks and statistical inference. In traditional approaches to neural networks, a single parameter vector $\mathbf{w}$ is optimized by maximum likelihood or penalized maximum likelihood. In the Bayesian interpretation, these optimized parameters are viewed as defining the mode of a posterior probability distribution $P(\mathbf{w}|D,\mathcal{H})$ (given data D and model assumptions $\mathcal{H}$), which can be approximated, with a Gaussian distribution $\tilde{P}$ for example (MacKay 1992b), in order to obtain predictive distributions and optimize model control parameters.

The new concept introduced by Hinton and van Camp (1993) is to work in terms of an approximating *ensemble* $Q(\mathbf{w};\theta)$, that is, a probability distribution over the parameters, and optimize the ensemble (by varying its own parameters θ) so that it approximates the posterior distribution of the parameters $P(\mathbf{w}|D,\mathcal{H})$ well. The objective function chosen to measure the quality of the approximation is a *variational free energy*,[1]

$$F(\theta) = -\int d^k\mathbf{w}\, Q(\mathbf{w};\theta) \log \frac{P(D|\mathbf{w},\mathcal{H})P(\mathbf{w}|\mathcal{H})}{Q(\mathbf{w};\theta)} \tag{1}$$

The free energy $F(\theta)$ is bounded below by $-\log P(D|\mathcal{H})$ and only attains this value for $Q(\mathbf{w};\theta) = P(\mathbf{w}|D,\mathcal{H})$. $F(\theta)$ can be viewed as the sum of $-\log P(D|\mathcal{H})$ and the Kullback-

[1] Variational free energy minimization is a well-established tool in statistical physics (Feynman 1972); 'mean field theory' is an important special case. The free energy can also be described in terms of description lengths.

Leibler divergence between $Q(\mathbf{w};\theta)$ and $P(\mathbf{w}|D,\mathcal{H})$. For certain models and certain approximating distributions, this free energy, and its derivatives with respect to the ensemble's parameters, can be evaluated.

Hinton and van Camp (1993) considered a regression network with one non-linear hidden layer and showed that a *separable* Gaussian approximating distribution $Q(\mathbf{w};\theta)$ can be optimized with a deterministic algorithm.

Hinton and Zemel (1994) have applied the same approach to the optimization of an autoencoder. The hidden-to-output part of an autoencoder is viewed as defining a generative model employing latent variables that live in the hidden layer of the model. The optimization of such a generative model is challenging, requiring, for every given data example, an implicit or explicit computation of the posterior probability distribution P of the latent variables. Hinton and Zemel (1994) view the input-to-hidden 'recognition' part of the autoencoder as defining an approximating distribution Q for this distribution P. A single objective function F can then be defined for simultaneous optimization of the generative model and the recognition model. The Helmholtz machine (Dayan *et al.* 1995) is a further generalization of these ideas.

In a broader statistical context, Neal and Hinton (1993) have shown that it is possible to view the Expectation-Maximization (EM) algorithm in terms of a free energy minimization. The deterministic Boltzmann machine can be derived as a free energy approximation to the Boltzmann machine (Radford Neal, personal communication). And MacKay (1995a) has obtained an algorithm for decoding certain binary codes by variational free energy minimization.

In this paper I study a free energy approximation for a linear regression model with an unknown hyperparameter. This model captures the essence of an important problem in neural networks, namely how to set regularization constants or weight decay parameters.

2 Inference of parameters and hyperparameters

There has been a debate over the appropriateness of the generalized maximum likelihood method, also known as the evidence framework (Gull 1989; MacKay 1992a), for controlling hyperparameters in linear and non-linear regression models (Wolpert 1993; MacKay 1994). In this section I demonstrate that, for linear models, a simple free energy minimization approximation reproduces the method of the evidence framework precisely. This demonstration further clarifies the sense in which the evidence approximation is a reasonable method of computationally tractable inference.

2.1 The linear regression model with regularization

The statistical model is as follows:

$$P(D,\mathbf{w},\alpha,\beta|\mathcal{H}) = P(D|\mathbf{w},\beta,\mathcal{H})P(\mathbf{w}|\alpha,\mathcal{H})P(\alpha,\beta|\mathcal{H}), \tag{2}$$

where D is the data, $\mathbf{w}$ is the parameter vector, of dimension k, β defines a noise variance $\sigma_\nu^2 = 1/\beta$, and α is a regularization constant. In a regression problem, for example, D might be a set of data points, $\{\mathbf{x},\mathbf{t}\}$, and the vector $\mathbf{w}$ might parameterize a function $f(\mathbf{x};\mathbf{w})$. The model $\mathcal{H}$ states that for some $\mathbf{w}$, the dependent variables $\{\mathbf{t}\}$ are given by adding noise to $\{f(\mathbf{x};\mathbf{w})\}$; the likelihood function $P(D|\mathbf{w},\beta,\mathcal{H})$ describes the assumed noise process, parameterized by a noise level $1/\beta$; the prior probability of the parameters, $P(\mathbf{w}|\alpha,\mathcal{H})$,

embodies assumptions about the spatial correlations and smoothness that the true function is expected to have, parameterized by a regularization constant α. The variables α and β are known as hyperparameters. Problems for which models can be written in the form (2) include linear interpolation with a fixed basis set (Gull 1988; MacKay 1992a), non-linear regression with a neural network (MacKay 1992b), and image deconvolution (Gull 1989).

In the simplest case (linear models, Gaussian noise), the first factor in (2), the likelihood, can be written in terms of a quadratic function of $\mathbf{w}$, $E_D(\mathbf{w})$:

$$P(D|\mathbf{w},\beta,\mathcal{H}) = \frac{1}{Z_D(\beta)} \exp(-\beta E_D(\mathbf{w})). \tag{3}$$

What makes the problem 'ill-posed' is that the hessian $\nabla\nabla E_D$ is ill-conditioned — some of its eigenvalues are very small, so that the maximum likelihood parameters depend undesirably on the noise in the data. The model is 'regularized' by the second factor in (2), the prior, which in the simplest case is a spherical Gaussian:

$$P(\mathbf{w}|\alpha,\mathcal{H}) = \frac{1}{Z_W(\alpha)} \exp(-\alpha \tfrac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}). \tag{4}$$

The regularization constant α defines the variance $\sigma_W^2 = 1/\alpha$ of the prior for the components w_i of $\mathbf{w}$.

Finally, a gamma distribution prior is assumed for α and β, $P(\alpha|\mathcal{H}) = \Gamma(\alpha; b_\alpha, c_\alpha)$, where this notation means:

$$\Gamma(\alpha; b_\alpha, c_\alpha) = \frac{1}{\Gamma(c_\alpha)} \frac{\alpha^{c_\alpha - 1}}{b_\alpha^{c_\alpha}} \exp\left(-\frac{\alpha}{b_\alpha}\right), 0 \leq \alpha < \infty$$

This distribution has mean $b_\alpha c_\alpha$ and variance $b_\alpha^2 c_\alpha$.

In what follows I will for simplicity assume that β is known and fixed, and that the model is indeed a linear model (*i.e.*, E_D is a quadratic function of $\mathbf{w}$).

2.2 Review of the evidence framework

The evidence framework divides our inferences into distinct 'levels of inference':
Level 1: Infer the parameters $\mathbf{w}$ for a given value of α:

$$P(\mathbf{w}|D,\alpha,\mathcal{H}) = \frac{P(D|\mathbf{w},\alpha,\mathcal{H})P(\mathbf{w}|\alpha,\mathcal{H})}{P(D|\alpha,\mathcal{H})}. \tag{5}$$

Level 2: Infer α:

$$P(\alpha|D,\mathcal{H}) = \frac{P(D|\alpha,\mathcal{H})P(\alpha|\mathcal{H})}{P(D|\mathcal{H})}. \tag{6}$$

Level 3: Compare models:

$$P(\mathcal{H}|D) \propto P(D|\mathcal{H})P(\mathcal{H}). \tag{7}$$

There is a pattern in these three applications of Bayes' rule: at each of higher levels 2 and 3, the data-dependent factor (*e.g.* in level 2, $P(D|\alpha,\mathcal{H})$) is the normalizing constant (the 'evidence') from the preceding level of inference.

The evidence framework obtains approximate inferences using the following procedure.

- The level 1 inference is approximated by making a quadratic expansion of $\log P(D|\mathbf{w},\alpha,\mathcal{H})P(\mathbf{w}|\alpha,\mathcal{H})$ around a maximum of $P(\mathbf{w}|D,\alpha,\mathcal{H})$; this expansion defines a Gaussian approximation to the posterior. The evidence $P(D|\alpha,\mathcal{H})$ is estimated by evaluating the appropriate determinant. For linear models the Gaussian approximation is exact.

- By maximizing the evidence $P(D|\alpha,\mathcal{H})$ at level 2, we find the most probable value of the regularization constant, $\alpha_{\rm MP}$, and error bars on it, $\sigma_{\log\alpha|D}$. (Because α is a positive scale variable, it is natural to represent its uncertainty on a log scale.)

- The value of $\alpha_{\rm MP}$ is substituted at level 1. This defines a probability distribution $P(\mathbf{w}|D,\alpha_{\rm MP},\mathcal{H})$ which is intended as a 'good approximation' to the posterior $P(\mathbf{w}|D,\mathcal{H})$; this distribution is a Gaussian around the maximum, $\mathbf{w}_{\rm MP|\alpha_{\rm MP}}$, with covariance matrix $\mathbf{\Sigma}$ defined by $\mathbf{\Sigma}^{-1} = -\nabla\nabla\log P(\mathbf{w}|D,\alpha_{\rm MP},\mathcal{H})$. Marginals for the components of $\mathbf{w}$ are easily obtained from this distribution.

- Predictive distributions $P(D_2|D,\mathcal{H})$ are approximated by using the posterior distribution with $\alpha{=}\alpha_{\rm MP}$:

$$P(D_2|D,\alpha_{\rm MP},\mathcal{H}) = \int d^k\mathbf{w}\ P(D_2|\mathbf{w},\mathcal{H})P(\mathbf{w}|D,\alpha_{\rm MP},\mathcal{H}). \tag{8}$$

  For a locally linear model with Gaussian noise, both the distributions inside the integral are Gaussian, and this integral is straightforward to perform.

As reviewed in MacKay (1992a), the most probable value of α satisfies a simple and intuitive implicit equation,

$$\frac{1}{\alpha_{\rm MP}} = \frac{\sum_1^k w_i^2}{\gamma} \tag{9}$$

where w_i are the components of the vector $\mathbf{w}_{\rm MP|\alpha_{\rm MP}}$ and γ is the *number of well-determined parameters*:

$$\gamma = k - \alpha{\rm Trace}\mathbf{\Sigma}. \tag{10}$$

This quantity is a number between 0 and k. Recalling that α can be interpreted as the variance σ_W^2 of the distribution from which the parameters w_i come, we see that equation (9) corresponds to an intuitive prescription for a variance estimator. The idea is that we are estimating the variance of the distribution of w_i from only γ well-determined parameters, the other $(k-\gamma)$ having been set roughly to zero by the regularizer and therefore not contributing to the sum in the numerator.

In principle, there may be multiple optima in α, but this is not the typical case for a model well matched to the data. Under general conditions, the error bars on $\log\alpha$ are $\sigma_{\log\alpha|D} \simeq \sqrt{2/\gamma}$ (MacKay 1992a; MacKay 1994). Thus $\log\alpha$ is well-determined by the data if $\gamma \gg 1$.

The central computation can be summarised thus:

> **Evidence approximation**: find the self-consistent solution $\{\mathbf{w}_{\rm MP|\alpha_{\rm MP}}, \alpha_{\rm MP}\}$ such that $\mathbf{w}_{\rm MP|\alpha_{\rm MP}}$ maximizes $P(\mathbf{w}|D,\alpha_{\rm MP},\mathcal{H})$ and $\alpha_{\rm MP}$ satisfies equation (9).

Justifications for this approximation are given in (MacKay 1995b; MacKay 1994), where correction terms of order $1/\sqrt{\gamma}$ are also given.

2.3 Free energy approximation

Let us consider approximating the joint distribution of $\mathbf{w}$ and α given the data,

$$P(\mathbf{w}, \alpha | D, \mathcal{H}) = \frac{P(D|\mathbf{w}, \mathcal{H}) P(\mathbf{w}|\alpha, \mathcal{H}) P(\alpha|\mathcal{H})}{P(D|\mathcal{H})}, \tag{11}$$

by a distribution $Q(\mathbf{w}, \alpha)$. I make one assumption only, namely that our approximation is separable into the form $Q(\mathbf{w}, \alpha) = Q_{\mathbf{w}}(\mathbf{w}) Q_\alpha(\alpha)$. *No functional form for these distributions is assumed.* We write down a variational free energy,

$$F(Q) = - \int d\mathbf{w}\, d\alpha\, Q_{\mathbf{w}}(\mathbf{w}) Q_\alpha(\alpha) \log \frac{P(D|\mathbf{w}, \mathcal{H}) P(\mathbf{w}|\alpha, \mathcal{H}) P(\alpha|\mathcal{H})}{Q_{\mathbf{w}}(\mathbf{w}) Q_\alpha(\alpha)}. \tag{12}$$

This functional is bounded below by the evidence for the model thus: $F \geq - \log P(D|\mathcal{H})$, with equality only attained if $Q(\mathbf{w}, \alpha) = P(\mathbf{w}, \alpha | D, \mathcal{H})$. We can find the optimal separable distribution Q by considering separately the optimization of F over $Q_{\mathbf{w}}(\mathbf{w})$ for fixed $Q_\alpha(\alpha)$, and then the optimization of $Q_\alpha(\alpha)$ for fixed $Q_{\mathbf{w}}(\mathbf{w})$.

2.4 Optimization of $Q_{\mathbf{w}}(\mathbf{w})$

As a functional of $Q_{\mathbf{w}}(\mathbf{w})$, F is:

$$\begin{aligned} F &= - \int d\mathbf{w}\, Q_{\mathbf{w}}(\mathbf{w}) \left[\int d\alpha\, Q_\alpha(\alpha) \log P(\mathbf{w}|\alpha) + \log P(D|\mathbf{w}, \mathcal{H}) - \log Q(\mathbf{w}) \right] + \text{const.} \\ &= \int d\mathbf{w}\, Q_{\mathbf{w}}(\mathbf{w}) \left[\int d\alpha\, Q_\alpha(\alpha) \alpha \frac{1}{2} \mathbf{w}\mathbf{w}^{\mathrm{T}} + \beta E_D(\mathbf{w}) + \log Q(\mathbf{w}) \right] + \text{const.}' \end{aligned}$$

The dependence on Q_α thus collapses down to a dependence simply on the mean $\bar{\alpha} \equiv \int d\alpha\, Q_\alpha(\alpha) \alpha$.

$$F = \int d\mathbf{w}\, Q_{\mathbf{w}}(\mathbf{w}) \left[\bar{\alpha} \frac{1}{2} \mathbf{w}\mathbf{w}^{\mathrm{T}} + \beta E_D(\mathbf{w}) + \log Q(\mathbf{w}) \right] + \text{const.}'$$

Noting that the $\mathbf{w}$-dependent terms $-\bar{\alpha} \frac{1}{2} \mathbf{w}\mathbf{w}^{\mathrm{T}} - \beta E_D(\mathbf{w})$ are the log of a posterior distribution, and using the theorem that a divergence $\int Q \log(Q/P)$ is minimized by setting $Q = P$, we can immediately write down the distribution $Q_{\mathbf{w}}(\mathbf{w})$ that minimizes this expression. Thus for given data D and Q_α, the optimizing distribution $Q_{\mathbf{w}}^{\text{opt}}(\mathbf{w})$ is a Gaussian identical to the posterior distribution for a particular value of $\alpha = \bar{\alpha}$.

$$Q_{\mathbf{w}}^{\text{opt}}(\mathbf{w}) = P(\mathbf{w}|D, \bar{\alpha}, \mathcal{H}) = \text{Normal}(\mathbf{w}_{\text{MP}|\bar{\alpha}}, \mathbf{\Sigma}). \tag{13}$$

2.5 Optimization of $Q_\alpha(\alpha)$

As a functional of $Q_\alpha(\alpha)$, F is:

$$\begin{aligned} F(Q) &= - \int d\alpha\, Q_\alpha(\alpha) \left[\int d\mathbf{w}\, Q_{\mathbf{w}}(\mathbf{w}) \log P(\mathbf{w}|\alpha, \mathcal{H}) + \log P(\alpha|\mathcal{H}) - \log Q_\alpha(\alpha) \right] + \text{const.} \\ &= \int d\alpha\, Q_\alpha(\alpha) \left[\frac{\alpha}{2} \int d\mathbf{w}\, Q_{\mathbf{w}}(\mathbf{w}) \mathbf{w}^{\mathrm{T}} \mathbf{w} - \frac{k}{2} \log \alpha - (c_\alpha - 1) \log \alpha + \frac{\alpha}{b_\alpha} + \log Q_\alpha(\alpha) \right] \\ &= \int d\alpha\, Q_\alpha(\alpha) \left[\left(\frac{1}{2} \mathbf{w}_{\text{MP}|\bar{\alpha}}{}^{\mathrm{T}} \mathbf{w}_{\text{MP}|\bar{\alpha}} + \frac{1}{2} \text{Trace} \mathbf{\Sigma} + \frac{1}{b_\alpha} \right) \alpha \right. \\ &\qquad \left. - \left(\frac{k}{2} + c_\alpha - 1 \right) \log \alpha + \log Q_\alpha(\alpha) \right] + \text{const.}' \end{aligned}$$

where c_α, b_α are the parameters of the gamma prior on α. Here, the α-dependent expression in the brackets can be recognized as the log of a gamma distribution, giving as the optimal distribution that minimizes F for fixed $Q_{\mathbf{w}}$:

$$Q_\alpha^{\rm opt}(\alpha) = \Gamma(\alpha; b', c') \tag{14}$$

where

$$\begin{aligned} 1/b' &= 1/b_\alpha + \tfrac{1}{2}\mathbf{w}_{{\rm MP}|\bar{\alpha}}{}^{\rm T}\mathbf{w}_{{\rm MP}|\bar{\alpha}} + \tfrac{1}{2}{\rm Trace}\boldsymbol{\Sigma} \\ c' &= k/2 + c_\alpha \end{aligned} \tag{15}$$

This completes our derivation of the free energy optimization. The optimal approximating distribution is given by finding the gamma distribution for α and the normal distribution for $\mathbf{w}$ that satisfy the simultaneous equations (13) and (15).

2.6 Comparison with evidence framework

To understand this result we complete the loop by evaluating the mean $\bar{\alpha}'$ for this optimized gamma distribution, which is:

$$\bar{\alpha}' = b'c' = \frac{\frac{k}{2} + c_\alpha}{\frac{1}{b_\alpha} + \frac{1}{2}\mathbf{w}_{{\rm MP}|\bar{\alpha}}{}^{\rm T}\mathbf{w}_{{\rm MP}|\bar{\alpha}} + \frac{1}{2}{\rm Trace}\boldsymbol{\Sigma}} \tag{16}$$

In the special case of an uninformative prior on α ($c_\alpha \to 0$ and $\frac{1}{b_\alpha} \to 0$) we obtain:

$$\bar{\alpha}' = \frac{k}{\mathbf{w}_{{\rm MP}|\bar{\alpha}}{}^{\rm T}\mathbf{w}_{{\rm MP}|\bar{\alpha}} + {\rm Trace}\boldsymbol{\Sigma}}. \tag{17}$$

Is this the same optimal α as that[2] found by the evidence approximationΓ The answer is yes. Substituting (equation 9) $\mathbf{w}_{{\rm MP}|\alpha_{\rm MP}}^{\rm T}\mathbf{w}_{{\rm MP}|\alpha_{\rm MP}} = \gamma/\alpha_{\rm MP}$, and using $\gamma = k - \alpha{\rm Trace}\boldsymbol{\Sigma}$, we find that if we set $\alpha = \bar{\alpha} = \alpha_{\rm MP}$ on the right hand side we obtain

$$\bar{\alpha}' = \frac{k}{\gamma/\bar{\alpha} + (k-\gamma)/\bar{\alpha}} = \bar{\alpha}. \tag{18}$$

Thus any optimum of the evidence approximation is also a minimum of the free energy.

2.7 Intuition for this relationship

These two approaches give complementary views of the task of inferring α given the data.

In the evidence framework we examine the optimized value of $\mathbf{w}$, $\mathbf{w}_{{\rm MP}|\alpha}$, and think of $(\mathbf{w}_{{\rm MP}|\alpha})^2$ as giving information about the variance σ_W^2 of the prior distribution of $\mathbf{w}$. The maximum likelihood estimator would be $\sigma_{W({\rm ML})}^2 = (\mathbf{w}_{{\rm MP}|\alpha})^2/k$, but the evidence framework modifies this estimator to take into account the fact that some of the k parameters have not been determined by the data, and have effectively been set to zero by the prior. Thus the variance estimate replaces k by the effective number of well determined parameters γ: $\sigma_{W({\rm MP})}^2 = (\mathbf{w}_{{\rm MP}|\alpha})^2/\gamma$.

The free energy minimization approach is like an EM algorithm, in which we wish to find the most probable α and do this by introducing an E-step in which a distribution over $\mathbf{w}$ is obtained. This distribution takes into account the ill-determinedness of the $k - \gamma$ ill-determined parameters by assigning each of them a variance of σ_W^2 in the matrix

[2] Or 'are *these* the same as *those* found by the evidence approximation?' if there are multiple optima.

Σ. Then when the M-step occurs, finding the optimal α, the maximum likelihood equation $\sigma^2_{W(\mathrm{ML})} = (\mathbf{w}_{\mathrm{MP}|\alpha})^2/k$ is modified by adding these variance terms to the numerator: $\sigma^2_{W(\mathrm{FE})} = \left[(\mathbf{w}_{\mathrm{MP}|\alpha})^2 + \mathrm{Trace}\boldsymbol{\Sigma}\right]/k$.

Thus evidence maximization decrements the denominator of the equation $\sigma^2_{W(\mathrm{ML})} = (\mathbf{w}_{\mathrm{MP}|\alpha})^2/k$ to take into account the smallness of the ill-determined parameters, whereas free energy minimization increments the numerator to take into account their variability. As we have seen, the two formulae converge on the identical result.

2.8 Further work on this model

There are two small differences between previous Bayesian results and the results of the free energy minimization.

1. The variance of the optimized gamma distribution for α is, in the limit of the uninformative prior,
$$\mathrm{var}(\alpha) = b'^2 c' = 2k/(k/\bar{\alpha})^2 = \bar{\alpha}^2/k \tag{19}$$
so that $\log \alpha$ has standard error $\sqrt{2/k}$. This contrasts with the result $\sqrt{2/\gamma}$ from the evidence framework.

2. This free energy approximation for $Q_{\mathbf{w}}(\mathbf{w})$ fails to produce the small order correction terms identified in (MacKay 1994), which arise because of the uncertainty in α.

It will be interesting to investigate whether a more complex approximating distribution Q might capture these terms.

2.9 Discussion

This result gives insight into the properties of both the evidence framework and ensemble learning. An additional spin-off is a convergence proof (at least for linear models) for a re-estimation formula for α (equation 16) which previous work on the evidence framework had not provided. The steps of re-estimating $\bar{\alpha}$ and computing the new distribution $Q_{\mathbf{w}}(\mathbf{w})$ both decrease F, and F is bounded below, so the iterative procedure converges.

3 Work in progress

In the final version of this paper (in preparation) I will describe work on two more simple models and one more complex model which capture the essence of other statistical problems of relevance to neural network regression models and classifiers:

1. **The inference of an unknown mean and standard deviation.** This example highlights the problem of inferring a noise level. Maximum likelihood noise level estimates are overconfident (hence the distinction between the σ_N and σ_{N-1} buttons on a calculator). I compare free energy approximations with the ideal solution obtained by Bayesian marginalization.

2. **The predictive distribution of a classifier whose parameters are uncertain.** In this problem marginalization is also important, but the outcome of the free energy approximation has a different character.

3. **Mixture models, including mixtures of Gaussians and the hierarchical mixture of experts.** This work starts from Neal and Hinton's view of the EM algorithm as a free energy minimization and generalizes it to include distributions over the parameters and hyperparameters too.

Acknowledgements

I thank Radford Neal, Geoff Hinton and Steve Waterhouse for helpful discussions.

References

DAYAN, P., HINTON, G. E., NEAL, R. M., and ZEMEL, R. S. (1995) The Helmholtz machine. *Neural Computation*. to appear.

FEYNMAN, R. P. (1972) *Statistical Mechanics*. W. A. Benjamin, Inc.

GULL, S. F. (1988) Bayesian inductive inference and maximum entropy. In *Maximum Entropy and Bayesian Methods in Science and Engineering, vol. 1: Foundations*, ed. by G. Erickson and C. Smith, pp. 53–74, Dordrecht. Kluwer.

GULL, S. F. (1989) Developments in maximum entropy data analysis. In *Maximum Entropy and Bayesian Methods, Cambridge 1988*, ed. by J. Skilling, pp. 53–71, Dordrecht. Kluwer.

HINTON, G. E., and VAN CAMP, D., (1993) Keeping neural networks simple by minimizing the description length of the weights. In: *Proceedings of COLT-93.*

HINTON, G. E., and ZEMEL, R. S. (1994) Autoencoders, minimum description length and Helmholtz free energy. In *Advances in Neural Information Processing Systems 6*, ed. by J. D. Cowan, G. Tesauro, and J. Alspector, San Mateo, California. Morgan Kaufmann.

MACKAY, D. J. C. (1992a) Bayesian interpolation. *Neural Computation* **4** (3): 415–447.

MACKAY, D. J. C. (1992b) A practical Bayesian framework for backpropagation networks. *Neural Computation* **4** (3): 448–472.

MACKAY, D. J. C., (1994) Hyperparameters: Optimize, or integrate outΓ Submitted to *Neural Computation.*

MACKAY, D. J. C. (1995a) Free energy minimization algorithm for decoding and cryptanalysis. *Electronics Letters* **31** (6): 446–447.

MACKAY, D. J. C. (1995b) Hyperparameters: Optimize, or integrate outΓ In *Maximum Entropy and Bayesian Methods, Santa Barbara 1993*, ed. by G. Heidbreder, Dordrecht. Kluwer.

NEAL, R. M., and HINTON, G. E. (1993) A new view of the EM algorithm that justifies incremental and other variants. *Biometrika.* submitted.

WOLPERT, D. H. (1993) On the use of evidence in neural networks. In *Advances in Neural Information Processing Systems 5*, ed. by C. L. Giles, S. J. Hanson, and J. D. Cowan, pp. 539–546, San Mateo, California. Morgan Kaufmann.