INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA

UMI®
800-521-0600

DISSERTATION

SADDLEPOINT METHODS IN NEURAL NETWORKS

Submitted by

Robert L. Paige

Department of Statistics

In partial fulfillment of the requirements

for the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 1999

COLORADO STATE UNIVERSITY

July 8, 1999

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY ROBERT L. PAIGE ENTITLED SADDLEPOINT METHODS IN NEURAL NETWORKS BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

Committee Member

Committee Member

Committee Member

Adviser

Department Head

ABSTRACT OF DISSERTATION

SADDLEPOINT METHODS IN NEURAL NETWORKS

Saddlepoint methods have found increased application in recent years. They provide fast and accurate approximations to some the most important integrals encountered in Statistics. In addition, when used to approximate probability distributions, they yield accurate tail probabilities. Competing methods, such as those based on stochastic simulation, are much slower and typically yield very poor approximations to the tail of a distribution. However, as we shall see in the forthcoming chapters, saddlepoint methods routinely provide virtually exact approximations to many quantities of interest besides probability distributions.

In the first chapter, Laplace's method is used to perform marginal inference in Bayesian neural networks. Accurate approximations for Bayes factors for model choice about the number of nonlinear sigmoidal terms; predictive densities for a future observable; Bayes estimates for the nonlinear regression function; and the marginal densities are given. Important use is made of the inherent partial linearity of the regression function and the lack of identifiability. The choice of prior and the use of an alternative sigmoidal lead to posterior invariance in the nonlinear parameter which is discussed in connection with the lack of sigmoidal identifiability. The accuracy of the Laplace approximations is illustrated in the context of two nonlinear data sets: a nonlinear regression model and a nonlinear autoregressive time series.

In chapter two, saddlepoint approximations and Laplace's method are used to study classification in the stochastic Hopfield model (SHM). First, the methodology is developed to provide saddlepoint approximations to classification time distributions. Secondly, saddlepoint approximations to the stationary distribution of the

Hopfield Markov chain, the Markov chain underlying the SHM's classification process, are presented. These approximations are particularly difficult to obtain since this stationary distribution has an intractable moment generating function which we approximate with Laplace's method. Lastly, a characterization of the set of possible absorbing states of the Hopfield Markov chain for the deterministic Hopfield model, a forerunner of the SHM, is provided. All of our contributions are a result of the lumpability of the Hopfield Markov chain which is rigorously derived and proven. The accuracy of the saddlepoint methods, in the above classification problems, is demonstrated on a SHM with $2^{10} = 1024$ states which reduced to a model with a mere 64 states via lumping.

Robert L. Paige
Department of Statistics
Colorado State University
Fort Collins, Colorado 80523
Summer 1999

Contents

1 Bayesian Inference in Neural Networks

1.1 Introduction

Neural network (NN) models have enjoyed considerable popularity in recent years. They have been applied to many nonlinear regression problems and have been used extensively for time series prediction, see Faraway and Chatfield (1998) and Weigend and Gershenfeld (1993). They also have been successfully applied in pattern recognition as reviewed in Bishop (1995) and Ripley (1996).

In the NN model, the expectation of observation y_i is a "quasi-linear" function of independent variable $x_i = (1, x_{i1}, \ldots, x_{im})^T$ for $i = 1, \ldots, n$. With $y = (y_1, y_2, \ldots y_n)^T$, the model is

$$y = X_\alpha \beta + \tau\varepsilon$$

where $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T \sim N(0, I_n)$, $\tau > 0$ is a scale parameter, X_α is an $n \times p$ design matrix of the form

$$X_\alpha = \begin{bmatrix} x_1^T & \sigma(x_1^T\alpha_1) & \cdots & \sigma(x_1^T\alpha_q) \\ x_2^T & \sigma(x_2^T\alpha_1) & \cdots & \sigma(x_2^T\alpha_q) \\ \vdots & \vdots & \ddots & \vdots \\ x_n^T & \sigma(x_n^T\alpha_1) & \cdots & \sigma(x_n^T\alpha_q) \end{bmatrix}, \tag{1.1}$$

and β is a $(p \times 1)$ linear regression vector with $!p = m + 1 + q$. The elements of X_α are functions of the nonlinear parameter matrix $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_q)^T$ where $\alpha_i = (\alpha_{i0}, \alpha_{i1}, \ldots, \alpha_{im})^T$ and $\sigma(x)$ is a sigmoidal usually taken to be the logistic function $(1 + e^{-x})^{-1}$. The convention has been to work with both y- and x-component values translated and scaled to be in the range $[0, 1]$.

In neural network parlance this particular type of model is known as a "feed-forward neural network with one hidden layer, one skip layer, and a linear output unit", as described in Ripley (1996). "Feed-forward" refers to the vector x_i being

mapped or "fed" forward to give $E[y_i]$. This expectation is a linear combination of the nonlinear sigmoidal terms, $\sigma(x_i^T\alpha_1), \ldots, \sigma(x_i^T\alpha_q)$, or "hidden layer" and the linear term, x_i^T, or "skip layer" which "skips" past the sigmoidal transformation. In addition, this model has a "linear output unit" since the expectation of y is linear in the regression function $X_\alpha\beta$.

The NN model is a highly flexible model. With sufficiently many sigmoidal terms, the regression function $X_\alpha\beta$ can approximate any continuous function uniformly over a compact set to an arbitrary degree of precision as first shown by Cybenko (1988). Usually relatively few sigmoidal terms are required in statistical problems. Nonetheless there is a need to avoid overfitting the data with either abruptly changing or "saturated" sigmoidal terms or too many sigmoidals. We use a prior distribution on (α, β, τ^2) that a prior favours smooth and non-abruptly changing sigmoidal terms. A similar approach has been used in Ripley's S-Plus routine "nnet" as described in Venables and Ripley (1997).

This paper addresses several aspects of marginal Bayesian inference for such NN models. First the issue of model choice is considered. If the model in (1.1) with q sigmoidals is referred to as M_q, then marginal posterior probabilities for the various models $\{M_q : q = 0, 1, \ldots\}$ are approximated using the method of Laplace. These computations determine approximate Bayes factors for use in model choice (determination of q) as well as in posterior mixing over models for prediction and estimation. We show in the numerical examples that the Bayes factors alone do not convey the information needed to select a parsimonious model with small q. Models with extra sigmoidal terms that nest this parsimonious choice are weighted as heavily as the parsimonious model. However, with the additional information about the modal estimates used in Laplace approximations for Bayes factors, this parsimonious choice can be made.

Approximate predictive densities for future observable Y_f given its associated independent variable x_f are computed using Laplace approximations. Also computed are approximate Bayes estimates for the nonlinear regression relationship and approximate marginal posteriors on model parameters.

Our marginal inference is facilitated by making important use of the "partial linearity" in the regression parameters. This term refers to the fact that if α were known, then the NN model would be a linear regression model in β. As such, any marginal inference may be simplified considerably by marginalizing first with respect to the linear and scale terms. Using standard conjugate priors on (β, τ^2), the exact marginal posterior density on nonlinear parameter α is easily computed analytically. Then Laplace's approximation need only be applied for marginalizing in α. Previous use of Laplace's method has not taken advantage of partial linearity to allow exact maginalization in the parameters (β, τ^2). The obvious benefits are simplicity as well as the greater expected numerical accuracy provided when Laplace's approximation is used in lower dimensional integration. Partial linearity has been noted in Ripley (1996) but has only been used for parameter estimation in frequentist models by Shepanski (1987) and Hrycej (1992).

The marginal posterior for α can have many local maxima. Therefore, when marginalizing in α, a sum of Laplace approximations at the various local maxima is required. This procedure was proposed for marginalizing over α and β in Buntine and Weigand (1991) and further discussed in Mackay (1992) and Ripley (1994*a*).

The methodology is illustrated on a small nonlinear regression data set from Bates and Watts (1988,§3.13). The form of the regression is not specified by the underlying theory. A second larger financial time series data set from Lee, White, & Granger (1993) is also modelled using a lag one nonlinear autoregressive model.

1.2 An Alternate Invariant Sigmoidal

There are several troubling aspects of these NN models related to our Bayesian computations. First, the parameter (α, β) lacks identifiability. To see this, consider a NN model with one independent variable, no linear term, and two sigmoidal terms. The expected value of y is given by

$$E[y] = \beta_0 + \beta_1 \sigma(\alpha_{10} + \alpha_{11} x) + \beta_2 \sigma(\alpha_{20} + \alpha_{21} x). \tag{1.2}$$

This expectation is unchanged if one interchanges the parameter sets $(\beta_1, \alpha_{10}, \alpha_{11})$ and $(\beta_2, \alpha_{20}, \alpha_{21})$. Identifiability however is not an issue in prediction or Bayes regression estimation where the parameter (α, β) is integrated out. We shall return to this issue later as it concerns marginal inference for the components of α.

Secondly, certain sets of parameter values are not equivalent but which should be. Consider a change in the sign of $(\beta_1, \alpha_{10}, \alpha_{11})$ while also adding β_1 to β_0. This parameter change leaves the expectation in (1.2) unchanged, or

$$\beta_0 + \beta_1 - \beta_1 \sigma(-\alpha_{10} - \alpha_{11} x) = \beta_0 + \beta_1 \sigma(\alpha_{10} + \alpha_{11} x).$$

These two sets of parameter values are said to be "expectation equivalent". These values receive equal weight from the data through the likelihood function but may not have equivalent posterior weight because they may receive different prior weights. This lack of "posterior equivalence" is troublesome because in effect one arbitrary parametrization is assigned a higher posterior probability than another.

This posterior inequivalence is corrected by working instead with the translated sigmoidal,

$$\sigma^*(x) = \sigma(x) - \frac{1}{2} = \frac{1}{2} \tanh\left(\frac{x}{2}\right)$$

which is an odd function about 0. The choice of prior for (α, β) also bears upon this issue and in this regard we suggest an exchangeable multivariate T prior with common

mean 0 in the next subsection. Such a prior is an even function in each component and guarantees prior equivalence for parameter values that are expectation equivalent; thus posterior and expectation equivalence are one in the same. This corrects the conceptual difficulty above.

For ease of discussion, we shall refer to σ^* as a sigmoidal. Strictly speaking it is a translated sigmoidal since sigmoidals must be distribution functions.

The use of σ^* together with our choice of prior greatly reduce the number of local maxima that must be averaged when using Laplace's approximation to marginalize in α. This occurs because they lead to a marginal posterior in α that is exchangeable and which therefore generates equivalence classes of local maxima over certain permutations and sign changes for α. This phenomenon has previously been pointed out by Bishop (1995) when using the tanh sigmoidal.

The use of sigmoidal σ^* also necessitates that we translate the y- and x-component values by subtracting $1/2$ so they now fall in the range $[-1/2, 1/2]$. The first row entries of $\{x_i\}$, which are all 1, are also replaced with the value $1/2$.

1.3 Choice of Prior

To motivate the choice of prior, we must first understand the nature of the sigmoidal data transformations. Consider the graph of $\sigma^*(ax)$ versus (a, x) for $-15 \leq a \leq 15$ and $-1/2 \leq x \leq 1/2$ as given in Figure 1 below. Values of a centered about 0 result in a functions which are flat or gradual whereas larger values of $|a|$ yield functions which are steep. As $|a| \to \infty$ the sigmoidal approaches a unit step function and is said to approach "saturation". More generally, $\beta\sigma^*(\alpha_0 + \alpha_1 x)$ will avoid saturation and be gradual and smooth if the values of α_0, α_1, and β fall within the range $(-6, 6)$. A summation of such terms yields a regression function which is also smooth and which tends not to overfit the data.

In this regard, we assume a marginal prior structure for (α, β) that is an exchangeable multivariate T and which gives each component the common mean 0 and variance 4. The support of each component is concentrated in the range $(-6, 6)$ so that prior weight is given to smooth regression functions. As previously mentioned, this prior is an even function in each component and assures that expectation and posterior equivalence are the same. Such a marginal prior is specified hierarchically using a conditional prior for (α, β) given τ^2 that is $N\left(0, (\tau^2/\lambda)\, I_{p+q(m+1)}\right)$ with the fixed value of $\lambda = 10^{-3}/4$ explained below. A conjugate prior is used for τ^2 which is the inverse gamma distribution $\Gamma^{-1}(\nu/2, \gamma/2)$ (see Lee, 1989, §A.5) with density

$$\pi\left(\tau^2\right) \propto \left(\tau^2\right)^{-\nu/2-1} \exp\left\{-\frac{\gamma}{2\tau^2}\right\} \tag{1.3}$$

and parameters $\nu = 3$ and $\gamma = 4\lambda$. The parameter choices represent vague prior knowledge in the following sense. The first parameter measures the prior information for τ^2 and $\nu = 3$ is its smallest integer value for which (1.3) has finite mean. The second parameter assures that $E[\tau^2] = 4\lambda$ so that the marginal prior mean for components of (α, β) is 0 with variance 4. These parameter choices assure a prior in which

$$\tau^2/\lambda \sim \Gamma^{-1}(3/2, 2)$$

for any choice of $\lambda > 0$.

Our choice of $\lambda = 10^{-3}/4$ has been motivated by the desire to seek prior choices using sigmoidal σ^* that are compatible with the penalty choices of Ripley (1996) when working with sigmoidal σ. Using $\lambda = 10^{-3}/4$, then our choice of prior combines with the normal likelihood to produce a posterior that depends on the penalized least squares error defined as

$$SSE^*(\alpha, \beta) = \|y - X_\alpha \beta\|^2 + \lambda \|(\alpha, \beta)\|^2 . \tag{1.4}$$

The latter penalty term in (1.4) is called a "weight decay". Expression (1.4) may be minimized over (α, β) by using Ripley's "nnet" routine. For x- and y-values in $[0, 1]$, and a moderate number of parameters, Ripley (1996) has suggested the penalty weight $\lambda \in (10^{-4}, 10^{-2})$. Our selected value $\lambda = 10^{-3}/4$ is consistent with Ripley's choice of $\lambda = 10^{-3}$ for the following reason. Our x- and y-values fall in the range $[-1/2, 1/2]$ instead of $[0, 1]$ with half the magnitude: this allows us to double the value of (α, β) but this must also be compensated for in the weight decay factor $\|(\alpha, \beta)\|^2$ by using a quarter of $\lambda \in (10^{-4}, 10^{-2})$.

To summarize, the choice of prior has been motivated in several ways. An exchangeable multivariate T prior on (α, β) with mean 0 and variance 4 that incorporates the penalty weight $\lambda = 10^{-3}/4$ makes saturation of sigmoidals difficult and leads to a smooth well-behaved marginal posterior surface in α with a decreased number of local maxima and few false maxima (saddlepoints). This smooth surface provides improved performance for quasi-Newton maximization routines when implementing Laplace's method in α. From a Bayesian perspective, the prior choice expresses an a priori preference for smooth regression functions, perhaps as an expression of Occam's razor for simple smooth relationships.

1.4 Posterior Calculation

The priors on (α, β, τ^2) combine with normal likelihood to yield posterior

$$\pi(\alpha, \beta, \tau^2 | y) \propto (\tau^2)^{-(n+p+q(m+1)+\nu)/2-1} \exp\left\{-\frac{SSE^*(\alpha, \beta) + \gamma}{2\tau^2}\right\}.$$

This expression admits explicit marginalization in β followed by τ^2. Completing the square in β and integrating out yields

$$\pi(\alpha, \tau^2 | y) \propto (\tau^2)^{-(n+q(m+1)+\nu)/2-1} \left|X_\alpha^T X_\alpha + \lambda I_p\right|^{-1/2} \exp\left\{-\frac{E(\alpha)}{2\tau^2}\right\}$$

with

$$E(\alpha) = s_\alpha^2 + B_\alpha + \gamma + \lambda \|\alpha\|^2 . \tag{1.5}$$

In (1.5), s_α^2 is the error sum of squares of linear regression treating α as fixed, or

$$s_\alpha^2 = y^T \left(I - X_\alpha X_\alpha^+\right) y$$

with X_α^+ as the Moore-Penrose inverse of X_α, and

$$B_\alpha = \hat{\beta}_\alpha^T X_\alpha^T \left[I_n - X_\alpha \left(X_\alpha^T X_\alpha + \lambda I_p\right)^{-1} X_\alpha^T\right] X_\alpha \hat{\beta}_\alpha \tag{1.6}$$

with $\hat{\beta}_\alpha$ as the least squares estimate of β with α fixed or $\hat{\beta}_\alpha = X_\alpha^+ y$. Marginalization in τ^2 yields

$$\pi(\alpha|y) \propto \bar{\pi}(\alpha|y) = c \left|X_\alpha^T X_\alpha + \lambda I_p\right|^{-1/2} E(\alpha)^{-(n+q(m+1)+\nu)/2} \tag{1.7}$$

where π is the true posterior, $\bar{\pi}$ is the unnormalized posterior from exact marginalization in β and τ^2, and

$$c = \Gamma\left(\tfrac{n+q+\nu}{2}\right) \lambda^{(p+q)/2} \gamma^{\nu/2} / \Gamma\left(\tfrac{\nu}{2}\right) \pi^{(n+q)/2}.$$

For large n, the term B_α is close to 0 and negligible so that it may be left out of posterior computation. This occurs with informative designs in which the likelihood contribution $X_\alpha^T X_\alpha$ "washes out" prior term λI_p in the centre matrix of (1.6). In this setting the centre matrix approximates the residual projection matrix that is orthogonal to the columns of X_α so that $B_\alpha \approx 0$. The expansion

$$\begin{aligned}
\left(X_\alpha^T X_\alpha + \lambda I_p\right)^{-1} &= \left(X_\alpha^T X_\alpha\right)^{-1/2} \left[I_n + \sum_{k=1}^{\infty} \left\{-\lambda \left(X_\alpha^T X_\alpha\right)^{-1}\right\}^k\right] \left(X_\alpha^T X_\alpha\right)^{-1/2} \\
&= \left(X_\alpha^T X_\alpha\right)^{-1} - \lambda \left(X_\alpha^T X_\alpha\right)^{-2} + \lambda^2 \left(X_\alpha^T X_\alpha\right)^{-3} + \cdots
\end{aligned}$$

when substituted into (1.6) gives

$$\begin{aligned}
B_\alpha &= 0 - \lambda \hat{\beta}_\alpha^T \hat{\beta}_\alpha + \lambda^2 \hat{\beta}_\alpha^T \left(X_\alpha^T X_\alpha\right)^{-1} \hat{\beta}_\alpha + \cdots \\
&= O(\lambda/n) + O\left((\lambda/n)^2\right) + \cdots .
\end{aligned}$$

Further simplification occurs with large n. The factor $\left|X_\alpha^T X_\alpha + \lambda I_{p+1}\right|$ is dominated by the $E(\alpha)$ term whose important components now include

$$\bar{\pi}(\alpha|\, y) \simeq c\left(s_\alpha^2 + \lambda \|\alpha\|^2\right)^{-(n+q(m+1)+\nu)/2}.$$

More generally, for any nonlinear normal regression model with a partially linear component, the marginal posterior on nonlinearity parameter α is given in (1.7) when computed with conjugate priors. Letting $\lambda \to 0$ and setting $\nu = 0 = \gamma$ corresponds to using Jeffreys' priors for which the posterior on α is

$$\bar{\pi}(\alpha|\, y) \simeq c\left|X_\alpha^T X_\alpha\right|^{-1/2} |s_\alpha|^{-(n+q(m+1))}.$$

Modal estimates of α essentially minimize the empirical error s_α^2 for large n.

1.5 Posterior Symmetry in α

The posterior on α is invariant under two groups of transformations. First, because the sigmoidals are not identified, it is invariant to the $q!$ permutations of the q parameter sets within common sigmoidals. Secondly, because of its symmetry about 0, it is invariant to the 2^q possible sign changes applied to these same parameter sets. Together the posterior is invariant under a single group of $2^q q!$ transformations. The symmetry of the marginal posterior on α can be easily seen by expressing it as the integral of likelihood $\times$ prior and noting that this integral is symmetric because of the choice of symmetric priors and the oddness property for σ^*.

The group under which $\pi(\alpha|\, y)$ is invariant forms a subgroup of the orthogonal group of transformations on α. This subgroup is the largest such subgroup which has relevance to the NN model. To show this, consider the nonlinear part of the design

matrix

$$\begin{bmatrix} \sigma\left(x_1^T\alpha_1\right) & \cdots & \sigma\left(x_1^T\alpha_q\right) \\ \sigma\left(x_2^T\alpha_1\right) & \cdots & \sigma\left(x_2^T\alpha_q\right) \\ \vdots & \ddots & \vdots \\ \sigma\left(x_n^T\alpha_1\right) & \cdots & \sigma\left(x_n^T\alpha_q\right) \end{bmatrix} := \sigma\left(\begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} \begin{bmatrix} \alpha_1 & \alpha_2 & \ldots & \alpha_q \end{bmatrix}\right).$$

The group maps $\alpha \rightarrow O\alpha$ and an orthogonal matrix O with more than one non-zero entry in the ith row will result in a transformed parameter matrix $O\alpha$ whose (i,j)th term is a linear combination of $\alpha_{1,j-1}, \alpha_{2,j-1}, \ldots, \alpha_{q,j-1}$ which is not permissible in the NN model. Matrix O must therefore have only one non-zero entry in each row which necessarily must be either "1" or " $-$ 1" if it is to be orthogonal: thus O is a member of the subgroup of size $2^q q!$.

The group invariance leads to equivalence classes of parameter values sharing common orbits under the invariant group of size $2^q q!$. Members of the same orbit are expectation equivalent parameter values. In each orbit, one member has location parameters for the q sigmoidals that satisfy

$$0 \leq \alpha_{10} \leq \ldots \leq \alpha_{q0} \tag{1.8}$$

and which can be used to identify the orbit. Starting with arbitrary matrix parameter α, it is transformed into its identified form by ordering the vectors according to their first components, $\{\alpha_i \text{sgn}\left(\alpha_{i0}\right) : i = 1, \ldots, q\}$, and letting these form the columns of the matrix α_{MI}^T. This identified version is the "maximal invariant" parameter under the invariant group. We shall denote the set of maximal invariant parameter values as $O = \{\alpha_{MI}\} \subset \Re^{q(m+1)}$. For parameter $\alpha_{MI} \in O$, the sigmoidals essentially end up numbered from left to right according to how the sigmoidals are centered in the regression.

1.6 Approximate Posterior Expectation in α

1.6.1 Laplace's Approximation with Multiple Modes

Laplace's method was introduced for Bayesian marginalization by Leonard (1982), Davison (1986), and Tierney and Kadane (1986). More recently it has found application as a means for computing approximate Bayes factors as discussed in Berger and Pericchi (1996), Kass and Raferty (1995), and DiCiccio et al. (1997). All of these authors, however, consider the unimodal situation and not the multimodal setting commonly arising with NN models. This latter setting has been discussed by Buntine and Weigand (1991), Mackay (1992), Bishop (1995), and Ripley (1994*a*). Multimodal marginalization has also been considered with MCMC by Neal (1993, 1995)

The posterior expectation of an arbitrary smooth positively-valued function $g(\alpha)$ is computed using Laplace's approximation adapted to deal with the multiple local maxima found in this context. Our context requires maginalization in α and the approximation is

$$\int g(\alpha)\,\bar{\pi}(\alpha|y)\,d\alpha \simeq \sum_{\hat{\alpha}} g(\hat{\alpha})\,L(\hat{\alpha}) \tag{1.9}$$

where

$$L(\hat{\alpha}) = \frac{c\,(E(\hat{\alpha}))^{-(n+q(m+1)+\nu)/2}}{\sqrt{|X_{\hat{\alpha}}^T X_{\hat{\alpha}} + \lambda I_p| \times \|H(\hat{\alpha})\|}}.$$

The values $\{\hat{\alpha}\}$ comprise the set of local maxima for the dominant portion of $\ln \bar{\pi}(\alpha|y)$ taken to be

$$-\tfrac{1}{2}(n + q(m+1) + \nu) \ln E(\alpha). \tag{1.10}$$

Matrix $H(\alpha)$ is the Hessian of (1.10). Determination of the collection of local maxima is the most computationally intensive part of this computation. For this we used Gauss-Newton searches for critical values starting the iterations at randomly selected values of α.

When $g(\cdot)$ is invariant under the permutation group of Section 5, the summation in (1.9) may be restricted to summing over only the identified local maxima parameters in O when the multiplicity factor $2^q q!$ is also used. Then (1.9) is

$$\int g(\alpha)\,\bar{\pi}(\alpha|y)\,d\alpha \simeq \sum_{\hat{\alpha}} g(\hat{\alpha})\,L(\hat{\alpha}) = 2^q q! \sum_{\hat{\alpha}\in O} g(\hat{\alpha})\,L(\hat{\alpha})\,. \tag{1.11}$$

Thus the search for local maxima $\hat{\alpha}$ is accordingly restricted to starting Gauss-Newton iteration within O.

1.6.2 Further Approximation Issues

Critical values of $\ln E(\cdot)$ require the computation of $\partial \ln E(\alpha)/\partial\alpha$ which is straightforward except for the dependence of $E(\cdot)$ on X_α^+. Golub and Pereyra (1973) determine the (i,j)th component of this derivative as

$$\begin{aligned}\partial X_\alpha^+/\partial\alpha_{ij} &= -X_\alpha^+\left(\partial X_\alpha/\partial\alpha_{ij}\right)X_\alpha^+ + X_\alpha^+\left(X_\alpha^+\right)^T\left(\partial X_\alpha/\partial\alpha_{ij}\right)^T\left(I - X_\alpha X_\alpha^+\right)\\ &+ \left(I - X_\alpha^+ X_\alpha\right)\left(\partial X_\alpha/\partial\alpha_{ij}\right)^T\left(X_\alpha^+\right)^T X_\alpha^+\end{aligned}$$

when $X_\alpha^T X_\alpha$ has locally constant rank which might not be its full rank. From this we are able to derive an exact expression for the gradient of $\ln E(\alpha)$. For Hessian computation, sufficient accuracy was obtained by using finite differences of the gradient.

Special care must be taken when working with the numerical Moore-Penrose inverse of X_α. This requires singular value decomposition and also a determination about which of the small singular values should be considered negligible and taken as 0. We used the IMSL routine DLSGRR and found that a tolerance for negligibility set at 10^{-8}, approximately the square root of machine precision, worked well. This setting was required due to the amount of relative error caused by the use of other numerical routines.

The stable estimate $\hat{\alpha}_s$ is defined as the dominant Bayesian mode with $\lambda = 0$ and

is nearly the dominant Bayesian mode, $\hat{\alpha}$ with $\lambda = 10^{-3}/4$ since

$$\hat{\alpha}_s = \hat{\alpha} + O\left(n^{-1}\right).$$

Despite this, it is still preferable to use a weight decay term for large n. The weight decay term makes the maximization of $-\ln E(\alpha)$ easier, gives fewer maxima, results in Hessians that are usually nonsingular and yields a surface which is more quadratic in appearance. This last point is particularly relevant for achieving accuracy with Laplace's approximation.

1.7 Marginal Inference

1.7.1 Model Choice

The NN model with q sigmoidal terms is denoted M_q and has Bayes factor

$$\Pr\{M_q|y\} \propto \pi(y|M_q)\,\pi(M_q).$$

where

$$\pi(y|M_q) = \int \bar{\pi}(\alpha|\, y, M_q)\, d\alpha \tag{1.12}$$

is approximated using Laplace's method with $g(\alpha) \equiv 1$. Variable selection from among the m independent variables in $\{x_i\}$ could also be implemented with or without sigmoidal selection but this has not been attempted.

In this context, the "best" model is not always that which maximizes $\Pr\{M_q|y\}$ in q. Suppose, for example, the true model has a single sigmoidal so $q = 1$. The NN models are nested with $M_0 \subset M_1 \subset M_2 \subset \cdots$ so that, for example, a model with 2 sigmoidals is a 1 sigmoidal model if one of the sigmoidals has all its non-intercept α-values set to zero. Thus there is no unique correct choice for q when M_1 is the true parsimonious model; the choice of any value of $q \in \{1, 2, \ldots\}$ is as good as any

other. We shall see the occurrence of this phenomenon in the examples where the Bayes factors for 2 and 3 sigmoidals are greater than that for 1 but the parsimonious choice of model is clearly a single sigmoidal. The Bayes factors in combination with their fitted parameter values need to be examined in order to make a parsimonious model choice for value q.

1.7.2 Predictive Distributions

The posterior density for future observable Y_f at y_f, given its associated variable $x_f = (1, x_{f1}, \ldots, x_{fm})$ and model M_q is easily computed by including y_f and x_f into the data set and marginalizing as in model choice. Then

$$\pi(y_f|x_f, y, M_q) = \frac{\pi(y, y_f|M_q)}{\pi(y|M_q)} = \frac{\int \bar{\pi}(\alpha|y, y_f, M_q)\, d\alpha}{\int \bar{\pi}(\alpha|y, M_q)\, d\alpha} \tag{1.13}$$

where both Laplace approximations take $g(\alpha) \equiv 1$. The denominator is the computation from model choice and the numerator is the same computation but including the prospective future observable y_f in the data. Further marginalization over the models $\{M_q\}$ with a priori weighting $\pi(M_q) \propto 1$ is determined as

$$\pi(y_f|x_f, y) = \frac{\sum_q \pi(y_f|x_f, y, M_q)\, \pi(y|M_q)}{\sum_q \pi(y|M_q)} = \frac{\sum_q \pi(y, y_f|M_q)}{\sum_q \pi(y|M_q)} \tag{1.14}$$

and may be based entirely upon Laplace approximations as used to compute (1.12) and (1.13). When a clear choice for a parsimonious model exists, as occurs in our examples, then the mixing of models in (1.14) is not much different from the use of (1.13) based upon the parsimonious choice. In such instances, use of the parsimonious model can greatly simplify the computations.

1.7.3 Bayes Estimation of the Regression Function

Taking $\theta = (\alpha, \beta, \tau^2)$, then the posterior expectation of the regression function is

$$\begin{aligned} E_y^{\theta}\left[X_\alpha \beta | \, y, M_q\right] &= \int X_\alpha \beta \, \pi\left(\theta | y, M_q\right) d\theta \\ &= \int X_\alpha \Sigma_1 X_\alpha^T X_\alpha \hat{\beta}_\alpha \pi\left(\alpha | \, y, M_q\right) d\alpha \end{aligned} \tag{1.15}$$

since

$$\beta | \alpha, \tau^2, y \sim N\left(\Sigma_1 X_\alpha^T X_\alpha \hat{\beta}_\alpha, \tau^2 \Sigma_1\right)$$

with $\Sigma_1 = \left(X_\alpha^T X_\alpha + \lambda I_p\right)^{-1}$. In terms of $\bar{\pi}$, (1.15) is

$$E_y^{\theta}\left[X_\alpha \beta | \, y, M_q\right] = \frac{\int X_\alpha \Sigma_1 X_\alpha^T X_\alpha \hat{\beta}_\alpha \bar{\pi}\left(\alpha | \, y, M_q\right) d\alpha}{\int \bar{\pi}\left(\alpha | \, y, M_q\right) d\alpha}. \tag{1.16}$$

Laplace's approximation in the numerator of (1.16) takes

$$g\left(\alpha\right) = X_\alpha \Sigma_1 X_\alpha^T X_\alpha \hat{\beta}_\alpha.$$

Since the regression function is invariant under the permutation group, Laplace's approximation as in (1.11) may be used. This invariance exists because expectation and posterior equivalence are the same in our framework. The denominator is $\pi\left(y | M_q\right)$ in (1.12) and its Laplace approximation also determines the Bayes factor for model M_q. Further marginalization over the models $\{M_q\}$ is possible and leads to a Bayes estimator that is a mixture of the Bayes estimates in (1.16) of the form (1.14). This is not particularly beneficial, however, when there is a clear cut choice for a parsimonious model.

1.7.4 Marginal Distributions

Inference about marginal densities in α is only meaningful once the various sigmoidals have been identified as described in Section 5. Consider such an identified NN model

in which $\alpha \in O$ has been partitioned into $\alpha_{(1)}$ and $\alpha_{(2)}$ where $\alpha_{(1)}$ is one-dimensional and $\alpha_{(2)}$ contains all the other components. The marginal posterior of $\alpha_{(1)}$ is

$$\pi\left(\alpha_{(1)}|y, M_q\right) = \frac{\int_{\mathcal{O}_1} \bar{\pi}\left(\alpha_{(1)}, \alpha_{(2)}|\, y, M_q\right) d\alpha_{(2)}}{\int_{\mathcal{O}} \bar{\pi}\left(\alpha|\, y, M_q\right) d\alpha}. \tag{1.17}$$

where $O_1 = \left\{\alpha_{(2)} : \left(\alpha_{(1)}, \alpha_{(2)}\right) \in \mathcal{O}\right\}$. A Laplace approximation to the denominator is the Bayes factor computation above and is summed over the identified members in O. An approximation to the numerator holds $\alpha_{(1)}$ fixed and applies Laplace's method in parameter $\alpha_{(2)}$ with $g = 1$. Only identified values of $\left(\alpha_{(1)}, \hat{\alpha}_{(2)}\right) \in O$ are used where $\hat{\alpha}_{(2)}$ is an identified maximum determined by holding $\alpha_{(1)}$ fixed.

1.8 Examples

1.8.1 Nonlinear Regression

Our first data set was analyzed in Bates and Watts (1988) and concerns the utilization of nitrate (y-variable) in bush beans as a function of light intensity (x-variable). The primary leaves of three 16-day-old bean plants were subjected to eight levels $(2.2, 5.5, 9.6, 17.5, 27.0, 46.0, 94.0, 170.0)$ of light intensity (μE/m^{2}s) and the nitrate utilizations (nmol/g hr) were measured. The experiment was performed on two different days. Even though no theoretical model has been proposed for this data, it has been hypothesized, by the researchers, that utilization should be zero for zero light intensity and should also approach an asymptote as light intensity increases. Incremental parameter effects for the two days were included in the analysis of Bates and Watts, for models having an asymptote, but were not found to be significant. Therefore, we combine the data from both days and fit a common model to the $n = 48$ cases of data.

We consider NN models with up to three sigmoidal terms using $\lambda = 10^{-3}/4$. The factors $\{\pi(y|M_q)\}$ which determine the posterior weights of the various models are

approximated using Laplace's method and displayed in Table 1. For each M_q, the local maxima were determined by fitting the model 100 times starting the Gauss-Newton iterations at randomly selected points in $\alpha \in [-9, 9]^{q(m+1)} \cap O$. The various maxima are given in maximal invariant form $\hat{\alpha}_{MI}$ along with the associated contributions to the Laplace approximation from the orbits of maxima. Starred entries are used in place of parameter estimates whose values are less than 10^{-12} in magnitude.

The Bayes factor totals favour 1, 2, and 3 nodes but these alone do not fully explain the regression relationship. An examination of the modal parameter estimates from Laplace's method is more informative about parsimonious model choice. There are two maxima when fitting the single node models in M_1. When fitting two node models, these same two nodes are essentially the two dominant maxima as may be seen from their weights in the first sigmoidal which are all less than 10^{-12} in magnitude. The percentage of the Bayes factor contributed by these two dominant maxima is 96.0%. Thus the fit of two node models also suggests that one sigmoidal is a parsimonious fit. The fit of three node models explains nothing that hasn't already been explained with two node models, since all three maxima are essentially those from the two node models. The two dominant maxima that represent a single node model combine for 97% of the Bayes factor when fitting three nodes.

To assess the accuracy of Laplace's method in determining the Bayes factors of the models above, we evaluated these quantities using numerical integration. For 1,2, and 3 node models these factors were 1.733×10^{20}, 1.577×10^{21}, and 6.865×10^{22}. They agree closely with the totals given in Table 1.

Figure 2 shows a scatterplot of the regression data (circles) along with several estimates of the regression function: (i) the true Bayes estimate (solid) as in (1.16) assuming a one node model as determined from numerical integration; (ii) Laplace's approximation to this estimate (dashed); and (iii) Laplace's approximation to the

mixture of Bayes estimates (dot-dash) where the mixing is over models with up to three nodes and weights are proportional to Bayes factors. Numerical integration and Laplace approximation yield indistinguishable estimates. A comparison of the dashed and dot-dashed lines reveals that very little is lost by restricting attention to the parsimonious model with a single node.

Figure 3 shows a predictive density for future value Y_f given $x_f = 187$ which provides an extrapolation for Y_f to the right of the data in the regression plot. The true predictive density based upon a single node (solid) is determined with numerical integration and its normalized Laplace approximation (dashed) is reasonably accurate.

The researchers who gathered the data expected a "levelling off" of the regression as light intensity increases. This can be seen in all the regression fits in Figure 2. It also is reflected in the location of the predictive density plot in Figure 3; the approximate mean of 1.8×10^4 is consistent with the asymptotic level determined from Figure 2. By contrast, the frequentist model eventually proposed by Bates and Watts did not suggest this "leveling off" behavior.

For the parsimonious fit of a single node, Figure 4 shows a surface plot for the marginal bivariate posterior of nonlinear parameters α_{10} (intercept) and α_{11}(slope). Figure 5 is the marginal posterior on α_{11}(slope) as determined from numerical integration (solid) and Laplace's method (dashed) when normalized. The Laplace approximation is extremely accurate.

1.8.2 Nonlinear Time Series

Our next data set consists of monthly U.S. personal income data from January, 1959 to July, 1990 (inclusive) for a total of 379 observations and is shown in Figure 6. It is taken from Lee, White, & Granger (1993) where tests for neglected nonlinearity

are compared against their newly proposed NN test. These authors transformed the personal income data $\{z_t\}$ into a stationary sequence of seasonally adjusted $y_t = \ln(z_{t+1}/z_t)$ values. A time plot for $\{y_t\}$ is shown in Figure 7. The y_t-values were then fit to an $AR(k)$ model where k was found to be one using the SIC criterion. All the tests for the presence of nonlinearity, including the Keenan, Tsay, Ramsy RESET (1 & 2), White (1, 2 & 3), Mcleod and Li, BDS, Bispectrum tests and their newly proposed NN test, gave p-values below 0.016 which strongly suggests the presence of neglected nonlinearity. We modelled this time series as a non-linear autoregressive process via a NN model with q sigmoidal terms where our distributions are conditional upon the observed value of Y_1. Once again we take $\lambda = 10^{-3}/4$. Since the data consist of a large number of observations, we might expect the prior to be dominated by the likelihood and so its choice is not so critical.

Our parameter estimation results are given in Table 2. A careful examination of the table suggests a parsimonious fit with one sigmoidal. As seen with the previous example, the best fits with one, two and three sigmoidals have increasing Bayes factors. Their associated parameter estimates, however, support the parsimonious fit of a single sigmoidal.

We may assess the accuracy of Laplace's method for determining Bayes factors in the models above by using numerical integration. For 1, 2, and 3 nodes these factors are 3.756×10^{114}, 1.278×10^{115}, and 1.523×10^{116} which show reasonable agreement with the totals in Table 2.

Figure 8 shows a scatterplot of the regression data (circles) along with several estimates of the regression function: (i) the true Bayes estimate (solid) as in (1.16) assuming a one node model as determined from numerical integration; (ii) Laplace's approximation to this estimate (dashed); and (iii) Laplace's approximation to the mixture of Bayes estimates (dot-dash) where the mixing is over models with up to

three nodes and weights are proportional to Bayes factors. Numerical integration and Laplace approximation yield indistinguishable estimates. A comparison of the dashed and dot-dashed lines reveals that virtually nothing is lost in working with the NN model with a single node.

Figure 9 shows a predictive density for future value Y_{378} given $y_{377} = 5.779 \times 10^{-3}$. The true predictive density based upon a single node (solid) is determined with numerical integration and its normalized Laplace approximation (dashed) is once again quite accurate. The predictive mean of approximately 0.007 appears to be a reasonable extrapolation of the data in Figure 7.

For the parsimonious fit of a single node, Figure 10 shows a surface plot for the marginal bivariate posterior of nonlinear parameters α_{10} (intercept) and α_{11} (slope). Figure 11 is the marginal posterior on α_{11} as determined from numerical integration (solid) and Laplace's method (dashed) when normalized. The Laplace approximation is extremely accurate.

Table 1. Contributions to the Bayes factors from Laplace's approximation associated with the various orbits of local maxima designated in the first six columns. Stars indicate that entries are smaller than 10^{-12} in magnitude and $\emptyset$ indicates the entry is not applicable.

Local maxima $\text{vec}(\hat{\alpha}_{MI}) = (\hat{\alpha}_1^T, \hat{\alpha}_2^T, \hat{\alpha}_3^T)$						Laplace Approx.
$\hat{\alpha}_{10}$	$\hat{\alpha}_{11}$	$\hat{\alpha}_{20}$	$\hat{\alpha}_{21}$	$\hat{\alpha}_{30}$	$\hat{\alpha}_{31}$	Contribution to $\pi(y\|M_q)$
Zero nodes ($q = 0$)						
$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	6.8648
One node ($q = 1$)						
$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	2.8398	5.0477	1.2686×10^{20}
$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	.78097	−5.0598	3.9796×10^{19}
$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	*	*	1.3280×10^{1}
					Total	1.6666×10^{20}
Two nodes ($q = 2$)						
$\emptyset$	$\emptyset$	*	*	2.8398	5.0477	9.7384×10^{20}
$\emptyset$	$\emptyset$	*	*	.78097	−5.0598	5.0505×10^{20}
$\emptyset$	$\emptyset$	.38741	−4.413	1.6894	3.3967	6.1946×10^{19}
$\emptyset$	$\emptyset$	*	*	*	*	5.1221×10^{1}
					Total	1.5408×10^{21}
Three nodes ($q = 3$)						
*	*	*	*	2.8398	5.0477	1.0849×10^{22}
*	*	*	*	.78097	−5.0598	9.3010×10^{21}
*	*	.38741	−4.413	1.6894	3.3967	6.3451×10^{20}
*	*	*	*	*	*	2.8668×10^{2}
					Total	2.0785×10^{22}

Table 2. Contributions to the Bayes factors from Laplace's approximation associated with the various orbits of local maxima designated in the first six columns.

Local maxima $\text{vec}(\hat{\alpha}_{MI}) = (\hat{\alpha}_1^T, \hat{\alpha}_2^T, \hat{\alpha}_3^T)$						Laplace Approx.
$\hat{\alpha}_{10}$	$\hat{\alpha}_{11}$	$\hat{\alpha}_{20}$	$\hat{\alpha}_{21}$	$\hat{\alpha}_{30}$	$\hat{\alpha}_{31}$	Contribution to $\pi(y\|M_q)$
Zero nodes ($q = 0$)						
$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	3.0268×10^{108}
One node ($q = 1$)						
$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	.15528	4.6874	3.8287×10^{114}
$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	*	*	3.9796×10^{108}
					Total	3.8287×10^{114}
Two nodes ($q = 2$)						
$\emptyset$	$\emptyset$	*	*	.15528	4.6874	2.1133×10^{115}
$\emptyset$	$\emptyset$	2.2567	4.3657	3.0945	−5.7259	3.7388×10^{112}
$\emptyset$	$\emptyset$	*	*	*	*	2.3839×10^{109}
					Total	2.1170×10^{115}
Three nodes ($q = 3$)						
*	*	*	*	.15528	4.6874	1.7407×10^{116}
*	*	2.2567	4.3657	3.0945	−5.7259	2.8953×10^{113}
*	*	*	*	*	*	1.4093×10^{110}
					Total	1.7435×10^{116}

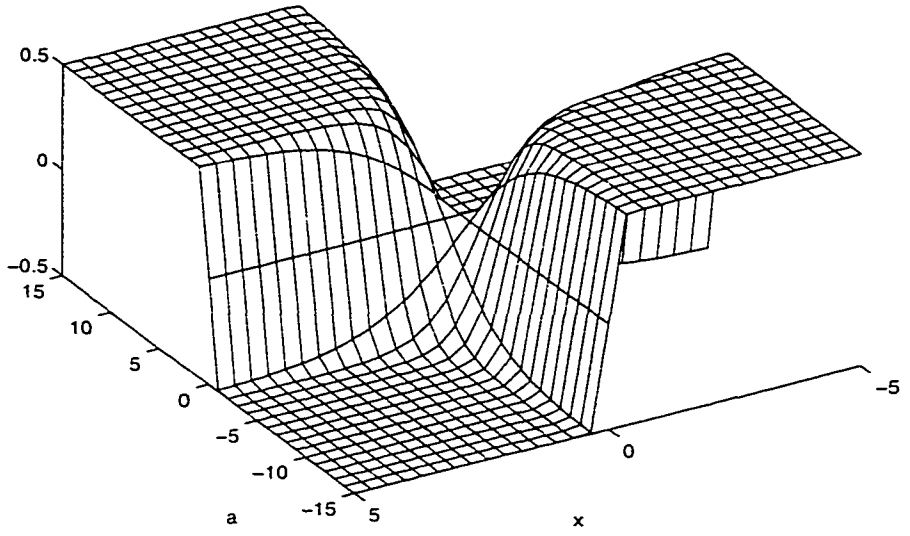

Fig. 1. Plot of $\sigma^*(\alpha x)$ versus (α, x).

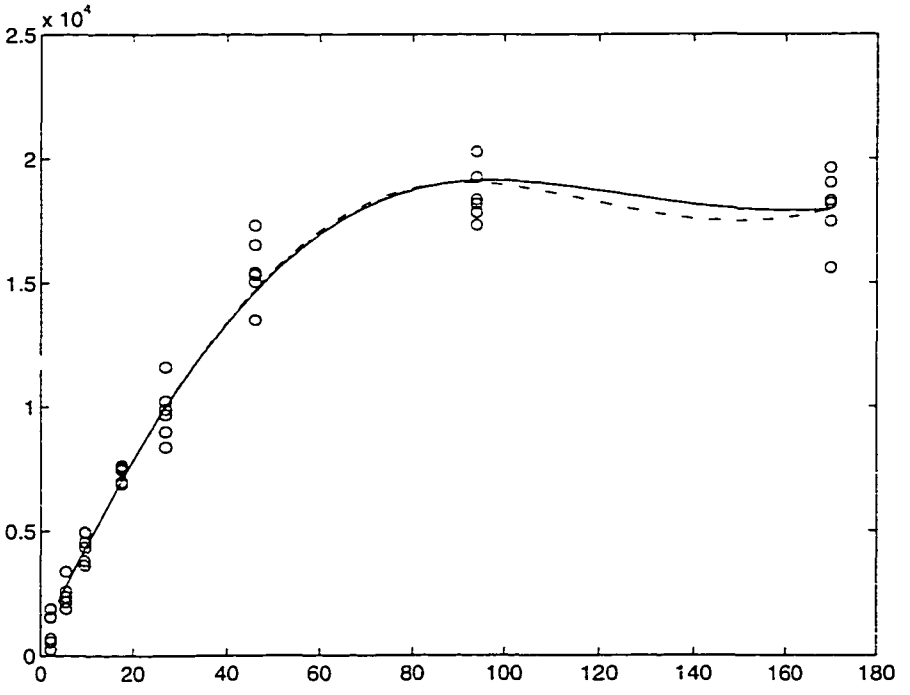

Fig. 2. Baye's estimates of the regression function: exact (solid) fitting a single node, its Laplace approximation (dashed), and a Laplace approximation to the estimate from mixing over models $M_0 - M_3$ (dot-dashed).

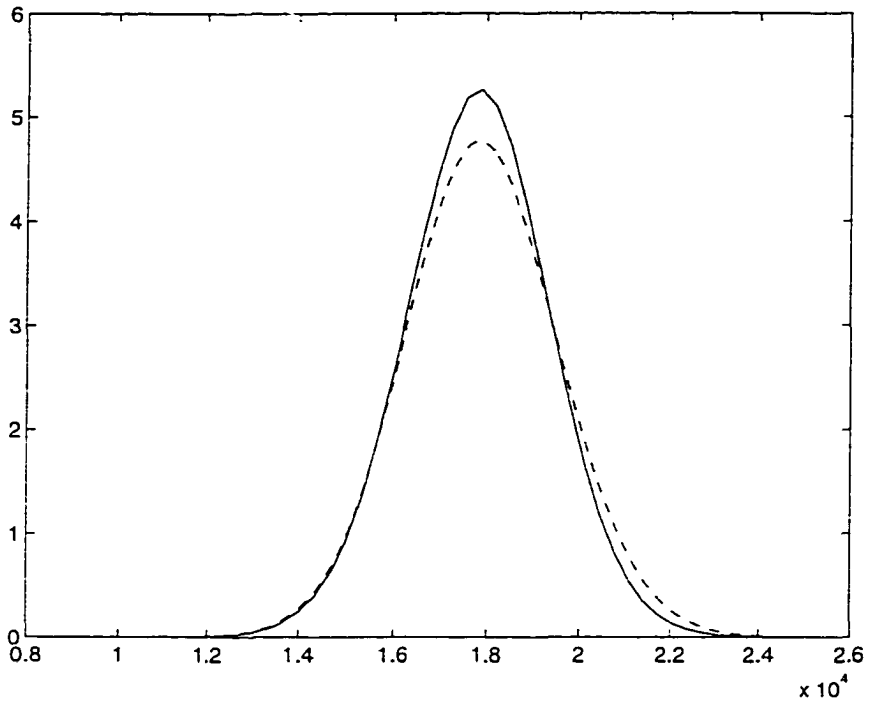

Fig. 3. True predicive density of Y_f for $x_f = 187$ (solid), assuming a single node, and its normalized Laplace approximation (dashed).

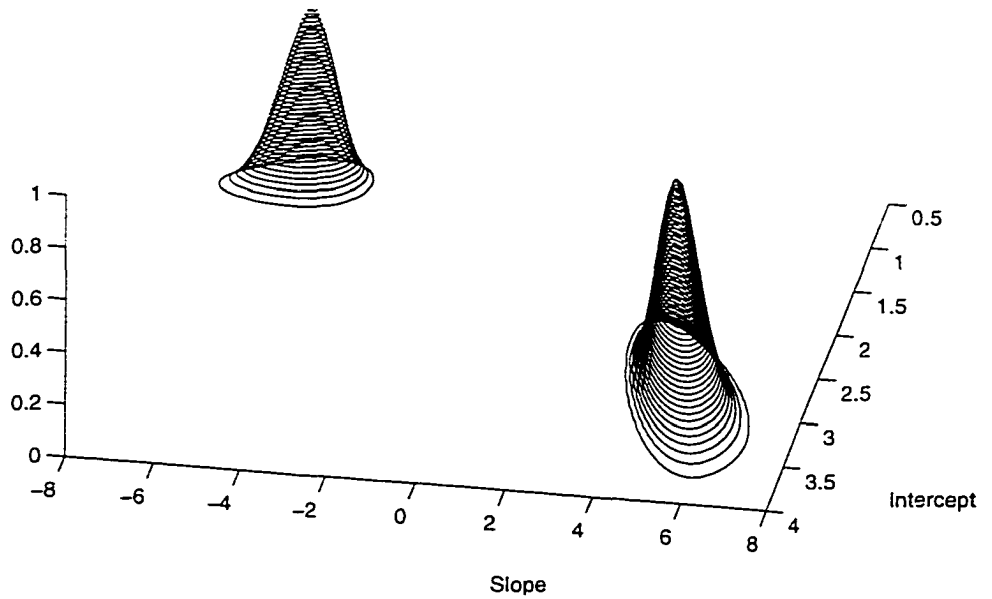

Fig. 4. Joint marginal posterior of the nonlinear parameters α_{10} (intercept) and α_{11}(slope) for a one node model.

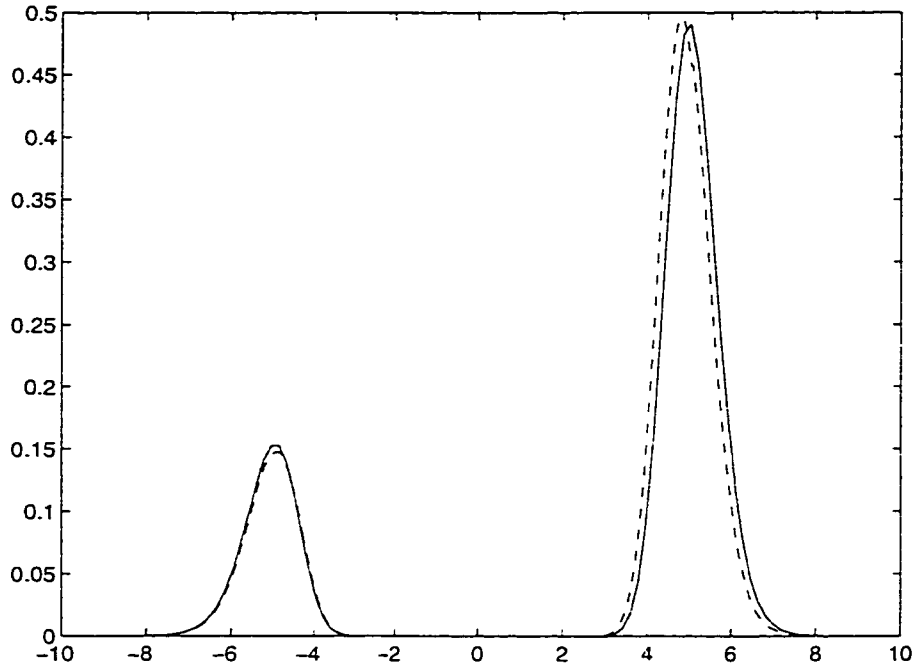

Fig. 5. The true marginal posterior on slope parameter α_{11} from Figure 4 (solid) and its Laplace's approximation (dashed).

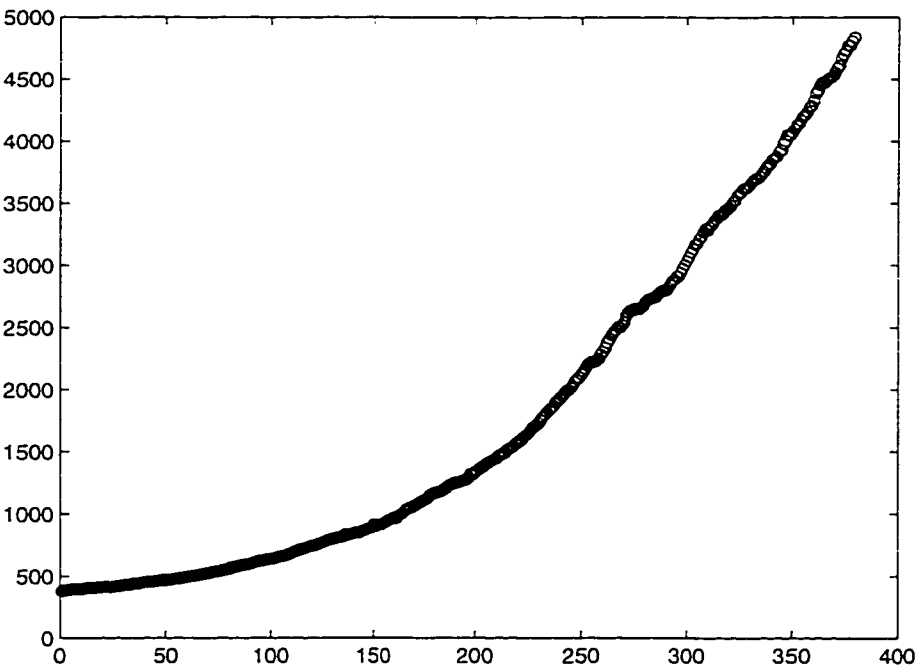

Fig. 6. Personal income data from Jan. 1959 to July 1990 (inclusive).

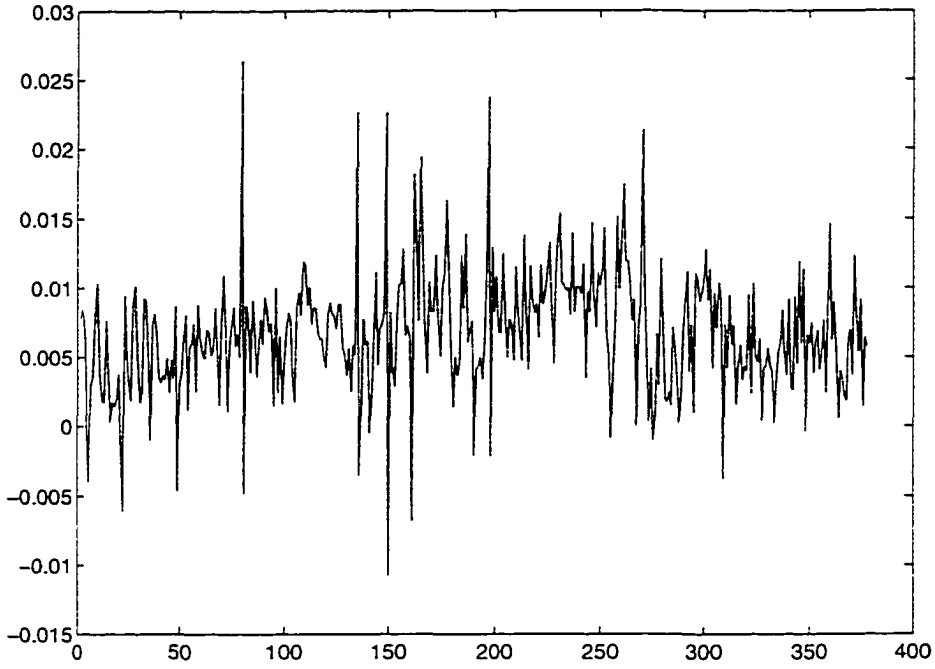

Fig. 7. The stationary time series.

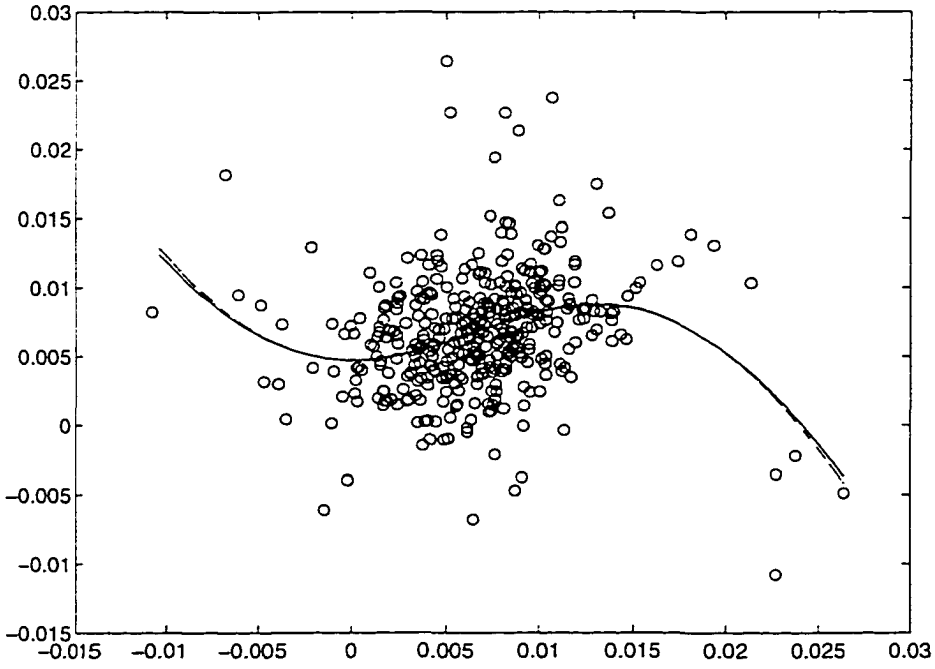

Fig. 8. Baye's estimates of the regression function: exact (solid) fitting a single node, its Laplace approximation (dashed), and a Laplace approximation to the estimate from mixing over models $M_0 - M_3$ (dot-dashed).

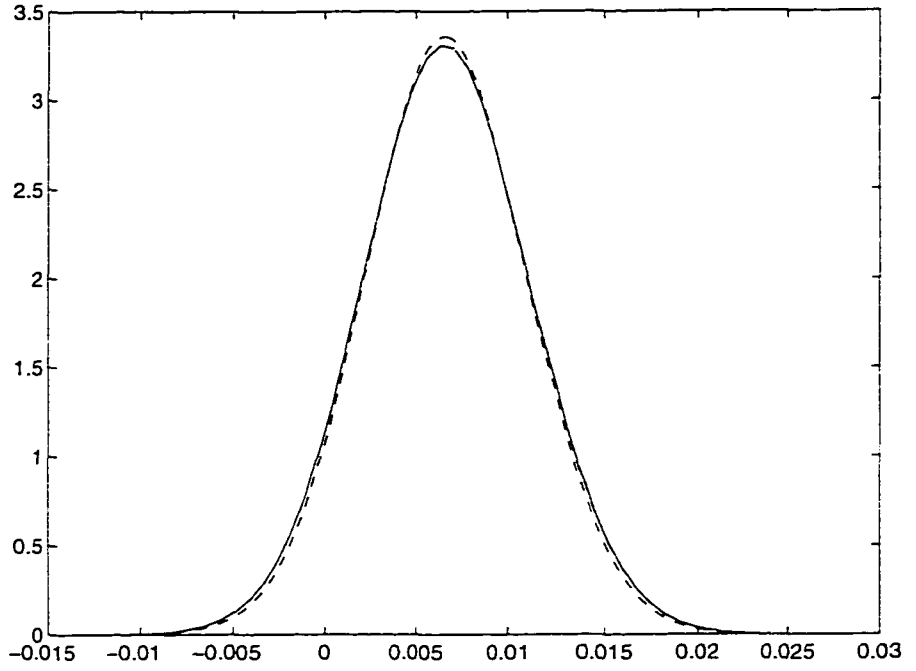

Fig. 9. True predicive density of Y_{378} for $y_{377} = 5.7794 \times 10^{-3}$ (solid), assuming a single node, and its renormalized Laplace approximation (dashed).

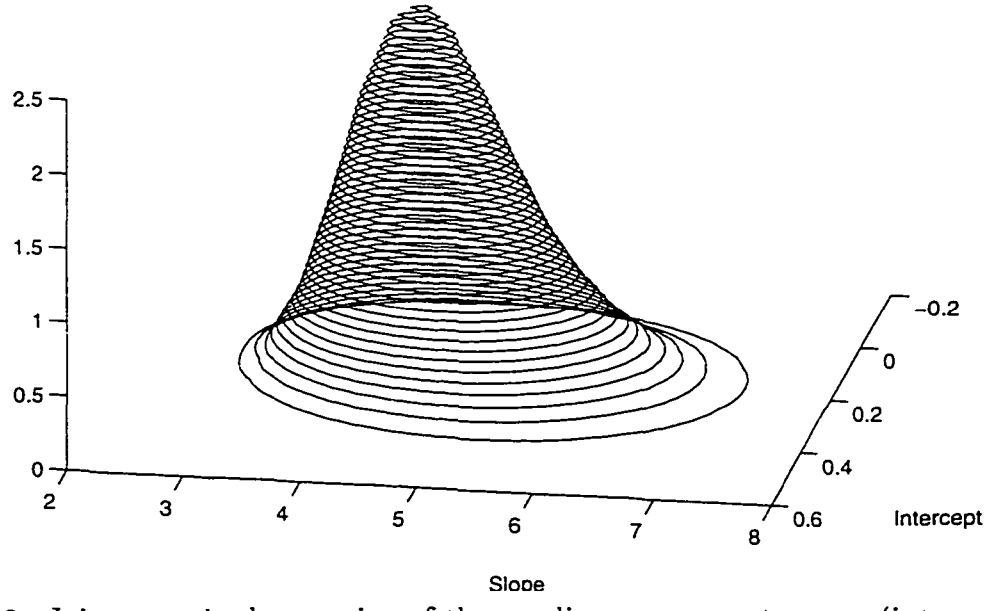

Fig. 10. Joint marginal posterior of the nonlinear parameters α_{10} (intercept) and α_{11}(slope) for a one node model.

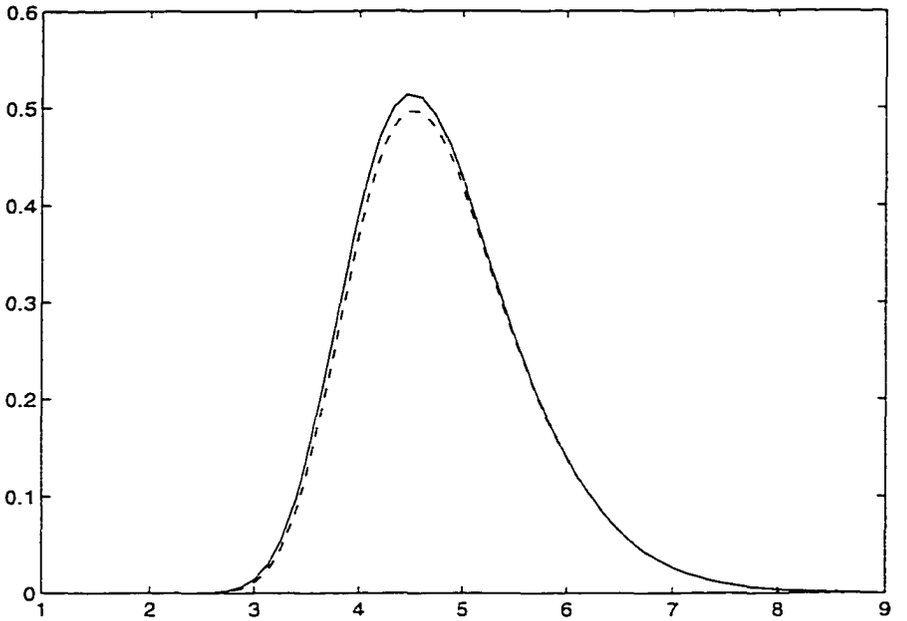

Fig. 11. The true marginal posterior on slope parameter α_{11} from Figure 10 (solid) and its Laplace's approximation (dashed).

2 Classification and Lumpability in the Stochastic Hopfield Model

2.1 Introduction

The stochastic Hopfield model (SHM) is a fundamental prototype of an artificial neural network. It can be used to associate the binary vector $\mathbf{u}_0$, of length n, with one of q binary exemplars. For this reason, it has been used to model the process of associative memory and more generally as a classification algorithm. It has a wide range of applicability and continues to motivate much research in the artificial neural network area.

The SHM and its variants have been applied to several combinatorial optimization problems such as graph bipartitioning, as well as the traveling salesman problem and the weighted matching problem. Furthermore, they have also been used as error-correcting algorithms in which $\mathbf{u}_0$ is modeled as an exemplar corrupted by noise. Along these lines, the SHM is a special case of a Gibbs sampler algorithm as described in Geman & Geman (1984). In addition, Whittle (1991) has pursed the connection between the SHM and hypothesis testing. The SHM has also found more biologically oriented applications in modeling bipolar disorder (Hoffman 1992) and the cyclic swimming pattern of the mollusk *Tritonia diomedea* (Kleinfield and Somopolinsky 1989). While applications for the SHM abound it remains, however, a rather difficult model to study.

This paper makes four contributions as outlined below. First, we simplify the complexity of this model using lumpability. The state space of the SHM has size 2^n and consists of the set of all binary vectors of length n. As a result of this size, many important calculations require $O(2^n)$ operations. State space size may be dramatically reduced by recognizing and proving the lumpability of the Markov chain underlying

this model's classification process in section 2.3.2. Lumping leads to a reduced or *lumped* SHM with a polynomial number of states in n, i.e., $O(n^p)$. When the lumped SHM is used to study classification in the full model, $O(n^p)$ operations are required. This simplifies many classical Markov chain analyses including the calculation of absorption probability and mean passage time.

Secondly, we provide saddlepoint approximations for the classification time distribution which are quick, accurate and easy to use. Access to saddlepoint approximations results from using Pyke's rule (1961) which provides the exact moment generating function (MGF) of the classification time distribution. Approximate inversion of this MGF is performed using with a saddlepoint approximation. Direct application of this method, without lumping, involves $O(2^n)$ operations; only $O(n^p)$ operations are required with lumping. The extreme accuracy attained by saddlepoint approximations is illustrated in section 2.4.3.

Aside from simulation studies, very little has been said about classification time distributions in the neural networks literature. One exception is the work of Kam and Cheng (1989) which provides upper and lower bounding curves on the cumulative distribution function (CDF) of time of classification to a single exemplar while artificially treating the other exemplars as non-absorbing states. Clearly, such calculations do not account for misclassification probabilities.

Thirdly, techniques are given for enumerating as well as approximating the stationary distribution of the energy function. This distribution is determined by the Markov chain underlying the SHM's classification process; therefore, direct enumeration of its probabilities requires $O(2^n)$ operations. Even though our work concerns the transient dynamics of this Markov chain, the energy function's stationary distribution is still of interest. The SHM was originally designed to ensure that energy function minimization would be equivalent to the classification of $\mathbf{u}_0$. As such, tail

probability calculations provide a rough measure of the additional reduction in the energy which is possible if the SHM's classification process were allowed to run for a very long time.

Even though lumping allows for enumeration of energy function probabilities with $O(n^p)$ operations, we also present a Laplace approximation to the MGF of its distribution for situations where this enumeration is burdensome. This MGF approximation may be inverted using a saddlepoint approximation. The accuracy of this method is illustrated in section 2.5.2.

To the best of our knowledge, the utility of energy function tail probabilities, as well as methods for approximating them, have not been addressed in the neural networks literature. First and second moment approximations have been considered as explained in and Hertz et al.(1991) and Kindermann & Snell (1980).

Our fourth contribution characterizes the set of possible absorbing states of the Markov chain underlying the *deterministic* Hopfield model, one of the first examples of a SHM. Our characterization specifies the absorbing states for fixed $\mathbf{E}$ and describes the functional structure of the absorption probabilities.A simple condition is also given which allows for the enumeration of all absorbing states. Previously, characterizations have assumed a random $\mathbf{E}$ where exemplar components are independent Bernoulli random variables as described by Amit (1989).

2.2 The Stochastic Hopfield Model

Hopfield (1982) proposed the deterministic Hopfield model as a means for approximating the process of associative memory. In this paper, the energy function along with the convention of using binary spin vectors, whose components assume the values "1" or "-1", were introduced. This choice of binary component variable was by design and ensured a direct correspondence between the deterministic Hopfield model

and the celebrated Ising model in statistical mechanics.

Classification in the deterministic Hopfield model proceeds through an iterative process starting at $\mathbf{u}_0$ and converging to an absorbing state of its underlying Markov chain. At each step in this procedure, a randomly selected component of $\mathbf{u} = (u_1, \ldots, u_n)^T$, the current state of the model, is updated in a deterministic way according to the rule:

$$\Pr\{u_i \longrightarrow \pm u_i\} = H(\pm u_i h_i). \tag{2.1}$$

Here $H(x)$ is the Heaviside function given as

$$H(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases},$$

h_i is the i^{th} component of $\mathbf{h} = \mathbf{W}\mathbf{u}$ and $\mathbf{W}$, the Hebb's rule weight matrix, is described as

$$\mathbf{W} = n^{-1}\mathbf{E}\mathbf{E}^T \tag{2.2}$$

where $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_q\}$ is the set of exemplars.

Often, another update rule is used that is identical with the one above except that the diagonal terms of weight matrix are set to zero. This results in a minor change in (2.1), i.e.,

$$\Pr\{u_i \rightarrow \pm u_i\} = H(\pm u_i h_i \mp q/n). \tag{2.3}$$

Our discussion below encompasses both update rules (2.1) and (2.3), however, we shall specifically address update rule (2.1) only.

Using either rule, the updating continues until the appearance of an absorbing state $\mathbf{u}_\infty \in \mathbf{U}_\infty$, where $\mathbf{U}_\infty$ is the set of all absorbing states. This state is now the Hopfield classification of $\mathbf{u}_0$. The set $\mathbf{U}_\infty$ is known to be non-empty and also have

an even number of states, however, there is no guarantee that any of these states are exemplars. When $\mathbf{u}_\infty$ is also in $\mathbf{E}$ we deem the classification of $\mathbf{u}_0$ correct if, among all exemplars, $\mathbf{u}_\infty$ and $\mathbf{u}_0$ are closest in Hamming distance, i.e., match in the most components. If $\mathbf{u}_\infty$ is not in $\mathbf{E}$, we shall conclude the classification process has failed.

Convergence to the set $\mathbf{E}$ can be ensured by replacing the deterministic update rule in (2.1) with a stochastic version and changing the criterion for classification to first passage to $\mathbf{E}$. Stochastic update rules were first introduced by Amit et al. (1985) where the deterministic change implemented by $H(x)$ was replaced by

$$\sigma(x) = (1 + e^{-2\beta x})^{-1} \quad \text{for} \quad 0 < \beta < \infty. \tag{2.4}$$

Hopfield's deterministic model may be recovered by passing to the limit as $\beta \to \infty$, and so, is a special case of their model. We call the model in which $\sigma(x)$ replaces $H(x)$, with $0 < \beta \leq \infty$, the stochastic Hopfield model.

From a probabilistic point of view, the SHM update dynamics represent a Markov process which we shall refer to as the *Hopfield* Markov chain (HMC). For finite β, this process is reversible, irreducible and has a stationary distribution given as

$$P(\mathbf{u}) = (Z(\beta))^{-1} \exp\left\{\frac{\beta}{2}\mathbf{u}^T\mathbf{W}\mathbf{u}\right\} \tag{2.5}$$

where $Z(\beta)$ the normalization or partition function given by

$$Z(\beta) = \sum_{\mathbf{u}\in\mathbf{U}} \exp\left\{\frac{\beta}{2}\mathbf{u}^T\mathbf{W}\mathbf{u}\right\}, \tag{2.6}$$

where $\mathbf{U}$ is the set of all spin vectors of length n. This distribution represents a linear exponential family in which $2^{-1}\mathbf{u}^T\mathbf{W}\mathbf{u}$ is the canonically sufficient statistic. The negative of this statistic is the energy function.

The HMC is simply too large to work with in practice. We, therefore, present a state space reduction that preserves the Markovian nature of the SHM as discussed in the next section.

2.3 The Lumped Stochastic Hopfield Model

2.3.1 Lumped Markov Processes

Suppose the states of a Markov chain are grouped into m disjoint megastates denoted by $\mathcal{L} = \{\mathcal{L}_0, \mathcal{L}_1, ..., \mathcal{L}_m\}$. The probability of transition from state k in $\mathcal{L}_i$ to some state in $\mathcal{L}_j$, is given by

$$p_{k,\mathcal{L}_j} = \sum_{l \in \mathcal{L}_j} p_{k,l}. \tag{2.7}$$

If

$$p_{k,\mathcal{L}_j} = p_{k',\mathcal{L}_j} \text{ for every state } k' \text{ in } \mathcal{L}_i, \tag{2.8}$$

one can consider the transition between megastates $\mathcal{L}_i$ and $\mathcal{L}_j$ without having to account for specific state to state transitions in the original Markov chain. Furthermore, if condition (2.8) is satisfied for any pair of megastates, then one can think about all transitions in the Markov chain as transitions between megastates. Fortunately, such a process on the megastates preserves the Markov property. Kemeny and Snell (1969) describe this process as being lumpable with respect to (w.r.t) megastates $\{\mathcal{L}_0, \mathcal{L}_1, ..., \mathcal{L}_m\}$. These authors also describe how to obtain the lumped transition matrix, $\mathbf{P}^L$, from $\mathbf{P}$, the original one, using the matrix operations:

$$\mathbf{P}^L = \mathbf{QPR}. \tag{2.9}$$

Here $\mathbf{Q}$ is a $m \times n$ matrix whose i^{th} row is zero except for uniform probability vector elements, $|\mathcal{L}_i|^{-1}$, in components corresponding to states in $\mathcal{L}_i$. $\mathbf{R}$, on the other hand, is an $n \times m$ matrix whose j^{th} column is zero everywhere, except for components corresponding to state in $\mathcal{L}_j$, where it assumes unity. Note that post-multiplication by $\mathbf{R}$ forms the megastate probabilities as in (2.7) and pre-multiplication by $\mathbf{Q}$ removes the redundancies found in $\mathbf{PR}$.

Many questions about a lumpable process can be answered with simpler computations based upon $\mathbf{P}^L$. Lumping can also provide new and interesting ways to look a problem. Unfortunately, in the analysis of a lumped Markov chain much information is lost about the original process, including perhaps first passage times. However, no information is lost about first passage times to megastates of size one. This is because, in a lumped Markov chain, the first passage time from state k in $\mathcal{L}_i$ to megastate $\mathcal{L}_j$ is the same for each k, as shown in Sumita and Rieders (1988). In the next section, we show how to lump the HMC so that exemplars are grouped by themselves. Therefore, the classification time for a SHM starting at $\mathbf{u}_0$ is equivalent to the first passage time from the megastate containing $\mathbf{u}_0$ to the set of megastates containing the exemplars.

2.3.2 The Lumped Hopfield Markov Chain

To show lumpability of the HMC, we start by partitioning the components of $\mathbf{e}_1$, the first exemplar. Denote ϵ_{p1} as the p^{th} component of $\mathbf{e}_1$ and the $(p,1)^{th}$ entry of $\mathbf{E}$. We define a *configuration* as a vector of sign differences that exist between e_{p1} and the remaining elements in the p^{th} row of $\mathbf{E}$. More formally, the sign differences in this row have configuration $l_p^*=(l_{p1}^*, l_{p2}^*, \ldots, l_{pq}^*)^T$ if

$$e_{p1} = l_{p2}^* e_{p2} = \ldots = l_{pq}^* e_{pq} \tag{2.10}$$

where $l_{pj}^* = e_{p1}e_{pj}$ and $l_{p1}^* \equiv 1$.

Suppose there are m distinct rows in $\mathbf{L}^*=[l_1^*, l_2^*, \ldots, l_n^*]^T$. Let A_1 denote the set of row indices that duplicate l_1^*, A_2 denote the set of row indices that duplicate the next distinct configuration, etc.. If we let n_i denote the cardinality of A_i, then clearly

$$\sum_{i=1}^{m} n_i = n.$$

Let the rows of $\mathbf{L} = [l_1, l_2, \ldots, l_m]^T$ be the m distinct configurations determined

from $\mathbf{L}^*$. A matrix $\mathbf{Q}$ can be determined such that

$$\mathbf{L} = \mathbf{Q}\mathbf{L}^* \tag{2.11}$$

and it is constructed in the following way: $\mathbf{Q}$ is a $m \times n$ matrix in whose i^{th} row consists of zeroes everywhere except for components whose indices are in A_i. Such components contain the elements of a uniform probability vector $n_i^{-1}\mathbf{1}$ where $\mathbf{1} = (1, \ldots, 1)^T$. Matrix $\mathbf{Q}$, in effect, collapses redundancies found in $\mathbf{L}^*$ according to the catalog of redundancies in $\{A_i\}$.

We now use this redundancy catalog to determine a set of megastates on $\mathbf{U}$. In order to motivate these megastates, however, we first consider how any spin vector $\mathbf{u}$ can be generated from $\mathbf{e}_1$. The first exemplar $\mathbf{e}_1$ provides a natural *point of reference* for this grouping since $\{A_i\}$ partition its components. These same megastates may also be determined using another exemplar as a point of reference as discussed in section 2.3.3.

In the mapping $\mathbf{e}_1 \longrightarrow \mathbf{u}$, let $k_i(\mathbf{u})$ denote the number of components of type A_i which must be changed in the conversion for $1 \leq i \leq m$. Now define

$$\mathbf{k}(\mathbf{u}) = (k_1(\mathbf{u}), k_2(\mathbf{u}), \ldots, k_m(\mathbf{u}))^T \tag{2.12}$$

to be the index vector for $\mathbf{u}$.

The vector of counted sign changes $\mathbf{k}(\mathbf{u})$ is now used to set up a relation among the spin vectors. Relation R on the set $\mathbf{U}$ is defined by

$$(\mathbf{u}_1, \mathbf{u}_2) \in R \text{ if } \mathbf{k}(\mathbf{u}_1) = \mathbf{k}(\mathbf{u}_2). \tag{2.13}$$

For each $\mathbf{u} \in \mathbf{U}$, let

$$\mathcal{L}_{\mathbf{u}} = \{\mathbf{v} : (\mathbf{u}, \mathbf{v}) \in R\} \tag{2.14}$$

be the megastate generated by spin vector $\mathbf{u}$. In addition, let $\mathcal{L}$ be the collection of all generated megastates, i.e.,

$$\mathcal{L} = \{\mathcal{L}_{\mathbf{u}} : \mathbf{u} \in \mathbf{U}\}.$$

Lemma 1 *Relation R in (2.13) is an equivalence relation on the spin vectors in $\mathbf{U}$. As such, the equivalence classes specified in (2.14) partition $\mathbf{U}$ and form our megastates.*

Proof Relation R is an equivalence relation if it is reflexive, symmetric and transitive.

i) R is reflexive since $(\mathbf{u}, \mathbf{u}) \in R$ for every $\mathbf{u} \in \mathbf{U}$.

ii) R is symmetric since for all $\mathbf{u}, \mathbf{v} \in \mathbf{U}$, if $(\mathbf{u}, \mathbf{v}) \in R$, then $(\mathbf{v}, \mathbf{u}) \in R$.

iii) R is transitive since for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbf{U}$, if $(\mathbf{u}, \mathbf{v})$ and $(\mathbf{v}, \mathbf{w}) \in R$, then $(\mathbf{u}, \mathbf{w}) \in R$.

$\diamond$

As a result of this equivalence relation, the indexing of equivalence classes $\mathcal{L}$ is based on $\mathbf{k}(\mathbf{u})$ and not $\mathbf{u}$. Let

$$\mathbf{k} = (k_1, k_2, ..., k_m)^T$$

where each component k_i assumes a value between 0 and n_i. In addition, let $\mathcal{K}$ be the set of all $\mathbf{k}$ vectors. Now, the set of megastates may be written as

$$\mathcal{L} = \{\mathcal{L}_{\mathbf{k}} : \mathbf{k} \in \mathcal{K}\}, \tag{2.15}$$

where $\mathcal{L}$ consists of

$$\prod_{i=1}^{m} (n_i + 1) \tag{2.16}$$

distinct megastates. A particular megastate $\mathcal{L}_{\mathbf{k}}$ consists of

$$N_{\mathbf{k}} = \prod_{i=1}^{m} \binom{n_i}{k_i} \tag{2.17}$$

states since in the mapping $\mathbf{e}_1 \longrightarrow \mathbf{u}$ and within each A_i, k_i out of n_i components are chosen for sign changes. Each state in $\mathcal{L}_{\mathbf{k}}$ is Hamming distance $\sum k_i$ from $\mathbf{e}_1$: therefore megastates are not characterized by their distance from $\mathbf{e}_1$. Megastates are, however, characterized by the index vector $\mathbf{k}$ which can be thought of as the collection of Hamming distances from the m subvectors which comprise $\mathbf{e}_1$, i.e., $\{\mathbf{e}_1|A_1, \mathbf{e}_1|A_2, \ldots, \mathbf{e}_1|A_m\}$ where $\mathbf{e}_1|A_i$ consists of the components in $\mathbf{e}_1$ whose indices are in A_i. Because of this characterization, we shall refer to $\mathbf{k}$ as the Hamming index vector of $\mathcal{L}_{\mathbf{k}}$.

Example An SHM has two exemplars given by

$$\mathbf{E}^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & 1 \end{bmatrix}.$$

There are two distinct configurations given by the rows of $\mathbf{L}$ where

$$\mathbf{L} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \tag{2.18}$$

and with repetitions $\mathbf{n} = (3, 1)$.

Space $\mathbf{U}$ consists of $2^4 = 16$ spin vectors which are partitioned into 8 megastates. The catalog of these megastates is as follows:

$\mathcal{L}_{00}$	$\mathcal{L}_{01}$	$\mathcal{L}_{10}$	$\mathcal{L}_{11}$
$(+1,+1,+1,+1)^T$	$(+1,+1,+1,-1)^T$	$(-1,+1,+1,+1)^T$	$(-1,+1,+1,-1)^T$
		$(+1,-1,+1,+1)^T$	$(+1,-1,+1,-1)^T$
		$(+1,+1,-1,+1)^T$	$(+1,+1,-1,-1)^T$

$\mathcal{L}_{20}$	$\mathcal{L}_{21}$	$\mathcal{L}_{30}$	$\mathcal{L}_{31}$
$(+1,-1,-1,+1)^T$	$(+1,-1,-1,-1)^T$	$(-1,-1,-1,+1)^T$	$(-1,-1,-1,-1)^T$
$(-1,+1,-1,+1)^T$	$(-1,+1,-1,-1)^T$		
$(-1,-1,+1,+1)^T$	$(-1,-1,+1,-1)^T$		

Note that the two exemplars form megastates of size one. In addition, the following lemma will be illustrated using this example.

Lemma 2 *With the system of indexing from* (2.12), $\mathbf{e}_j$ *has a particularly simple form given as*

$$\mathbf{e}_j \equiv \mathcal{L}_{2^{-1}\mathbf{D_n}(\mathbf{1}-\bar{l}_j)}$$

where $\mathbf{D_n} = diag\{n_1, n_2, \ldots, n_m\}$ *is a diagonal matrix whose* $(i,i)^{th}$ *entry is* n_i *and* $\bar{l}_j$ *is the* j^{th} *column of* $\mathbf{L}$.

◇

Example (continued) The index vector for $\mathbf{e}_1$ should be

$$2^{-1}\mathbf{D_n}\left(\mathbf{1}-(1,1)^T\right) = (0,0)^T$$

and for $\mathbf{e}_2$ the index should be

$$2^{-1}\mathbf{D_n}\left(\mathbf{1}-(-1,1)^T\right) = (3,0)^T$$

Proof For every $p \in A_i$, $e_{pj} = l_{ij}e_{p1}$. Therefore, l_{ij} is the indicator of sign agreement between $\mathbf{e}_1|A_i$ and $\mathbf{e}_j|A_i$. If $l_{ij} = 1$, then $\mathbf{e}_j|A_i$ and $\mathbf{e}_1|A_i$ must match in all n_i places so that k_i must be n_i. On the other hand, if $l_{ij} = -1$, then $\mathbf{e}_j|A_i$ and $\mathbf{e}_1|A_i$ match in none of the n_i places, meaning that $k_i = 0$. This relationship between k_i and l_{ij} can be more succinctly written as $k_i = n_i(1 - l_{ij})/2$ which in turn yields $\mathbf{k} = 2^{-1}\mathbf{D_n}\left(\mathbf{1}-\bar{l}_j\right)$.

◇

If a state's Hamming index vector is such that $k_i = 0$ or n_i for all i, we shall call it a *corner* state. In addition, we shall denote the set of all corner states by $\mathbf{C}$. Corner states consist of megastates of size one and include the exemplars, $\mathbf{E}$, the negative exemplars, $-\mathbf{E}$, as well as many other states of interest which will be described in section 2.3.4.

We now say a little more about the relationship between spin vectors and index vectors.

Lemma 3 *For any* $\mathbf{u} \in \mathbf{U}$

$$\mathbf{E}^T\mathbf{u} = \mathbf{L}^T(\mathbf{n}-2\mathbf{k}) \tag{2.19}$$

Proof The p^{th} component of $\mathbf{E}^T\mathbf{u}$ is

$$\mathbf{e}_p^T\mathbf{u} = n - 2d(\mathbf{u}, \mathbf{e}_p) \tag{2.20}$$

where $d(\mathbf{u}, \mathbf{e}_p)$ is the Hamming from $\mathbf{u}$ to $\mathbf{e}_p$. The discrepancy between the above expression and $d(\mathbf{u}, \mathbf{e}_p)$ results from the use of spin vectors for which component disagreement yields "-1" instead of zero. From (2.20), expression (2.19) is simplified to

$$\mathbf{E}^T\mathbf{u} = n\mathbf{1}-2\mathbf{d}. \tag{2.21}$$

Distance vector $\mathbf{d}$ can be written in entirely terms of $\mathbf{n}$ and $\mathbf{k}$ by showing

$$d(\mathbf{u}, \mathbf{e}_j) = d(\mathbf{e}_1, \mathbf{e}_j) + \sum_{i=1}^{m} l_{ij}k_i. \tag{2.22}$$

To show this, consider the process of sign changes required to map $\mathbf{e}_1$ into $\mathbf{u}$. Start with $\mathbf{u} = \mathbf{e}_1$ so that $d(\mathbf{u}, \mathbf{e}_j) = d(\mathbf{e}_1, \mathbf{e}_j)$. Next, change k_i components in A_i for $i = 1, \ldots, m$. If $l_{ij} = -1$, so that $\mathbf{e}_1|A_i$ and $\mathbf{e}_j|A_i$ differed in all components prior to the change, then $d(\mathbf{u}, \mathbf{e}_j)$ must decrease by k_i. In a similar fashion, if $l_{ij} = 1$ then $d(\mathbf{u}, \mathbf{e}_j)$ must increase by k_i. Expression (2.22) simply implements these corrections.

The Hamming distance from $\mathbf{e}_1$ to $\mathbf{e}_j$ is the component sum of the Hamming index vector. Therefore, lemma 2 gives

$$d(\mathbf{e}_1, \mathbf{e}_j) = 2^{-1}\sum_{i=1}^{m} n_i(1 - l_{ij}). \tag{2.23}$$

Combining (2.22) and (2.23) has the simple form

$$d(\mathbf{u}, \mathbf{e}_j) = 2^{-1}\left\{n - \bar{l}_j^T(\mathbf{n}-2\mathbf{k})\right\} \tag{2.24}$$

which in turns yields

$$\mathbf{d} = 2^{-1}\left\{n\mathbf{1} - \mathbf{L}^T(\mathbf{n}-2\mathbf{k})\right\}.$$

Substitution of this expression into (2.21) proves the lemma.

◇

Lemma 4 will be required in the proof of lumpability.

Lemma 4 *For p an element of A_i, let h_p be the p^{th} entry of $\mathbf{h} = \mathbf{Wu}$, then*

$$h_p = e_{p1} h_i^{\mathcal{L}}$$

where $h_i^{\mathcal{L}}$ is the i^{th} entry of

$$\mathbf{h}^{\mathcal{L}} = \mathbf{Q}\mathbf{D}_{\mathbf{e}_1}\mathbf{h}, \tag{2.25}$$

and $\mathbf{D}_{\mathbf{e}_1} = diag\{e_{11}, e_{21}, \ldots, e_{n1}\}$.

Proof From lemma 3.

$$\mathbf{h} = n^{-1}\mathbf{E}\mathbf{E}^T\mathbf{u} = n^{-1}\mathbf{E}\mathbf{L}^T(\mathbf{n}-2\mathbf{k}). \tag{2.26}$$

From this expression, we see that the p^{th} entry of $\mathbf{h}$ is

$$h_p = n^{-1}\bar{\mathbf{e}}_p\mathbf{L}^T(\mathbf{n}-2\mathbf{k})$$

where $\bar{\mathbf{e}}_p$ is the p^{th} row of $\mathbf{E}$. When $p \in A_i$, configuration l_i records the sign differences in $\bar{\mathbf{e}}_p$; therefore

$$\bar{\mathbf{e}}_p = e_{p1}\bar{l}_i^T.$$

This in turn yields

$$h_p = n^{-1} e_{p1} l_i^T \mathbf{L}^T (\mathbf{n}-2\mathbf{k}) .$$

which is e_{p1} times the i^{th} entry of

$$\mathbf{h}^{\mathcal{L}} = n^{-1} \mathbf{L}\mathbf{L}^T (\mathbf{n}-2\mathbf{k}) . \tag{2.27}$$

Lemma 3 and formula (2.11) yield the following simplifications of the above expression:

$$\begin{aligned} \mathbf{h}^{\mathcal{L}} &= n^{-1} \mathbf{L}\mathbf{E}^T \mathbf{u} \\ &= n^{-1} \mathbf{Q}\mathbf{L}^* \mathbf{E}^T \mathbf{u}. \end{aligned}$$

Expression (2.25) results after noting that

$$\mathbf{L}^* = \mathbf{D}_{\mathbf{e}_i} \mathbf{E}.$$

◇

Note the similarities in the first part of expression (2.26) and the equality in (2.27) . Their similarity provides an interpretation of the lumped SHM is a *SHM-type* model defined by a weight matrix, given by Hebb's rule, and using generalized spin vectors $\mathbf{u}^{\mathcal{L}} = \mathbf{n}-2\mathbf{k}$.

The sequence of lemmas allow us to prove our main result; that the HMC preserves its Markov property when grouped into the megastates as previously described.

Theorem 1 *Transitions among the megastates defined in Lemma* 1 *preserve the Markov property. Therefore, the Hopfield Markov chain is lumpable with respect to these megastates. Transition probabilities for the lumped Markov chain are given by*

the following formulas:

$$P_{\mathbf{k},\mathbf{k}_i(-1)} = \frac{k_i}{n}\sigma\left(h_i^{\mathcal{L}}\right) \tag{2.28}$$

$$P_{\mathbf{k},\mathbf{k}_i(+1)} = \frac{n_i - k_i}{n}\sigma\left(-h_i^{\mathcal{L}}\right) \tag{2.29}$$

for $1 \leq i \leq m$ *where*

$$\mathbf{k}_i(\pm 1) = (k_1, \ldots, k_i \pm 1, \ldots, k_m)$$

and $\sigma(\cdot)$ *is given in* (2.4).

Proof We first need to show that the transition probabilities associated with the collection of megastates satisfy the grouping condition in (2.7). Let $\mathbf{u}$ be an arbitrary vector in $\mathcal{L}_{\mathbf{k}}$. The probability of transition from $\mathbf{u}$ to the megastate $\mathcal{L}_{\mathbf{k}_i(-1)}$ is given as

$$\Pr\left\{\mathbf{u} \longrightarrow \mathcal{L}_{\mathbf{k}_i(-1)}\right\} = \sum_{j=1}^{N_{\mathbf{k}_i(-1)}} \Pr\left\{\mathbf{u} \longrightarrow \mathbf{v}_j\right\} \tag{2.30}$$

where $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{N_{\mathbf{k}_i(-1)}}\}$ is the set of vectors in $\mathcal{L}_{\mathbf{k}_i(-1)}$. A transition to $\mathcal{L}_{\mathbf{k}_i(-1)}$ implies a component in A_i which previously did not match e_{p1} has to be changed to match. In the above sum, there are k_i non-zero probabilities each of which is given as

$$\Pr\{u_p \to -u_p\} = n^{-1}\sigma(-u_p h_p) \tag{2.31}$$

by the SHM version of (2.3). The "n^{-1}" term results from our initial randomization in choosing a component of $\mathbf{u}$ for updating. Lemma 4 states that $h_p = e_{p1}h_i^{\mathcal{L}}$ when $p \in A_i$ and $-u_p e_{p1} = 1$ since u_p does not match e_{p1}. Therefore, we can write the above expression as

$$\Pr\{u_p \to -u_p\} = n^{-1}\sigma\left(h_i^{\mathcal{L}}\right). \tag{2.32}$$

Since (2.32) is summed k_i times in (2.30) the final transition probability is given as in (2.28). Similar arguments give the complementary transition probability (2.29). The transition probabilities in (2.28) and (2.29) only depend upon the membership of $\mathbf{u}$ in $\mathcal{L}$. Therefore, condition (2.7) is satisfied proving the HMC is lumpable w.r.t $\mathcal{L}$.

$\diamond$

Example (continued) The lumped Markov chain determined from this technique is given as

$$
\begin{array}{c}
 \\
\mathcal{L}_{00} \\
\mathcal{L}_{01} \\
\mathcal{L}_{10} \\
\mathcal{L}_{11} \\
\mathcal{L}_{20} \\
\mathcal{L}_{21} \\
\mathcal{L}_{30} \\
\mathcal{L}_{31}
\end{array}
\begin{array}{c}
\begin{array}{cccccccc}
\mathcal{L}_{00} & \mathcal{L}_{01} & \mathcal{L}_{10} & \mathcal{L}_{11} & \mathcal{L}_{20} & \mathcal{L}_{21} & \mathcal{L}_{30} & \mathcal{L}_{31}
\end{array} \\
\left(\begin{array}{cccccccc}
* & \frac{1}{4}\bar{\sigma}\left(\frac{1}{2}\right) & \frac{3}{4}\bar{\sigma}\left(\frac{3}{2}\right) & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{4}\bar{\sigma}\left(\frac{1}{2}\right) & * & 0 & \frac{3}{4}\bar{\sigma}\left(\frac{3}{2}\right) & 0 & 0 & 0 & 0 \\
\frac{1}{4}\sigma\left(\frac{1}{2}\right) & 0 & * & \frac{1}{4}\bar{\sigma}\left(\frac{1}{2}\right) & \frac{1}{2}\bar{\sigma}\left(\frac{1}{2}\right) & 0 & 0 & 0 \\
0 & \frac{1}{4}\sigma\left(\frac{1}{2}\right) & \frac{1}{4}\bar{\sigma}\left(\frac{1}{2}\right) & * & 0 & \frac{1}{2}\bar{\sigma}\left(\frac{1}{2}\right) & 0 & 0 \\
0 & 0 & \frac{1}{2}\bar{\sigma}\left(\frac{1}{2}\right) & 0 & * & \frac{1}{4}\bar{\sigma}\left(\frac{1}{2}\right) & \frac{1}{4}\sigma\left(\frac{1}{2}\right) & 0 \\
0 & 0 & 0 & \frac{1}{2}\bar{\sigma}\left(\frac{1}{2}\right) & \frac{1}{4}\bar{\sigma}\left(\frac{1}{2}\right) & * & 0 & \frac{1}{4}\sigma\left(\frac{1}{2}\right) \\
0 & 0 & 0 & 0 & \frac{3}{4}\bar{\sigma}\left(\frac{3}{2}\right) & 0 & * & \frac{1}{4}\bar{\sigma}\left(\frac{1}{2}\right) \\
0 & 0 & 0 & 0 & 0 & \frac{3}{4}\bar{\sigma}\left(\frac{3}{2}\right) & \frac{1}{4}\bar{\sigma}\left(\frac{1}{2}\right) & *
\end{array}\right)
\end{array}
$$

For notational convenience, we have defined $\bar{\sigma}(\cdot)$ as

$$\bar{\sigma}(x) = \sigma(-x) = 1 - \sigma(x).$$

Also, the starred diagonal entries of $\mathbf{P}^{\mathcal{L}}$ have been omitted to yield a more concise description. Such entries are simply sums of remaining row probabilities but with a sign change in the argument of $\sigma(\cdot)$, i.e.,

$$\Pr\{\mathcal{L}_{00} \to \mathcal{L}_{00}\} = \frac{1}{4}\sigma\left(\frac{1}{2}\right) + \frac{3}{4}\sigma\left(\frac{3}{2}\right).$$

In the next section, we consider the possibility of further lumping in the presence of symmetry.

2.3.3 Further Lumping

In the above example, there exists a symmetry among the transition probabilities given by

$$P_{(i,j),(k,l)} = P_{(n_1-i,j),(n_1-k,l)}. \tag{2.33}$$

As such, we may further lump $\mathcal{L}$ while keeping the exemplars isolated. The new partition is $\mathcal{L}^{(1)} = \left\{\mathcal{L}_{00}^{(1)}, \mathcal{L}_{01}^{(1)}, \mathcal{L}_{10}^{(1)}, \mathcal{L}_{11}^{(1)}\right\}$ where $\mathcal{L}_{00}^{(1)} = \mathcal{L}_{00} \cup \mathcal{L}_{30} = \{\mathbf{e}_1, \mathbf{e}_2\}$, $\mathcal{L}_{01}^{(1)} = \mathcal{L}_{01} \cup \mathcal{L}_{31}$, $\mathcal{L}_{10}^{(1)} = \mathcal{L}_{10} \cup \mathcal{L}_{20}$ and $\mathcal{L}_{11}^{(1)} = \mathcal{L}_{11} \cup \mathcal{L}_{21}$. The transition matrix for the process on $\mathcal{L}^{(1)}$ is given as

$$\begin{array}{c} \\ \mathcal{L}_{00}^{(1)} \\ \mathcal{L}_{01}^{(1)} \\ \mathcal{L}_{10}^{(1)} \\ \mathcal{L}_{11}^{(1)} \end{array} \begin{array}{c} \begin{array}{cccc} \mathcal{L}_{00}^{(1)} & \mathcal{L}_{01}^{(1)} & \mathcal{L}_{10}^{(1)} & \mathcal{L}_{11}^{(1)} \end{array} \\ \begin{pmatrix} * & \frac{1}{4}\bar{\sigma}\left(\frac{1}{2}\right) & \frac{3}{4}\bar{\sigma}\left(\frac{3}{2}\right) & 0 \\ \frac{1}{4}\bar{\sigma}\left(\frac{1}{2}\right) & * & 0 & \frac{3}{4}\bar{\sigma}\left(\frac{3}{2}\right) \\ \frac{1}{4}\sigma\left(\frac{1}{2}\right) & 0 & \frac{3}{4} & \frac{1}{4}\bar{\sigma}\left(\frac{1}{2}\right) \\ 0 & \frac{1}{4}\sigma\left(\frac{1}{2}\right) & \frac{1}{4}\bar{\sigma}\left(\frac{1}{2}\right) & \frac{3}{4} \end{pmatrix} \end{array}$$

This transition matrix also has a symmetry we can describe as

$$P_{(i,j),(k,l)}^{(1)} = P_{(i,n_2-j),(k,n_2-l)}^{(1)} \tag{2.34}$$

Therefore, if we expand the exemplar set to include $-\mathbf{E}$, as is sometimes done in the neural network literature, then even further lumping is possible which isolates the exemplars. In general, we can always place $\mathbf{e}_i$ and $-\mathbf{e}_i$ in the same megastate for expanded exemplar sets. Now we have only two megastates $\mathcal{L}^{(2)} = \left\{\mathcal{L}_{00}^{(2)}, \mathcal{L}_{10}^{(2)}\right\}$ where $\mathcal{L}_{00}^{(2)} = \mathcal{L}_{00}^{(1)} \cup \mathcal{L}_{01}^{(1)} = -\mathbf{E} \cup \mathbf{E}$ and $\mathcal{L}_{10}^{(2)} = \mathcal{L}_{10}^{(1)} \cup \mathcal{L}_{11}^{(1)}$ with associated transition matrix

$$\begin{array}{c} \\ \mathcal{L}_{00}^{(2)} \\ \mathcal{L}_{10}^{(2)} \end{array} \begin{array}{c} \begin{array}{cc} \mathcal{L}_{00}^{(2)} & \mathcal{L}_{10}^{(2)} \end{array} \\ \begin{pmatrix} 1-\frac{3}{4}\bar{\sigma}(3) & \frac{3}{4}\bar{\sigma}(3) \\ \frac{1}{4}\sigma(1) & 1-\frac{1}{4}\sigma(1) \end{pmatrix} \end{array}$$

Through this lumping process, we are able to determine that the classification time distribution to $-\mathbf{E} \cup \mathbf{E}$ from the set of non-exemplars is Geometric with parameter $\frac{1}{4}\sigma(1)$. In general, lumping alone will not determine classification time distributions. A lumped transition matrix, however, can be used with Pyke's (1961) rule to calculate exact classification time MGFs, as described in section 2.4.1.

In the context of treating two exemplars only, one can always perform further lumping which combines exemplar megastates. This is a consequence of the mutual orthogonality of the rows in $\mathbf{L}$ as seen in (2.18). When these rows are orthogonal (2.27) simplifies to

$$\mathrm{h}^{\mathcal{L}} = \frac{q}{n}(\mathrm{n}-2\mathrm{k}). \tag{2.35}$$

As a result, lumped transition probabilities will satisfy equality (2.33) since $h_i^{\mathcal{L}}$ is a function of only n_i and k_i. When dealing with three or more exemplars, further lumping is possible when $\mathbf{L}$ has mutually orthogonal rows.

In the previous section, we used $\mathbf{e}_1$ as the point of reference when generating megastates as in (2.14). It turns out that one would calculate the same megastates if another exemplary point of reference were used.

Theorem 2 *The same megastates are generated in* (2.14) *for any point of reference vector chosen in* $\mathbf{E}$.

Proof Consider $\mathbf{u} \in \mathcal{L}_{\mathbf{k}}$ where the Hamming index vector $\mathbf{k}$ comes from using point of reference $\mathbf{e}_1$ in (2.14). We analyze how $\mathbf{k}$ would be different if $\mathbf{e}_j$ were used as a point

of reference instead. The Hamming index k_i describes the number of components in which $\mathbf{u}|A_i$ and $\mathbf{e}_1|A_i$ differ. If $l_{ij} = 1$, then $\mathbf{e}_1|A_i = \mathbf{e}_j|A_i$ and k_i is the same for both points of reference. On the other hand, if $l_{ij} = -1$, then $\mathbf{e}_1|A_i = -\mathbf{e}_j|A_i$ and k_i would be $n_i - k_i$ for the following reason. The configuration from the perspective of $\mathbf{e}_j$ is minus the perspective from $\mathbf{e}_1$ thereby reversing the roles k_i and $n_i - k_i$. From this discussion, we see that changing the point of reference changes the Hamming index vector but leaves the megastates intact.

$\diamond$

2.3.4 Absorbing States

It is well-known the deterministic Hopfield model will always have at least two absorbing states. In addition, the number of absorbing states must be even since if the vector $\mathbf{u}$ is absorbing then $-\mathbf{u}$ must also be absorbing.

For a deterministic Hopfield model with one exemplar, the set of absorbing states, $\mathbf{U}_\infty$, will consist of the exemplars and the negative exemplars, i.e., $-\mathbf{E} \cup \mathbf{E}$, as proven in Hertz et al. (1991). When this model has two exemplars, it is still true that $\mathbf{U}_\infty = -\mathbf{E} \cup \mathbf{E}$ as shown below in theorem 5. Unfortunately, for a SHM with three or more exemplars, $\mathbf{U}_\infty$ is not so simple. Typically $\mathbf{U}_\infty$ will contain many non-exemplary states and may contain no exemplars at all. For example, if a deterministic Hopfield model uses the set of exemplars from section 2.4.3, $\mathbf{U}_\infty$ consists of two states neither of which are exemplars. These two states are

$$\mathbf{u} = (+1, +1, +1, +1, +1, +1, +1, +1, +1, -1)^T$$

and

$$-\mathbf{u} = (-1, -1, -1, -1, -1, -1, -1, -1, -1, +1)^T,$$

and were determined by inspecting transition probabilities. For general q, $\mathbf{E}$ is simply not big enough to contain $\mathbf{U}_\infty$.

Theorem 3 *The set of corner states,* $\mathbf{C}$*, contains* $\mathbf{U}_\infty$*.*

Proof For a deterministic Hopfield model, exit probabilities from megastate $\mathcal{L}_\mathbf{k}$ are determined from theorem 1 as

$$\begin{aligned} P_{\mathbf{k},\mathbf{k}_i(-1)} &= \frac{k_i}{n} H\left(h_i^{\mathcal{L}}\right) \\ P_{\mathbf{k},\mathbf{k}_i(+1)} &= \frac{n_i - k_i}{n} H\left(-h_i^{\mathcal{L}}\right) \end{aligned} \tag{2.36}$$

where $1 \leq i \leq m$. If $\mathbf{u}$ is in $\mathcal{L}_\mathbf{k}$ as well as $\mathbf{U}_\infty$, then all exit probabilities must vanish. Clearly, at least one of the quantities $H\left(h_i^{\mathcal{L}}\right)$ or $H\left(-h_i^{\mathcal{L}}\right)$ must be unity for any given $h_i^{\mathcal{L}}$. Therefore, the only way all of the above probabilities could be zero is if k_i is 0 or n_i for $i = 1, \ldots, m$. This occurs only when $\mathbf{u} \in \mathbf{C}$.

◇

Theorem 4 *When the configurations of* $\mathbf{E}$ *are mutually orthogonal, or equivalently* $\mathbf{L}$ *has orthogonal rows, then* $\mathbf{U}_\infty$ *and* $\mathbf{C}$ *are identical.*

Proof From (2.35), $h_i^{\mathcal{L}}$ takes on a very special form;

$$h_i^{\mathcal{L}} = \begin{cases} n^{-1}qn_i & \text{when } k_i = 0 \\ -n^{-1}qn_i & \text{when } k_i = n_i \end{cases} \tag{2.37}$$

in expression (2.36). Now, it is easy to verify a corner state's exit probabilities must be zero.

◇

It should be clear from the previous proof that $\mathbf{U}_\infty \subseteq \mathbf{C}$ could not be developed without theorem 1 and the lumped transition probabilities.

Theorem 5 *If the deterministic Hopfield model contains two exemplars, then* $\mathbf{U}_\infty$ *and* $-\mathbf{E} \cup \mathbf{E}$ *are identical.*

Proof If the two exemplars are linearly dependent, then it necessarily follows that $\mathbf{e}_1 = -\mathbf{e}_2$. As such, $\mathbf{E}$ has a single configuration and $h_1^{\mathcal{C}}$ assumes the following values:

$$h_1^{\mathcal{C}} = \begin{cases} n^{-1}2n_1 & \text{when } k_1 = 0 \\ -n^{-1}2n_1 & \text{when } k_1 = n_1 \end{cases} . \tag{2.38}$$

Therefore, $h_1^{\mathcal{C}}$ is a special case of (2.37) where $i = 1$ and $q = 2$. As such, by the proof of theorem 4, since all exit probabilities for states in $-\mathbf{E} \cup \mathbf{E}$ must be zero.

On the other hand, if these two exemplars are linearly independent the matrix $\mathbf{L}$ has one of two forms:

$$\mathbf{L} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \text{ or } \mathbf{L} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

From this, is it clear that

$$\mathbf{C} = \{(0,0), (0,n_2), (n_1,n_2), (n_1,0)\} = -\mathbf{E} \cup \mathbf{E}$$

and either $\mathbf{L}$ has orthogonal rows. Therefore, $\mathbf{U}_\infty = \mathbf{C}$, by theorem 4, and our result is proven.

◇

2.3.5 Energy Function Probabilities

It is possible to simplify the partition function, and thus the stationary spin vector distribution $P(\mathbf{u})$ in (2.5), using the results from lumping. From lemma 2, we rewrite $\mathbf{u}^T\mathbf{W}\mathbf{u}$ as

$$\mathbf{u}^T\mathbf{W}\mathbf{u} = n^{-1}\mathbf{u}^T\mathbf{E}\mathbf{E}^T\mathbf{u} = n^{-1}\,(\mathbf{n}-2\mathbf{k})^T\,\mathbf{L}\mathbf{L}^T\,(\mathbf{n}-2\mathbf{k})\,. \tag{2.39}$$

Note that two spin vectors in the same megastate have the same energy function value since this value depends only on the vector $\mathbf{k}$. This expression, furthermore, allows us to list the values of $\mathbf{u}^T\mathbf{W}\mathbf{u}$ and $N_{\mathbf{k}}$ with $O(n^m)$ operations as

$$\left\{\left(-\frac{\beta}{2n}(\mathbf{n}-2\mathbf{k})^T\mathbf{L}\mathbf{L}^T(\mathbf{n}-2\mathbf{k}), N_{\mathbf{k}}\right)\right\} \text{ for } \mathbf{k}\in\mathcal{K} \tag{2.40}$$

where $\mathcal{K}$ has size $O(n^m)$ and is given in (2.16). With this list, the partition function is calculated as

$$Z(\beta) = \sum_{\mathbf{k}\in\mathcal{K}} N_{\mathbf{k}} \exp\left\{\frac{\beta}{2n}(\mathbf{n}-2\mathbf{k})^T\mathbf{L}\mathbf{L}^T(\mathbf{n}-2\mathbf{k})\right\}. \tag{2.41}$$

Notice that the quadratic form in (2.39) is unchanged by the transformation

$$\mathbf{k} \longrightarrow \mathbf{n}-\mathbf{k}. \tag{2.42}$$

As such, the list in (2.40) be generated while only enumerating roughly half of the k_j values for some j. This list is now written as

$$\left\{\left(-\frac{\beta}{2n}(\mathbf{n}-2\mathbf{k})^T\mathbf{L}\mathbf{L}^T(\mathbf{n}-2\mathbf{k}), N_{\mathbf{k}}\right)\right\} \text{ for } \mathbf{k}\in\mathcal{K}^* \tag{2.43}$$

where

$$\mathcal{K}^* = \{\mathbf{k} : k_i = 0,\ldots,n_i^*(j)\}$$

and

$$n_i^*(j) = \begin{cases} n_i & \text{if } i\neq j \\ \lfloor \frac{n_i}{2} \rfloor & \text{if } i=j \end{cases}.$$

When $\mathbf{L}$ has orthogonal rows, even further reduction is possible. Write (2.41) as

$$Z(\beta) = \sum_{\mathbf{k}\in\mathcal{K}}\prod_{i=1}^{m}\binom{n_i}{k_i}\exp\left\{\frac{q\beta}{2n}(\mathbf{n}-2\mathbf{k})^T(\mathbf{n}-2\mathbf{k})\right\} \tag{2.44}$$

This sum may be broken into a product of m sums each of which depends on a single value of k_i, i.e.,

$$Z(\beta) = 2^m \prod_{i=1}^{m} \left(\sum_{k_i=0}^{\lfloor \frac{n_i}{2} \rfloor} N_{k_i} \exp\left\{ \frac{q\beta}{2n} (n_i - 2k_i)^2 \right\} \right)$$

where

$$N_{k_i} = \begin{cases} \frac{1}{2}\binom{n_i}{k_i} & \text{if } k_i = \frac{n_i}{2} \\ \binom{n_i}{k_i} & \text{otherwise} \end{cases} .$$

This yields a $O(2^m)$ reduction in the number of operations required to evaluate the partition function.

One can also obtain an equivalent reduction in a listing of possible values of the energy function. Now this list can be written as

$$\left\{ -\frac{\beta}{2n} (\mathbf{n}-2\mathbf{k})^T (\mathbf{n}-2\mathbf{k}) \right\} \text{ where } k_i = 0, 1, \ldots, \left\lfloor \frac{n_i}{2} \right\rfloor \text{ for } i = 1, 2, \ldots, m. \quad (2.45)$$

This simplification is a result the *m-fold* symmetry of $(\mathbf{n}-2\mathbf{k})^T (\mathbf{n}-2\mathbf{k})$. This quantity is invariant under the transformation

$$k_i \longrightarrow n_i - k_i \text{ for any } i \in \{1, 2, \ldots, m\}. \quad (2.46)$$

Unfortunately, we cannot include $N_{\mathbf{k}}$ in this list since it is typically invariant under the transformation in (2.42) and not the one in (2.46).

The above reductions allow calculation of the energy function's PMF in $O(n^m)$ operations as

$$\begin{aligned} P\left(-\frac{1}{2}\mathbf{u}^T\mathbf{W}\mathbf{u} = x \right) &= \sum_{\left\{ \mathbf{u}: -\frac{1}{2}\mathbf{u}^T\mathbf{W}\mathbf{u} = x \right\}} P(\mathbf{u}) \\ &= \frac{\exp\{-\beta x\} N(x)}{Z(\beta)} \end{aligned} \quad (2.47)$$

where $N(x)$ is given by

$$N(x) = \sum_{\mathbf{k} \in \mathcal{K}(x)} N_{\mathbf{k}}$$

and

$$\mathcal{K}(x) = \{\mathbf{k} \in \mathcal{K} : -2^{-1}\mathbf{u}^T\mathbf{W}\mathbf{u} = x \text{ for any } \mathbf{u} \in \mathcal{L}_{\mathbf{k}}\}. \tag{2.48}$$

2.4 MGF Calculation and Saddlepoint Approximation

Classification times in the SHM are random variables whose MGFs we now compute. We use Pyke's (1961) rule to determine the MGF of the lumped Markov chain and invert this MGF with a saddlepoint approximation to yield nearly exact answers.

2.4.1 Pyke's Rule

The MGF of this classification time distribution is determined from Pyke's rule as discussed in Butler and Huzurbazar (1999). Suppose $\mathbf{T}(s) = e^s\mathbf{P}$ is the transmittance matrix of the Hopfield Markov chain, e.g., $\mathbf{P}$ is the transition probability matrix and e^s is the MGF of the degenerate distribution with time one. Define the *equivalent* transmittance matrix as

$$\mathfrak{M}(s) = (\mathfrak{M}_{ij}(s)) = \mathbf{T}(s)[\mathbf{I} - \mathbf{T}(s)]^{-1}\left(\left\{[\mathbf{I} - \mathbf{T}(s)]^{-1}\right\}_d\right)^{-1} \quad \text{for } s \neq 0 \tag{2.49}$$

where $\{\cdot\}_d$ is the matrix operator which replaces all non-diagonal elements with zero.

Consider the SHM with one exemplar which is treated as an absorbing state. According to Pyke's rule, $\mathfrak{M}_{i\mathbf{e}_1}(0)$ is the probability of classifying state i as exemplar $\mathbf{e}_1$ in a finite amount of time. In this notation, $\mathbf{e}_1$ stands for an integer index which would be used in the matrix of transmittances in (2.49). Furthermore, Pyke's rule states that $\mathfrak{M}_{i\mathbf{e}_1}(s)/\mathfrak{M}_{i\mathbf{e}_1}(0)$ is the MGF of classification time conditional upon such classification occurring in finite time. At first glance, it would appear these formulas are not well-defined since Pyke's rule (2.49) has a singularity at zero. This singularity

is removable and Butler & Huzurbazar (1999) show its removal is automatic when destination states are made absorbing, as in our setting.

When the SHM contains more than one exemplar, Pyke's rule can still be used. With all exemplars treated as absorbing states, define the classification transmittance from state i to the set of exemplars as

$$\mathfrak{M}_{i\mathbf{E}}(s) = \sum_{j=1}^{q} \mathfrak{M}_{ie_j}(s). \tag{2.50}$$

Now, $\mathfrak{M}_{i\mathbf{E}}(0)$ is the probability of passage from state i to the set of exemplars in a finite amount of time, and $\mathfrak{M}_{i\mathbf{E}}(s)/\mathfrak{M}_{i\mathbf{E}}(0)$ is the MGF of classification time conditional upon such passage occurring in finite time. The quantity $\mathfrak{M}_{i\mathbf{E}}(0)$ is particularly relevant to MGF calculations in the deterministic Hopfield model in which convergence to the set of exemplars is not guaranteed. It can be used to calculate the probability of failure in the deterministic Hopfield classification algorithm.

The concept of lumpability can be extended to transmittance matrices. The lumped transmittance matrix, $\mathbf{T}^{\mathcal{L}}(s)$ is related to $\mathbf{T}(s)$, the unlumped one, through $\mathbf{Q}$ and $\mathbf{R}$ matrices as defined in section 2.3.1, i.e.,

$$\mathbf{T}^{\mathcal{L}}(s) = \mathbf{Q}\mathbf{T}(s)\mathbf{R}. \tag{2.51}$$

Therefore, lumping of the HMC yields a *lumped* transmittance matrix which describes the transient dynamics of the lumped SHM.

Throughout our computations we shall use $\mathbf{T}^{\mathcal{L}}(s)$ in Pyke's rule to calculate exact classification time MGFs. These MGFs are used in saddlepoint approximations, as described in the next section.

2.4.2 Saddlepoint Approximations

Classification time X has MGF determined from Pyke's rule as

$$M_{i\mathbf{E}}(s) = \mathfrak{M}_{i\mathbf{E}}(s)/\mathfrak{M}_{i\mathbf{E}}(0). \tag{2.52}$$

where $\mathfrak{M}_{i\mathbf{E}}(s)$ is determined from (2.50). The mass function of X, $p_k = \Pr\{X = k\}$, is determined by the derivative

$$p_k = \frac{1}{k!}\frac{d^k}{ds^k} M_{i\mathbf{E}}(\ln s)\,|_{s=0}. \tag{2.53}$$

However, the HMC typically has many states, even after lumping, and as a result one cannot hope for simple expressions for $M_{i\mathbf{E}}(s)$. Without such, it becomes difficult to calculate the higher order derivatives of (2.52). It is possible, however, to approximate p_k.

An alternative to exact calculation in (2.53) is the saddlepoint approximation to the Fourier inversion integral. The saddlepoint approximation was first presented in Daniels (1954) and is based upon the cumulant generating function (CGF) $K_{i\mathbf{E}}(s) = \ln M_{i\mathbf{E}}(s)$. If k is in the interior of the convex hull of the support for the classification time distribution and $\hat{s}$ is the unique solution to the saddlepoint equation

$$K'_{i\mathbf{E}}(\hat{s}) = k, \tag{2.54}$$

then

$$\hat{p}_k = \frac{1}{\sqrt{2\pi K''_{i\mathbf{E}}(\hat{s})}} \exp\left\{K_{i\mathbf{E}}(\hat{s}) - \hat{s}k\right\}. \tag{2.55}$$

The smallest possible classification time, $k_{\min}$, is on the boundary of the support and therefore its probability, $p_{\min} = P\{X = k_{\min}\}$, does not have a saddlepoint approximation. In most instances, however, it can be calculated exactly.

In general, $p_{\min} + \sum \hat{p}_k$ is not equal to unity. Therefore, the saddlepoint approximation can be improved, by renormalization, as discussed in Reid (1988);

$$\bar{p}_k = \frac{\hat{p}_k}{p_{\min} + \sum \hat{p}_k}. \tag{2.56}$$

We can also easily obtain an approximation to the discrete CDF of X, for $k \neq E[X] = K'_{i\mathbf{E}}(0)$, as described in Daniels (1987). This approximation differs according to whether k is below or above the mean and is given by

$$\begin{aligned} P\{X \geq k\} &= 1 - \Phi(\hat{w}) - \phi(\hat{w})\left(\tfrac{1}{\hat{w}} - \tfrac{1}{\hat{u}}\right) \quad \text{if} \quad k > E[X] \\ P\{X \leq k\} &= \Phi(\hat{w}) + \phi(\hat{w})\left(\tfrac{1}{\hat{w}} - \tfrac{1}{\hat{u}}\right) \qquad \text{if} \quad k < E[X] \end{aligned}$$

where Φ denotes the standard normal CDF, ϕ the standard normal density,

$$\hat{w} = \text{sgn}\,(\hat{s})\,(2\,\{\hat{s}k - K_{i\mathbf{E}}\,(\hat{s})\})^{-1}$$

and

$$\hat{u} = \begin{cases} (1 - \exp\{-\hat{s}\})\,\sqrt{K''_{i\mathbf{E}}\,(\hat{s})} & \text{if } k > E[X] \\ (\exp\{\hat{s}\} - 1)\,\sqrt{K''_{i\mathbf{E}}\,(\hat{s})} & \text{if } k < E[X] \end{cases}.$$

The accuracy of these saddlepoint approximations is illustrated in the next section.

2.4.3 Example

Consider a SHM with the three exemplars of length ten given as

$$\mathbf{E}^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 & -1 & 1 \end{bmatrix}.$$

With this exemplar matrix, there are four distinct configurations given by the rows of

$$\mathbf{L} = \begin{bmatrix} 1 & -1 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & 1 \end{bmatrix}$$

with repetitions $\mathbf{n} = (7, 1, 1, 1)$. The original HMC has $2^{10} = 1,024$ states and the lumped HMC has 64 megastates.

First, the classification time distribution with $\beta = 0.5$ and $\mathbf{u}_0 = (-1, \ldots, -1)^T$ is studied. The starting vector is Hamming distance two from its nearest exemplary neighbors, $\mathbf{e}_2$ and $\mathbf{e}_3$. Pyke's rule gives the mean and variance of classification time

as 74.75 and 7571.13 respectively. This mean and variance are exact and agree with such calculations from classical Markov chain analysis.

Next, $p_{\min}$ is calculated exactly as

$$p_{\min} = \frac{2}{100}\left[\sigma(-1)\,\sigma(2/5) + \sigma(1/5)\,\sigma(-4/5)\right] = 6.6295 \times 10^{-3}. \tag{2.57}$$

The minimal classification time occurs in two steps: $p_{\min}$ is, therefore, determined as the sum of probabilities over four distinct paths which transform $\mathbf{u}_0$ to $\mathbf{e}_2$ or $\mathbf{e}_3$. The summed contribution from $\mathbf{e}_2$ is the same as that from $\mathbf{e}_3$ with the probability for each given by theorem 1 as one half of (2.57).

The probability of correct classification, without regard to step number, is computed from as 0.84 using classical Markov chain analysis.

Saddlepoint approximations to the PMF and CDF of the classification time distribution are calculated. Figure 12 presents a smoothed, renormalized saddlepoint PMF approximation (dashed) and a kernel density estimate (solid), using a bi-weight kernel and that is based upon one million simulated classification times. The saddlepoint approximation appears to do extremely well throughout the visible support of X and does particularly well in the right tail. Figure 14 gives a smoothed saddlepoint CDF approximation (dashed) and the integrated kernel density estimate. Again, the saddlepoint approximation performs very well throughout the visible support of X.

The dependence on β of mean time to classification and probability of correct classification is studied in our final computations. Such a performance study is important in designing classifiers based upon the SHM since β can be chosen to match some performance criterion. Our criteria are averaged over the set all possible starting vectors using a uniform distribution on $\mathcal{U}$. Such randomization at the component level leads to a megastate distribution, $P\{\mathbf{u}_0 \in \mathcal{L}_\mathbf{k}\} = N_\mathbf{k}/2^n$ for all $\mathcal{L}_\mathbf{k} \in \mathcal{L}$ so that averaging need only be done over megastates and not individual states. Figure 15 presents the mean time to classification for $\beta \in [0, 3]$. As β increases this mean time will also increase to infinity. This is a consequence of the appearance of absorbing states besides the exemplars when β when is infinity. Figure 16 presents the probability

of correct classification for $\beta \in [0,3]$. Numerical experimentation verifies that, as β increases, this probability will increase to 0.66, the probability of correct classification at $\beta = \infty$.

2.5 The Double Laplace Approximation

The HMC's stationary distribution is a regular exponential family in which the energy function is a sufficient statistic. Therefore, the stationary distribution's normalization constant, or partition function, forms the kernel for the MGF of the energy function. Unfortunately, exact calculation of the partition function requires $O(2^n)$ operations. We provide an approximation for it in this section using Laplace's method which leads to an approximation of the energy function MGF and access to saddlepoint approximations.

2.5.1 MGF Approximation

The stationary distribution of the HMC is a one-dimensional regular exponential family, so that the MGF of the energy function is

$$M(s) = Z(\beta - s)/Z(\beta) \tag{2.58}$$

for all real s. In principle, the MGF tells us everything we would want to know about the stationary distribution of the energy function. Unfortunately, there is no closed form expression for the transform in (2.58) and exact enumeration of probabilities may be too cumbersome. It is possible, however, to approximate (2.58) with the method of auxiliary variables as described in Amit (1989).

From (2.39), the exponent in the stationary distribution (2.5) may be written as

$$\exp\left\{\frac{\beta}{2}\mathbf{u}^T\mathbf{W}\mathbf{u}\right\} = \exp\left\{\frac{\beta}{2n}\left(\mathbf{L}^T\mathbf{u}^{\mathcal{L}}\right)^T\left(\mathbf{L}^T\mathbf{u}^{\mathcal{L}}\right)\right\}. \tag{2.59}$$

where $\mathbf{u}^{\mathcal{L}} = \mathbf{n} - 2\mathbf{k}$. It would be trivial to evaluate the sum of all terms of the form (2.59) if the exponent were linear and not quadratic in $\mathbf{u}^{\mathcal{L}}$. The method of auxiliary

variables replaces (2.59) with an integral whose log-integrand is linear in $\mathbf{u}^{\mathcal{C}}$; thus the summation can be performed in closed form inside the integral. This integral is the multivariate Gaussian identity

$$\exp\left\{\frac{\beta}{2n}\left(\mathbf{L}^T\mathbf{u}^{\mathcal{C}}\right)^T\left(\mathbf{L}^T\mathbf{u}^{\mathcal{C}}\right)\right\} = c(\beta)\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left\{-\frac{n\beta}{2}\|\mathbf{x}\|^2+\beta\mathbf{x}^T\mathbf{L}^T\mathbf{u}^{\mathcal{C}}\right\}d\mathbf{x} \tag{2.60}$$

where $c(\beta) = (2\pi/\beta n)^{-\frac{q}{2}}$ and $\mathbf{x} = (x_1, x_2, \ldots, x_q)^T$ is a set of auxiliary variables. Substituting (2.60) into (2.59) and performing the sum over $\mathcal{K}$ allows (2.41) to be expressed as

$$Z(\beta) = c(\beta)\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left\{-\frac{n\beta}{2}\|\mathbf{x}\|^2\right\}\sum_{\mathbf{k}\in\mathcal{K}}N_{\mathbf{k}}\exp\left\{\beta\left(\mathbf{L}\mathbf{x}\right)^T(\mathbf{n}-2\mathbf{k})\right\}d\mathbf{x}. \tag{2.61}$$

This summation may be factored into a product m sums each of which is over a single k_i to give

$$Z(\beta) = c(\beta)\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left\{-\frac{n\beta}{2}\|\mathbf{x}\|^2\right\}\prod_{i=1}^{m}\left(\sum_{k_i=0}^{n_i}\binom{n_i}{k_i}\exp\left\{\beta l_i^T\mathbf{x}\left(n_i-2k_i\right)\right\}\right)d\mathbf{x}.$$

The binomial theorem now allows explicit summation over each k_i to give

$$Z(\beta) = c(\beta)\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left\{-nf_\beta(\mathbf{x})\right\}d\mathbf{x} \tag{2.62}$$

where

$$f_\beta(\mathbf{x}) = \frac{\beta\|\mathbf{x}\|^2}{2} - \sum_{i=1}^{m}\frac{n_i}{n}\left[\beta l_i^T\mathbf{x} + \ln\left(1+\exp\left\{-2\beta l_i^T\mathbf{x}\right\}\right)\right]. \tag{2.63}$$

From this development, we see that lumping has provided us with an integral expression for the partition function whose approximation is now amenable to Laplace's method.

Laplace's method requires that we determine all local minima of $f_\beta(\mathbf{x})$. The approximation in this context is

$$\hat{Z}(\beta) = \left(\frac{2\pi\beta}{n}\right)^{-\frac{q}{2}}\sum_{i=1}^{N(\beta)}\left|f_\beta''\left(\tilde{\mathbf{x}}_\beta^{(i)}\right)\right|^{-\frac{1}{2}}\exp\left\{-nf_\beta\left(\tilde{\mathbf{x}}_\beta^{(i)}\right)\right\}$$

where $\left\{\tilde{\mathbf{x}}_{\beta}^{(i)}, i = 1, \ldots, N(\beta)\right\}$ is the set of local minima of $f_{\beta}(\mathbf{x})$.

For $s < \beta$, an approximation to $M(s)$, the MGF of the energy function at s, is given as a ratio Laplace approximations:

$$\hat{M}(s) = \hat{Z}(\beta - s)/\hat{Z}(\beta). \tag{2.64}$$

This approximation is not defined for $s > \beta$. However, this presents no problem, as the tail probabilities of interest to us concern small saddlepoint values as illustrated in the next section.

2.5.2 Example

The SHM from section 2.4.3 is used here with $\beta = 2.0$. Figure 17 plots the exact CGF (solid) and its double Laplace approximation (dashed). The approximation appears to be reasonable over most of the range of s, however, as s approaches β the approximation loses accuracy. This occurs because, at each minima, the function $\exp\{-nf_{\beta-s}(\mathbf{x})\}$ becomes less peaked and therefore less Gaussian-shaped. The practical consequence of this is that Laplace's method becomes less accurate far into the right tail of $-\frac{1}{2}\mathbf{u}^T\mathbf{W}\mathbf{u}$. Mean and variance approximations may be obtained by numerically differentiating the approximate CGF. This results in values -9.25 and 0.41 which should be compared to the true mean and variance, -9.26 and 0.32.

The accuracy in using (2.64) for saddlepoint approximations is illustrated in Table 3 which lists exact probabilities $\left(P\{-\frac{1}{2}\mathbf{u}^T\mathbf{W}\mathbf{u} = k\}\right)$, saddlepoint mass function approximations $\left(\hat{P}\{-\frac{1}{2}\mathbf{u}^T\mathbf{W}\mathbf{u} = k\}\right)$, and the associated saddlepoints, $\hat{s}$, for each value of the energy function. The lowest and highest energy values do not have saddlepoint approximations since they are on the boundary of the support. Greater accuracy is seen in the left tail which contains the lower energy values.

The accuracy of the Lugannani-Rice distribution approximation is displayed in Table 4. Again, the saddlepoint approximation is most accurate in the left tail. Left tail probabilities are of interest for the following reason. The determination of the

weight matrix from Hebb's rule in (2.2), i.e.,

$$\mathbf{W} = n^{-1}\mathbf{E}\mathbf{E}^T$$

was originally developed with the hope of ensuring that all exemplars have minimum energy function values, as described in Hertz et al. (1991). Therefore, a left tail probability of the energy function, at an exemplar energy level, can describe to what extent this criterion is met. This computation describes the chance of finding an energy level lower than that of an exemplar when a SHM is run for a very long time and exemplars in this SHM are not treated as absorbing states. In effect, left tail probabilities provide an informal indicator of the extent to which Hebb's rule assigns exemplars low energy. For instance, if exemplar $\mathbf{e}_i$ has very low energy, then the associated left tail probability will be very small.

In our example, the energy function values of the three exemplars are displayed in the table below:

$\mathbf{u}$	$-\frac{1}{2}\mathbf{u}^T\mathbf{W}\mathbf{u}$
$\mathbf{e}_1$	-8.6
$\mathbf{e}_2$	-7.2
$\mathbf{e}_3$	-7.2

While none of these have the smallest possible energy, they all have quite low energy values. If we were unable to enumerate the energy function values, the saddlepoint approximations Table 2, could be used to indicate the existence of smaller energy values.

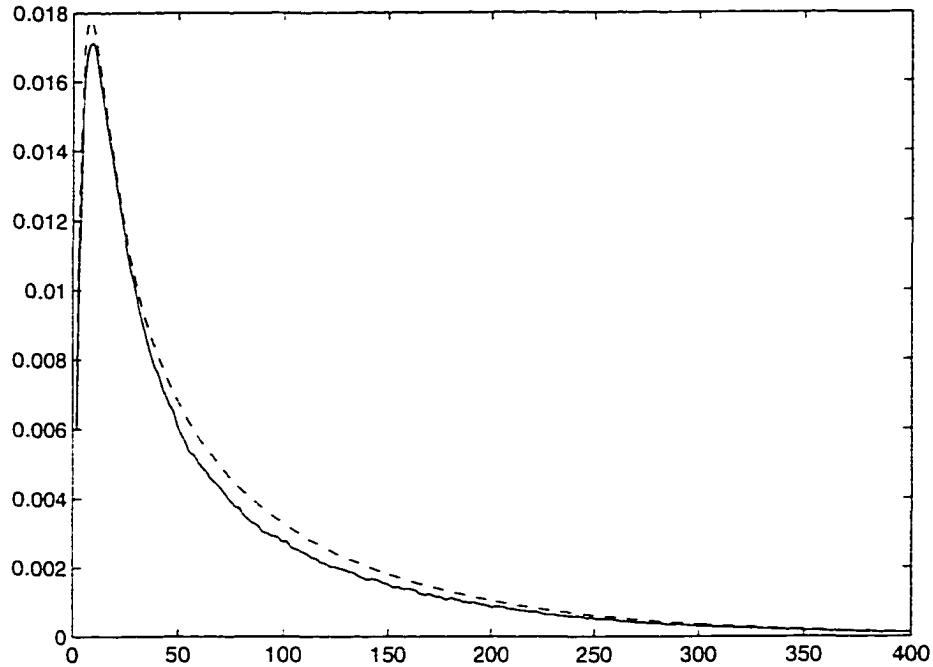

Fig. 12. The smoothed PMF from simulation (solid) and the saddlepoint approximation (dashed).

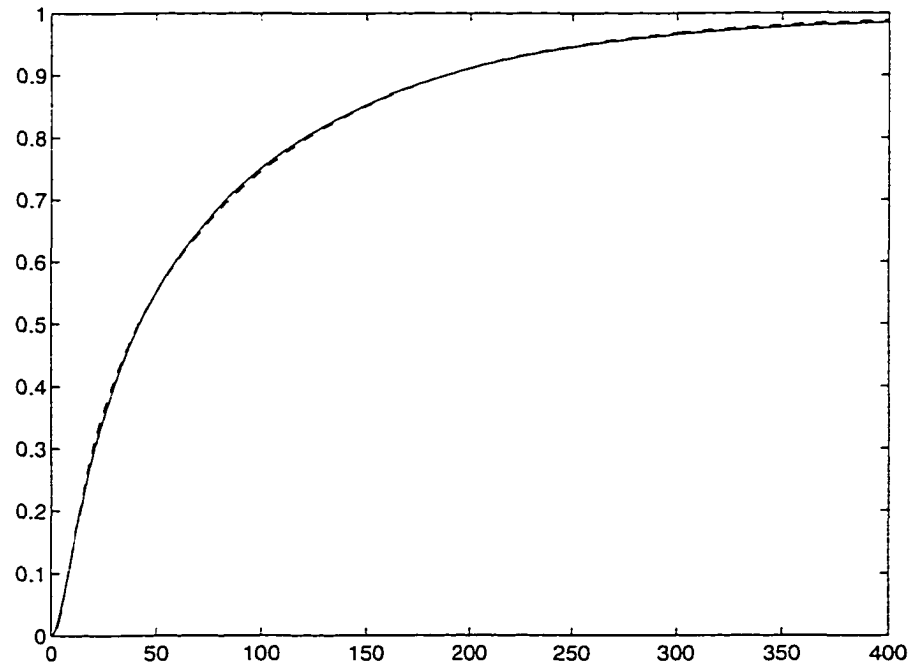

Fig. 13. The smoothed CDF from simulation (solid) and the saddlepoint approximation (dashed).

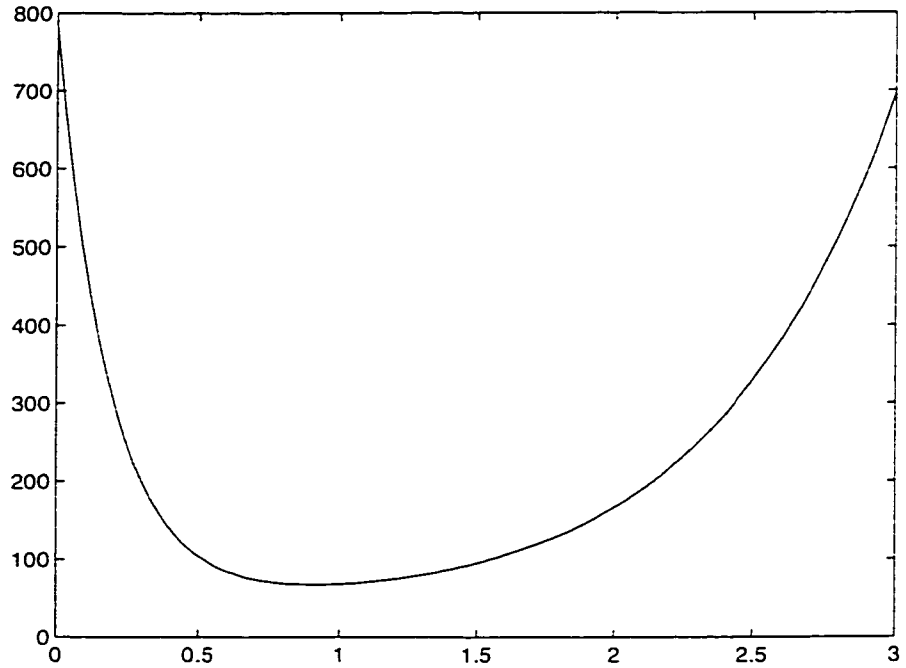

Fig. 14. The Mean time to classification for $\beta \in [0, 3]$.

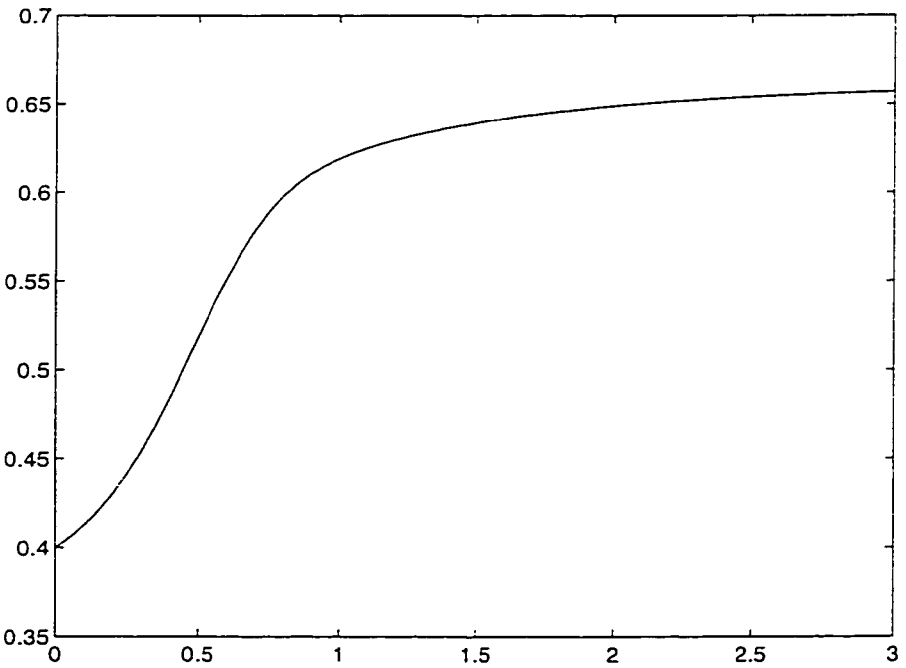

Fig. 15. The probability of correct classification for $\beta \in [0, 3]$.

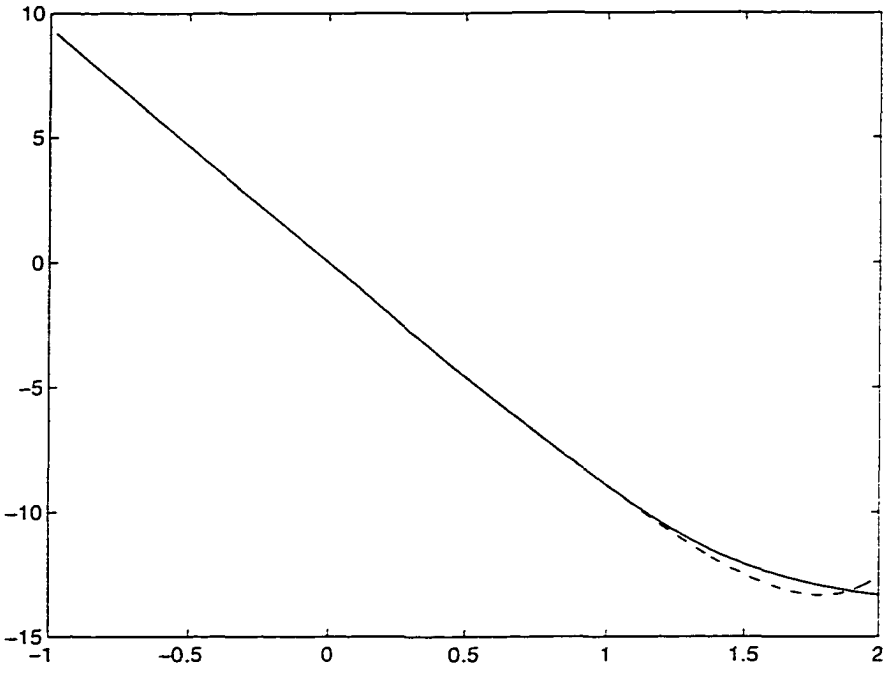

Fig. 16. The CGF from exact enumeration using lumping (solid) and Laplace's approximation (dashed).

Table 3. The value of the energy function, it's exact probability, the associated saddlepoint approximation, and the saddlepoint.

k	$P\{-\frac{1}{2}\mathbf{u}^T\mathbf{W}\mathbf{u} = k\}$	$\hat{P}\{-\frac{1}{2}\mathbf{u}^T\mathbf{W}\mathbf{u} = k\}$	$\hat{s}$
0.0	1.1195×10^{-7}	*	*
−0.6	1.7097×10^{-6}	3.8372×10^{-7}	0.8802
−0.8	1.6634×10^{-6}	5.5309×10^{-7}	0.8760
−1.6	4.9435×10^{-6}	2.3438×10^{-6}	0.8582
−2.2	1.6413×10^{-5}	6.8072×10^{-6}	0.8433
−2.4	1.0883×10^{-5}	9.6536×10^{-6}	0.8379
−3.8	1.3421×10^{-4}	1.0451×10^{-4}	0.7971
−4.8	9.9173×10^{-4}	5.3633×10^{-4}	0.7625
−5.4	1.2544×10^{-3}	1.3956×10^{-3}	0.7389
−7.2	1.7215×10^{-2}	2.3210×10^{-2}	0.6397
−8.6	0.2831	0.2228	0.4332
−9.6	0.6973	*	*

Table 4. The value of the energy function, it's exact right tail value and associated saddlepoint approximation.

k	$P\{-\frac{1}{2}\mathbf{u}^T\mathbf{W}\mathbf{u} > k\}$	$\hat{P}\{-\frac{1}{2}\mathbf{u}^T\mathbf{W}\mathbf{u} > k\}$
0.0	0.0	*
−0.6	1.1195×10^{-7}	6.5223×10^{-8}
−0.8	1.8217×10^{-6}	9.4892×10^{-8}
−1.6	3.4851×10^{-6}	4.2365×10^{-7}
−2.2	8.4286×10^{-6}	1.2381×10^{-6}
−2.4	2.4842×10^{-5}	1.7754×10^{-6}
−3.8	3.5725×10^{-5}	1.9694×10^{-5}
−4.8	1.6993×10^{-4}	1.0385×10^{-4}
−5.4	1.1617×10^{-3}	2.7013×10^{-4}
−7.2	2.4161×10^{-3}	3.9916×10^{-3}
−8.6	1.9631×10^{-2}	3.0532×10^{-2}
−9.6	0.3027	*

References

[1] Amit, D.J. (1989). *Modeling Brain Function.* Cambridge: Cambridge University Press.

[2] Amit, D.J., Gutfreund, H., & Sompolinsky, H. (1985). Spin-glass models for neural networks, *Physical Review A*, **32,** 1007-1018.

[3] Bates, M.B. and Watts, D.G. (1988). *Nonlinear Regression Analysis and its Applications.* New York: Wiley.

[4] Berger, J.O and Pericchi, L.R (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association.*, **91**, 109-122.

[5] Bishop, C.M. (1995). *Neural Networks for Pattern Recognition.* Oxford : Clarendon Press.

[6] Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis.* Reading MA: Addision-Wesley.

[7] Buntine, W.L. and Weigend, A.S. (1991). Bayesian back-propagation. *Complex Systems*, **5**, 603-643.

[8] Butler, R.W. & Huzurbazar, A.V (1999). Saddlepoint Methods for Bayesian Prediction in Applied Probability Models. *Canadian Journal of Statistics.* to appear.

[9] Cybenko, G. (1988). Continuous valued neural networks with two hidden layers are sufficient. Technical Report, Dept. of Computer Science, Tufts University.

[10] Daniels, H. (1954). Saddlepoint approximations in statistics. *Annals of Mathematical Statistics*, **25**, 631-650.

[11] Daniels, H. (1987). Tail probability approximations. *International Statistical Review*, **55**, 37-48.

[12] Davison, A.C.(1986). Approximate predictive likelihood. *Biometrika*, **73**, 323-332.

[13] DiCiccio, T, Kass, R, Raftery, A, and Wasserman (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, **92**, 903-915.

[14] Faraway, J. and Chatfield, C. (1998). Time series forecasting with neural networks: a comparative study using the airline data. *Applied Statistics*, **47**, 231-250.

[15] Feller, W. (1968). *An Introduction to Probability Theory and its Applications.* New York: Wiley.

[16] Geman, S. & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 720-741.

[17] Golub, G.H. and Pereyra, V. (1973). The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate, *Siam Journal on Numerical Analysis*, **10**, 413-432.

[18] Hertz, J.A., Krogh, A.S., & Palmer, R.G (1991). *Introduction to the Theory of Neural Computation.* New York: Addison-Wesely.

[19] Hoffman, R. E. (1992). Attractor neural networks and psychotic disorders. *Psychiatric Annals*, **22**, 119-124.

[20] Hrycej, T. (1992). *Modular Learning in Neural Networks. A Modularized Approach to Neural Network Classification.* New York: Wiley.

[21] Kam, M.& Cheng, R. (1989). Decision-making with the Boltzmann Machine. Proceedings of the American Control Conference. Vol. 1, pp. 902 - 907, Pittsburgh, PA, June 1989.

[22] Kass, R and Raferty, A. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773-795.

[23] Kemeny, J. & Snell, L. (1969). *Finite Markov Chains.* Princeton, N.J. : Van Nostrand.

[24] Kindermann, R. & Snell, L. (1980). *Markov Random Fields and Their Applications.* Contemporary Mathematics, vol. 1. Providence: American Mathemaical Society.

[25] Kleinfield, D. & Somopolinsky, H. (1989). Associative Neural Models for Central Pattern Generators. In *Methods in Neuronal Modeling: From Synapses to Networks,* eds. C. Koch & I. Segev, 195-246. Cambridge: MIT Press.

[26] Lee, T.H., White, H., and Granger, C. (1993). Testing for neglected nonlinearity in time series models. A comparison of neural network methods and alternative tests. *Journal of Econometrics*, **56**, 269-290.

[27] Leonard, T. (1982). Comment on "A simple predictive distribution function". *Journal of the American Statistical Association*, **77**, 657-87.

[28] Lugananni, R. & Rice, S. (1980). Saddlepoint approximations for the distribution of the sum of independent random variables. *Advances in Applied Probability*, **12**, 475-490.

[29] Mackay, D.J.C (1992). A practical Bayesian framework for backpropagation networks. *Neural Computation*, **4**, 448-472.

[30] Müller, B., Reinhardt, J. and Strickland, M.T. (1995). *Neural Networks : an introduction*. Heidelberg : Springer-Verlag.

[31] Neal, R. (1993). Bayesian learning via stochastic dynamics. NIPS5, 475-482.

[32] Neal, R. (1995). Bayesian Learning for Neural Networks, Ph.D.thesis, Department of Computer Science, University of Toronto.

[33] Pyke, R. (1961). Markov renewal processes with finitely many states. *Annals of Mathematical Statistics*, **32**, 1243-1259.

[34] Reid, N. (1988). Saddlepoint Methods and statistical inference. *Statistical Science*, **3**, 213-238.

[35] Ripley, B.D. (1987). *Stochastic Simulation*. New York: Wiley.

[36] Ripley, B.D. (1994*a*). Flexible non-linear approaches to classification, In *From Statistics to Neural Networks. Theory and Pattern Recognition Applications.*,1994, eds. Cherkassy, V., Friedman, J.H. & Wechsler, H., 105-126, Berlin: Springer.

[37] Ripley, B.D. (1994*b*). Neural networks and related methods for classification (with discussion). *Journal of the Royal Statistical Society, Series B*, **56**, 409-456.

[38] Ripley, B.D. (1995). Statistical ideas for selecting neural network architectures. In *Neural Networks: Artificial Intelligence and Industrial Applications*, eds. B. Kappen and S. Gielen. London: Springer.

[39] Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

[40] Seber, G.A.F and Wild, C.J. (1989). *Nonlinear Regression*. New York: Wiley.

[41] Shepanski, J.F. (1987). Fast learning in artificial neural systems: multilayer perceptron training using optimal estimation. In *Proceedings of IEEE First International Conference on Neural Networks*, San Diego, 1987, eds. M. Caudill and C. Butler I. 465 – 472. Long Beach, CA: IEEE Press.

[42] Sumita, U. & Rieders, M. (1988). First Passage Times and Lumpability of Semi-Markov Processes. *Journal of Applied Probability*, **25**, 675-687.

[43] Whittle, P. (1991). Neural Nets and Implicit Inference. *Annals of Applied Probability*, **1**, 173-188.