LETTER Communicated by Shun-ichi Amari and Todd Leen

Nonmonotonic Generalization Bias of Gaussian Mixture Models

Shotaro Akaho
Electrotechnical Laboratory, Information Science Division, Ibaraki 305-8568, Japan

Hilbert J. Kappen
RWCP Theoretical Foundation SNN, Department of Medical Physics and Biophysics,
University of Nijmegen, NL 6525 EZ Nijmegen, The Netherlands

Theories of learning and generalization hold that the generalization bias, defined as the difference between the training error and the generalization error, increases on average with the number of adaptive parameters. This article, however, shows that this general tendency is violated for a gaussian mixture model. For temperatures just below the first symmetry breaking point, the effective number of adaptive parameters increases and the generalization bias decreases. We compute the dependence of the neural information criterion on temperature around the symmetry breaking. Our results are confirmed by numerical cross-validation experiments.

1 Introduction

An important problem for learning is to optimize the model such that the best generalization performance is obtained. Generalization performance depends on the number of adaptive parameters in the model. We can reduce the training error as much as needed by increasing the number of adaptive parameters. However, the performance of the model should be evaluated by the generalization error, which measures the error for test samples. The best generalization performance is usually not obtained by maximizing the number of adaptive parameters (Vapnik, 1984; Rissanen, 1986).

The generalization bias is defined as the difference between the generalization error and the training error. The expected generalization bias can be measured approximately by the neural information criterion (NIC) or numerically through cross-validation on test data (Moody, 1992; Amari, 1993; Murata, Yoshizawa, & Amari, 1991, 1994). Typically, if the number of model parameters increases, the training error decreases because a better fit can be obtained. At the same time, the generalization bias is expected to increase due to the larger model variability. In this article we present an example where this intuition is violated: the case that generalization bias *decreases* when at the same time the number of model parameters increases.

We consider radial basis Boltzmann machines (RBBM), a special class of gaussian mixture models (Kappen, 1995). Gaussian mixture models have attracted a lot of attention in the neural network community (Barkai, Seung, & Sompolinsky, 1993; Titterington, 1985) because they are closely related to several neural network models such as radial basis function (RBF) networks (Poggio & Girosi, 1990) and hierarchical mixture of experts (HME) networks (Jacobs & Jordan, 1991; Jordan & Jacobs, 1994).

For gaussian mixtures, complexity and generalization performance are controlled by the number of mixture components. In RBBMs, the complexity of the model is controlled by a continuous parameter β. When β is small, the maximum likelihood (ML) solution of the RBBM degenerates into one gaussian. At a critical value of β, the ML solution becomes a mixture of several gaussians. This phenomenon of symmetry breaking is repeated recursively for increasing β. Thus, β controls the effective number of mixture components and, in this way, the generalizaton performance. In this article we study the generalization bias for RBBMs around the first symmetry breaking point.

In section 2, we define RBBMs and show the symmetry breaking mechanism. In section 3, we derive the condition when the symmetry breaking is two-way or h-way, where h is the number of mixture components in the RBBM. In section 4, we show an analytical result that the generalization bias of the model, as measured by the NIC, decreases if the symmetry breaking is two-way, even though the superficial effective number of parameters increases. If we can assume that training error does not change significantly around the symmetry breaking point, this anomaly implies that the ML solution just below the critical temperature is expected to realize a smaller generalization error than just above the critical temperature. In section 5, we show that this effect is confirmed numerically.

2 Radial Basis Boltzmann Machines

Let us consider a gaussian mixture model with equal priors in which the variance of all components is spherically symmetric and identical:

$$p(\boldsymbol{x} \mid W; \beta) = \frac{1}{h} \sum_{i=1}^{h} \sqrt{\frac{\beta}{\pi}} \exp(-\beta \, \|\boldsymbol{x} - \boldsymbol{w}_i\|^2). \tag{2.1}$$

W denotes the set of adaptive parameters $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_h\}$. The complexity of the model is determined by h, the number of gaussian components, and β, a control parameter called the inverse temperature in physics. Statistically β is equal to half the inverse variance. We chose spherical covariance matrices, since we are interested in the relation between complexity and symmetry breaking. We expect that our results can be generalized to mixture models with full covariance matrices.

This unsupervised model was originally proposed by Rose, Gurewitz, and Fox (1990) and generalized by Kappen (1993, 1995; Nijman & Kappen, 1997) to the supervised case. The model is related to fuzzy clustering as well (Bezdek, 1980). In this article, we refer to model 2.1 as a RBBM.

Equation 2.1 can be written as

$$p(\boldsymbol{x}) = \sum_{i=1}^{h} p(\boldsymbol{x}, i) = \sum_{i=1}^{h} p(\boldsymbol{x}|i)p(i),$$

with i labeling the individual clusters and $p(i) = 1/h$ the prior probability of each cluster. $p(\boldsymbol{x}|i)$ is simply a gaussian distribution in $\boldsymbol{x}$ centered on $\boldsymbol{w}_i$. For fixed β, the ML solution for the cluster means is easily derived and is given by Rose et al. (1990),

$$\boldsymbol{w}_i = \frac{\langle \boldsymbol{x} p(i|\boldsymbol{x})\rangle}{\langle p(i|\boldsymbol{x})\rangle},$$

where $\langle \cdot \rangle$ denotes expectation values with respect to $q(\boldsymbol{x})$,

$$\langle \cdot \rangle \equiv \int \cdot\, q(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x},$$

and $q(\boldsymbol{x})$ is the target distribution from which the training samples and test samples are generated. These coupled equations can be solved by a variety of methods such as the expectation maximization (EM) algorithm.

An example of the ML solutions as a function of the temperature is shown in Figure 1. The training data are generated with equal probability from two distributions: $u[0.5, 1.5]$ and $N[-1, 0.3^2]$, where $u[a, b]$ is the uniform distribution on $[a, b]$ and $N[\mu, v]$ is the gaussian distribution with mean μ and variance v. The number of training samples is 100, and $h = 100$.

For small β, the ML solution is of the form $\boldsymbol{w}_1 = \boldsymbol{w}_2 = \cdots = \boldsymbol{w}_h$, that is, the solution corresponds to one cluster. At a critical value of β, the cluster splits into smaller parts. These symmetry breakings reoccur recursively at higher values of β. Therefore, the number of gaussian components can be controlled by adjusting the temperature in this model. So at any β, although the total number of kernels is h, only a smaller effective number of kernels is used. For this reason, we assume h to be sufficiently large. The complexity is then controlled by β only.

3 Symmetry Breaking Point of the RBBM

Since the effective number of parameters changes at the symmetry breaking points (SBPs), we study the symmetry breaking process in detail. Although the behavior of symmetry breaking is complicated to analyze, we have some results on the first SBP, the highest temperature at which the phase transition occurs. These results are expected to be applicable for the other SBPs qualitatively.

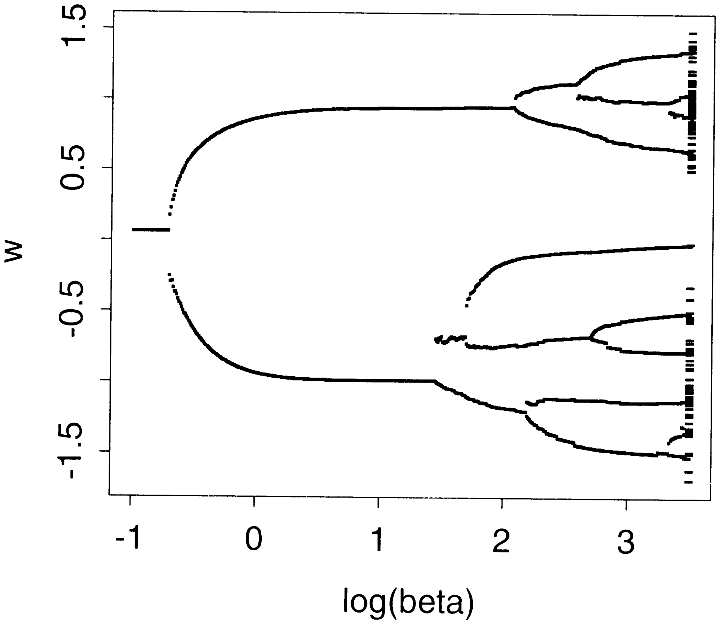

Figure 1: An example of the ML solution and the symmetry breaking phenomenon. Horizontal axis: $\log(\beta)$; vertical axis: w or x. Dots show the ML solution for each temperature; dashes on the right show the training samples.

When the temperature is higher than $1/\beta_c$, all the gaussian components are degenerated into one gaussian, and the ML solution is given by $\boldsymbol{w}_i = \langle \boldsymbol{x} \rangle$. One can compute the first SBP analytically as

$$\beta = \beta_c \equiv \frac{1}{2\lambda_1}, \tag{3.1}$$

where λ_1 is the maximal eigenvalue of the covariance matrix of $\boldsymbol{x}$. This means that the first symmetry breaking occurs when β is equal to the variance of samples along the first principal component axis (Rose et al., 1990).

This result is considered to be applicable to other SBPs as follows. After breaking, the two clusters drift apart as a function of β. If the distance between the clusters is sufficiently large, each data point contributes to the covariance matrix of the closest cluster only. The eigenvalues of this reduced matrix determine the next critical temperature, as in Equation 3.1.

3.1 Below the First SBP. We can characterize the behavior of the symmetry breaking under the following assumption:

Assumption. *The target distribution $q(x)$ is defined on $\mathcal{R}$ (one dimension) and is assumed to be symmetric. The number of gaussian components h in the RBBM is taken to be even.*

Let $\sigma^2 \equiv \langle (x - \langle x \rangle)^2 \rangle$ and $s_4 \equiv \langle (x - \langle x \rangle)^4 \rangle$ denote, respectively, the second- and fourth-order moment of the distribution $q(x)$. The fourth-order cumulant is defined as $\kappa_4 \equiv s_4 - 3(\sigma^2)^2$ (κ_4/s_4 is called the kurtosis in statistics).

Proposition 1. *Under the assumption, the behavior of the first symmetry breaking can be classified into the following two cases:*

1. *If $\kappa_4 \neq 0$ the symmetry breaking is two-way: the components split into two clusters. The relation between the ML solution w_i and β in the neighborhood of the first SBP β_c is given by*

$$\Delta\beta \simeq \frac{s_4}{6(\sigma^2)^4}(\Delta w_i)^2, \tag{3.2}$$

   where $\Delta\beta = \beta - \beta_c > 0$, $\Delta w_i = w_i - \langle x \rangle$.

2. *If $\kappa_4 = 0$ the symmetry breaking is h-way: the individual components Δw_i are underdetermined up to the third-order approximation that we computed. This suggests that no preferred breaking pattern exists, although fourth-order corrections may somewhat limit this freedom. The relation between the ML solution w_i and β in the neighborhood of the first SBP β_c is written as*

$$\Delta\beta \simeq \frac{1}{2(\sigma^2)^2}\sigma_w^2, \tag{3.3}$$

   where $\sigma_w^2 = \frac{1}{h}\sum_i \Delta {w_i}^2$, $\Delta w_i = w_i - \langle x \rangle$.

κ_4 is equal to zero when $q(x)$ is gaussian; hence the condition represents the similarity between $q(x)$ and gaussian in the sense of the fourth-order cumulant. An outline of the proof of proposition 1 is given in appendix A.

This result can be intuitively understood as follows. When β increases toward the critical point, the stability of the symmetric solution $w_i = \langle x \rangle$ decreases. The stability is measured by the eigenvalues of the Hessian, which is the matrix of second derivatives of the likelihood. At $\beta = \beta_c$ some eigenvalues of the Hessian become zero. Thus, the symmetric solution becomes unstable. When $\kappa_4 \neq 0$, only one eigenvalue becomes zero. The corresponding eigenvector is in the direction of symmetry breaking, and the breaking is twofold: one cluster moves in the positive eigendirection and one in the negative eigendirection. When $\kappa_4 = 0$, however, all eigenvalues become zero simultaneously, and thus all directions become unstable. The breaking is h-fold, because the initial movement of each cluster center can be in any arbitrary direction.

Equation 3.3 is interpreted as follows: Suppose that we have an infinite number of kernels. The model $p(x \mid w) = (1/h)\sum_{i=1}^{h}\phi(x - w_i, \beta)$, with $\phi(x, \beta)$ a gaussian density function with mean 0 and variance $1/2\beta$, can be written as

$$p(x \mid w) = \int p(w)\phi(x - w, \beta)\,\mathrm{d}w,$$

with $p(w)$ some distribution over the kernel means. As an example of $\kappa_4 = 0$, let us consider the case that the target distribution $q(x)$ is gaussian. Therefore, at large β, $p(x \mid w)$ will approach a gaussian distribution. Since $p(x \mid w)$ is the convolution of $p(w)$ with a gaussian distribution, it follows that $p(w)$ for large β will also approach a gaussian distribution.

4 Nonmonotonic Generalization Bias

Since for the RBBM the effective number of adaptive parameters changes as a function of temperature, we must find the temperature that gives the best generalization. Overfitting results in a suboptimal effective number of components and a suboptimal clustering.

Although the generalization bias is a statistical quantity and unknown in general, we can estimate it through cross-validation. Alternatively, we can approximate the mean generalization bias, which is known as the NIC (Murata et al., 1991, 1994) or the effective number of parameters (Moody, 1992). Using this value, we can select the model that minimizes the sum of the training likelihood and the mean generalization bias. Another well-known generalization bias is Akaike's information criterion (AIC), which is given by the number of independent parameters. However, AIC assumes that the target distribution is an RBBM, which is a poor assumption around the SBPs.

4.1 Neural Information Criterion. Given a set of P training samples $X^{(P)} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_P\}$ from a target probability distribution $q(\boldsymbol{x})$, the ML solution $W = W^{(P)}$ maximizes the empirical likelihood over the training samples,

$$R_{\mathrm{emp}}^{(P)}(W; \beta) \equiv \frac{1}{P}\sum_{i=1}^{P}\log p(\boldsymbol{x}_i \mid W; \beta). \tag{4.1}$$

The true likelihood is defined as the expectation value of the likelihood over the target distribution $q(\boldsymbol{x})$:

$$R_{\mathrm{exp}}(W; \beta) \equiv \langle \log p(\boldsymbol{x} \mid W; \beta)\rangle. \tag{4.2}$$

The generalization bias can be approximated by making a Taylor expansion of $R_{\mathrm{emp}}^{(P)}(W; \beta)$ around $W^{(P)}$ and the fact that the asymptotic distribution of

$W^{(P)}$ is locally a gaussian distribution (Murata et al., 1991, 1994). Consequently, when P is large enough, the mean generalization bias is asymptotically given by

$$\left\langle R_{\mathrm{emp}}^{(P)}(W^{(P)};\beta)\right\rangle - \left\langle R_{\exp}(W^{(P)};\beta)\right\rangle \simeq \frac{h_{\mathrm{NIC}}}{P}\ , \tag{4.3}$$

where the average is taken over all training sets of size P, $X^{(P)}$. h_{NIC} is the NIC, defined by

$$h_{\mathrm{NIC}}(\beta) \equiv \mathrm{Tr}[H(W^*)^{-1}D(W^*)], \tag{4.4}$$

where W^* denotes the ML solution of the true likelihood. $H(W)$ and $D(W)$ are the following matrices,

$$H_{ij}(W) \equiv -\langle \partial_i\partial_j \log p(\boldsymbol{x}; W, \beta)\rangle, \tag{4.5}$$

$$D_{ij}(W) \equiv \langle \partial_i \log p(\boldsymbol{x} \mid W, \beta)\, \partial_j \log p(\boldsymbol{x} \mid W, \beta)\rangle, \tag{4.6}$$

where $\partial_i \equiv \frac{\partial}{\partial w_i}$. If $q(\boldsymbol{x})$ belongs to the model set, NIC is equal to AIC because $H(W^*) = D(W^*)$.

In practice, when we apply the NIC in selecting a model, we need to evaluate the bias by using only one training set. In this case, the bias is given by

$$R_{\mathrm{emp}}^{(P)}(W^{(P)};\beta) - R_{\exp}(W^{(P)};\beta) \simeq \frac{h_{\mathrm{NIC}}}{P} + \frac{U}{\sqrt{P}}\ , \tag{4.7}$$

where $U = \sqrt{P}\{R_{\mathrm{emp}}^{(P)}(W^*;\beta) - R_{\exp}(W^*;\beta)\}$ is a random variable of order 1 with zero mean. It can be shown that U is the same for all the models within a nested set of models (Murata et al., 1994), which holds for the RBBM. Therefore, although the $U/\sqrt{P}$ term dominates the NIC term, it does not affect the model selection.

In the following sections, we compute the behavior of the NIC near the SBP for the RBBM. The effective number of adjustable parameters in the RBBM is constant between symmetry breakings and increases stepwise for increasing β. Since the NIC measures the complexity of the model, it is expected to increase as β increases. Our analysis indeed shows a linear increase of NIC with β just before the symmetry breaking ($\beta < \beta_c$). However, just after the symmetry breaking ($\beta > \beta_c$), our analysis shows a decrease of the NIC depending on κ_4.

4.2 Above the First SBP.

When there is only one cluster ($\beta < \beta_c$), we can compute h_{NIC} explicitly. The following proposition shows that the generalization bias increases linearly in proportion to β:

Proposition 2. *If $\beta < \beta_c$, NIC is given by*

$$h_{\mathrm{NIC}}(\beta) = 2\beta \mathrm{Tr}[V_{\boldsymbol{x}}], \tag{4.8}$$

where $V_{\boldsymbol{x}}$ is the covariance matrix of $q(\boldsymbol{x})$.

An outline of the proof of proposition 2 is given in appendix B. Intuitively, the NIC measures the sensitivity of the likelihood against sample fluctuation. Proposition 2 explains this intuition because for small β, one has broad kernels whose centers are less sensitive to sample fluctuations.

4.3 Below the First SBP.

Because the situation below the critical temperature is very complicated, we analyze the NIC under the same assumption as in section 3.1:

Proposition 3. *Under the assumption and also if $\kappa_4 \neq 0$ and $s_4 \neq (\sigma^2)^2$,*

$$\lim_{\beta \downarrow \beta_c} \frac{\partial}{\partial \beta} h_{\mathrm{NIC}}(\beta) = -\infty. \tag{4.9}$$

The condition $s_4 \neq (\sigma^2)^2$ applies to all distributions except for a mixture distribution of 2 δ-functions. If $q(x) = (\delta(x-1) + \delta(x+1))/2$ one obtains $\partial h_{\mathrm{NIC}}(\beta_c)/\partial\beta = -4$.

An outline of the proof of proposition 3 is given in appendix C. Proposition 3 states that the NIC of the RBBM decreases even if the effective number of parameters increases when the symmetry breaking in the first SBP is two-way.

Since the training error is approximately constant around the SBP, we conclude that this model gives slightly better generalization error (in terms of NIC) just below the critical temperature than at or just above the critical temperature. Whether this anomalous behavior affects the optimal value of β depends on the functional dependence of the training error on β. Although this effect causes a local minimum in the generalization error as a function of β around the SBP, its global minimum might be attained at much higher or lower values of β.

It is not easy to analyze the case $\kappa_4 = 0$, since the symmetry breaking is h-way and the ML solution is underdetermined by the third-order approximation as shown in proposition 1.

5 Experiments

In this section, we show some computer simulation results for the two cases with different fourth-order cumulants presented in proposition 1. In both cases, the target distribution $q(x)$ is created so that the mean is 0.0 and the variance is 1.0. Therefore, $\log \beta_c \simeq -0.693$, and the ML solution for training

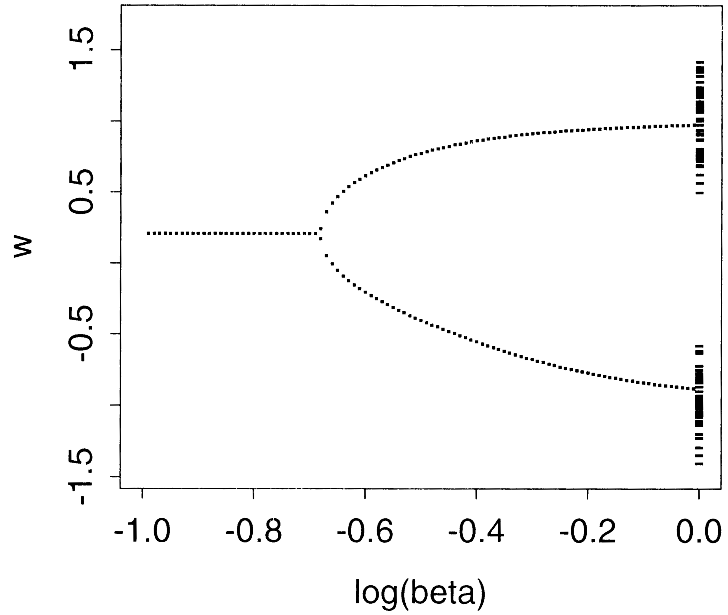

Figure 2: ML solution versus log(β). Data are generated from a distribution with $\kappa_4 \neq 0$. Horizontal axis: log(β); vertical axis: w or x. Dots show the ML solution for each temperature; dashes on the right show the training samples.

samples breaks around this value. Both the number of training samples (P) and the number of gaussian components (h) is set to 100. We use the EM algorithm (Dempster, Laird, & Rubin, 1977) to optimize the empirical likelihood. We initialize the EM algorithm with one gaussian component on each of the training samples. The test set consists of 100,000 samples, which are generated from the same distribution as training samples.

5.1 The Case of $\kappa_4 \neq 0$. The target distribution is a mixture of two gaussians,

$$q(x) = \frac{1}{2}\sqrt{\frac{C_1}{\pi}}\left[\exp\left\{-C_1(x - C_2)^2\right\} + \exp\left\{-C_1(x + C_2)^2\right\}\right],$$

where $C_1 = 12.5$, $C_2 = \sqrt{0.96}$. C_1 and C_2 are chosen such that the variance of $q(x)$ is 1.0. An example of the ML solution as a function of the temperature around the first SBP is shown in Figure 2. The approximate generalization bias, as given by the NIC, as a function of temperature is shown in Figure 3. Note that the vertical tangent at the breaking point is in agreement with proposition 3.

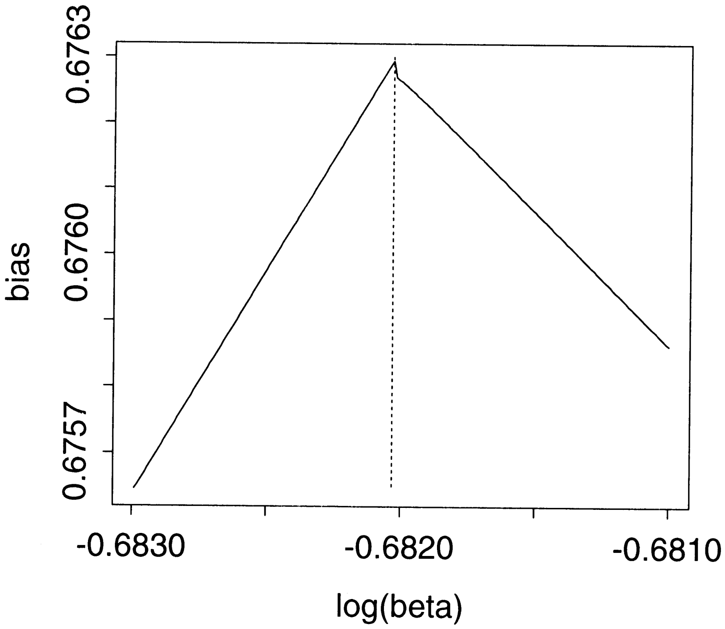

Figure 3: NIC as a function of $\log(\beta)$ (magnified around the symmetry breaking point). Data are generated from a distribution with $\kappa_4 \neq 0$. Horizontal axis: $\log(\beta)$; vertical axis: NIC multiplied by the number of training samples; dashed line: β_c

The generalization bias averaged over 50 experiments with different training sets is shown in Figure 4. It differs from Figure 3 in two ways. It does not show the vertical tangent as in the individual runs because the vertical tangent appears only very close to the symmetry breaking point and the symmetry breaking point fluctuates for different training sets. Second, the term U in equation 4.7 is rather different for different training sets, giving a smoothing effect on the average generalization bias.

5.2 The Case of $\kappa_4 = 0$. The target distribution is one gaussian with a unit variance,

$$q(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2).$$

In this case, the convergence of the EM algorithm is much more unstable than in the case of the previous section, because of the reasons mentioned in proposition 1. A typical ML solution as a function of temperature is shown in Figure 5.

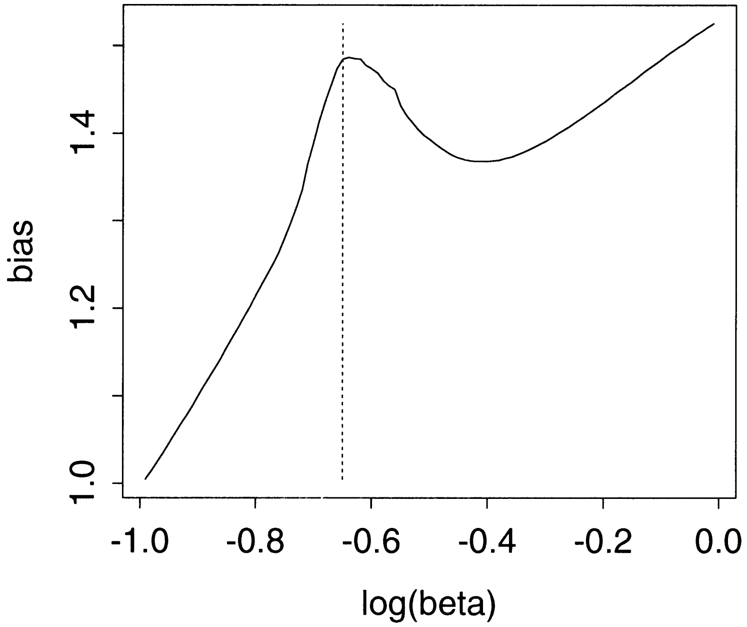

Figure 4: Generalization bias from cross-validation, averaged over 50 experiments, as a function of log(β). Data are generated from a distribution with $\kappa_4 \neq 0$. Horizontal axis: log(β); vertical axis: generalization bias multiplied by the number of training samples; dashed line: average of empirical value of β_c's.

The generalization bias averaged over 50 experiments with different random numbers are shown in Figure 6. We did not observe a single instance where the likelihood is maximal around the first SBP as in the case of $\kappa_4 \neq 0$.

6 Conclusion

We have shown the nonmonotonical behavior of the generalization bias of a special class of gaussian mixture models, called the radial basis Boltzmann machines. For high temperature (large variance in the gaussian kernels), the generalization error increases linearly with β. On the other hand, below the critical temperature, the symmetry breaking phenomenon depends critically on the value of the fourth cumulant κ_4.

If $\kappa_4 \neq 0$, the generalization bias decreases with β below the critical temperature. This means that the NIC decreases during the symmetry breaking process. After symmetry breaking is complete, NIC increases again. While NIC decreases, the effective number of adaptive parameters increases because the kernels split up. It is normally assumed that NIC measures the

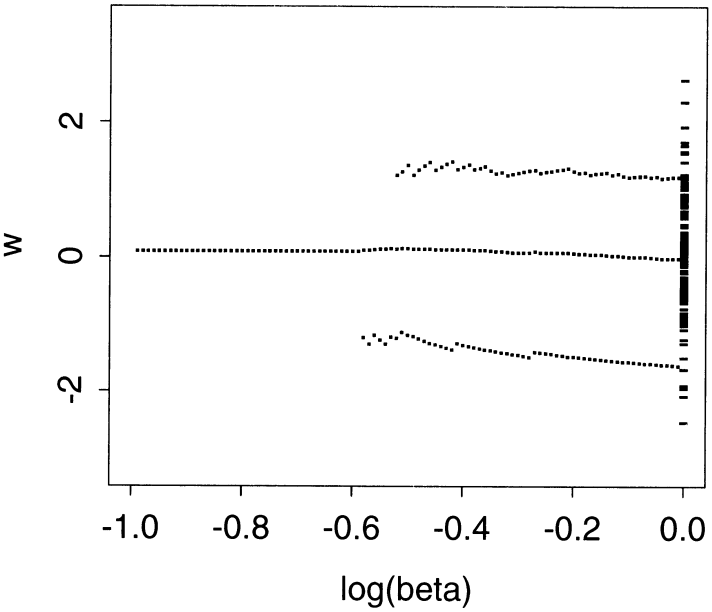

Figure 5: ML solution versus $\log(\beta)$. Data are generated from a gaussian distribution ($\kappa_4 = 0$). Horizontal axis: $\log(\beta)$; vertical axis: w or x. Dots show the ML solution for each temperature; dashes on the right show the training samples.

effective number of adaptive parameters. We conclude that this relation is violated around the SBPs. This anomalous behavior affects the optimal model selection, which in this article is the choice of the optimal β. When the optimal generalization error occurs around an SBP, increasing β will decrease (or at least not increase) training error as well as decrease NIC.

If $\kappa_4 \simeq 0$ we predict theoretically that symmetry breaking is h-way, but this is not observed numerically. We attribute this discrepancy to the delicacy of the symmetry breaking process and the numerical instability of the optimization procedure.

Although our analysis was restricted to one dimension, we expect our results to hold in higher dimensions as well. If the number of kernels is large, the condition of an even number of kernels is not very strict. If the target distribution is nonsymmetric and contains odd moments, other results could be observed.

Appendix A: Outline of the Proof of Proposition 1

We assume $\langle x \rangle = 0$ without loss of generality. Since we assume an even number of gaussian components, let $h = 2h'$.

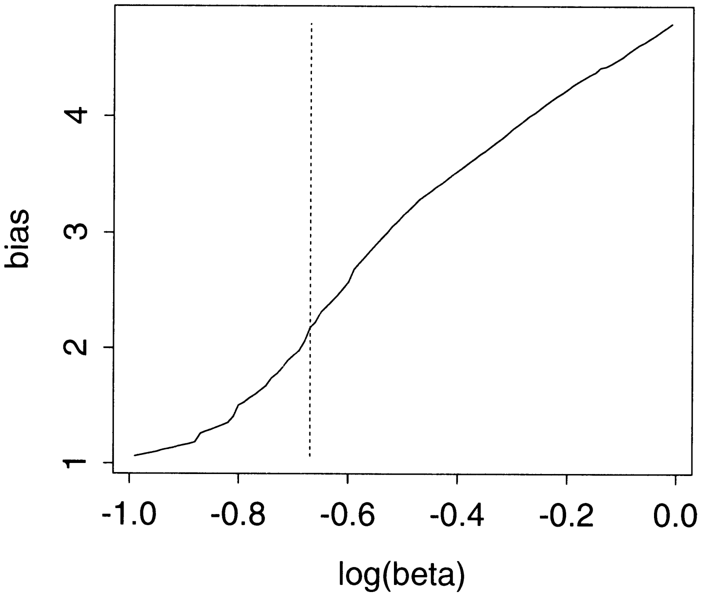

Figure 6: Generalization bias from cross-validation, averaged over 50 experiments, as a function of $\log(\beta)$. Data are generated from a gaussian distribution ($\kappa_4 = 0$). Horizontal axis: $\log(\beta)$; vertical axis: generalization bias multiplied by the number of training samples; dashed line: average of experimental values of β_c's.

Since the target distribution is symmetric and the number of gaussians is even, the symmetry breaking will be symmetric and we can write

$$p(x \mid W; \beta) = \frac{1}{2h'} \sum_{i=1}^{h'} \sqrt{\frac{\beta}{\pi}}\, p_i(x \mid w_i, \beta), \tag{A.1}$$

where

$$p_i(x \mid w_i, \beta) \equiv \exp(-\beta(x - w_i)^2) + \exp(-\beta(x + w_i)^2). \tag{A.2}$$

The derivative of the log-likelihood is defined by

$$L_i(x \mid W, \beta) \equiv \frac{\partial}{\partial w_i} \log p(x \mid W, \beta) = \frac{\partial_i p_i(x \mid w_i, \beta)}{p(x \mid W, \beta)}\ . \tag{A.3}$$

The expectation value of L_i under the target distribution q is equal to zero at the ML solution,

$$R_{\exp}(W^*, \beta) = \langle L_i(x \mid W^*, \beta) \rangle = 0. \tag{A.4}$$

Above the critical temperature, $W = 0$. Below the critical temperature, W will become nonzero. Let Δw_i denote the weight vector describing the mean of kernel i. In order to obtain a nontrivial solution of Δw_i as a function of $\Delta\beta = \beta - \beta_c$, we expand $L_i(x \mid W, \beta)$ to third order in W and to first order in β around $\beta = \beta_c$ and $W = 0$:

$$\begin{aligned}\langle L_i(x \mid W, \beta)\rangle &= \frac{2}{h'}\Delta\beta\Delta w_i - \frac{1}{2}\sum_{j\neq i}\left(\frac{s_4}{h'^2(\sigma^2)^2} - 1\right)\Delta w_i \Delta w_j^2 \\ &\quad + \frac{1}{6h'(\sigma^2)^2}\left\{\frac{s_4}{(\sigma^2)^2} - 3 - \frac{3}{h'}\left(\frac{s_4}{(\sigma^2)^2} - 1\right)\right\}\Delta w_i^3 \\ &\quad + \text{ higher order terms.}\end{aligned} \tag{A.5}$$

Setting $\langle L_i(x \mid W, \beta)\rangle = 0$ and neglecting higher-order terms, we obtain h' simultaneous equations of $\Delta\beta$ and Δw_i.

Since a solution $\Delta w_i = 0$ gives a local minimum of the likelihood, we can assume $\Delta w_i \neq 0$. If $s_4 \neq 3(\sigma^2)^2$, we obtain $\Delta w_i^2 = \Delta w_j^2$ and $\Delta\beta = \{s_4/6(\sigma^2)^4\}\Delta w_i^2$, which is the first case of proposition 1.

On the other hand, if $s_4 = 3(\sigma^2)^2$, the simultaneous equations degenerate to the equation $\Delta\beta = \sum_i \Delta w_i^2/\{2h'(\sigma^2)^2\}$. It means w_i cannot be determined uniquely from β when we neglect higher-order terms.

Appendix B: Outline of the Proof of Proposition 2

If the temperature is higher than the first SBP, all gaussians degenerate to one gaussian; therefore the only thing we should do is to calculate NIC for one gaussian model.[1]

We derive $D(W^*)$ and $H(W^*)$ for one gaussian model as follows,

$$D_{ij}(W^*) = 4\beta^2 V_{ij}, \tag{B.1}$$

$$H_{ij}(W^*) = 2\beta\delta_{ij}, \tag{B.2}$$

where V_{ij} is a covariance between x_i and x_j and δ_{ij} is Kronecker's δ. Therefore NIC is given by

$$h_{\text{NIC}}(\beta) = \text{Tr}[H(W^*)^{-1}D(W^*)] = 2\beta\text{Tr}[V_{\boldsymbol{x}}]. \tag{B.3}$$

Appendix C: Outline of the Proof of Proposition 3

Similar to appendix A, we assume $\langle x\rangle = 0$ without loss of generality. The symmetry breaking is two-way from the assumption of proposition 3. As a

[1] This is not strictly true, because the matrices H and D are still h-dimensional. However, one can show that this h dependence drops out in equation 4.4.

result, one can show that the expansion of h_{NIC} around the SBP is identical to the case of only two gaussian kernels. Therefore, we analyze the NIC of the model of two gaussians.

The model of two gaussians is written as

$$p(x \mid w_1, w_2; \beta) = \frac{1}{2}\sqrt{\frac{\beta}{\pi}}\left[\exp\{-\beta(x-w_1)^2\} + \exp\{-\beta(x+w_2)^2\}\right]. \tag{C.1}$$

$D(w_1, w_2)$ and $H(w_1, w_2)$ can be calculated from their definition. At the ML solution, we have $w_1 = w_2 = w$. Therefore we obtain

$$D(w, w) = \begin{bmatrix} d_0 & d_2 \\ d_2 & d_0 \end{bmatrix}, \tag{C.2}$$

$$H(w, w) = \begin{bmatrix} d_1 & d_2 \\ d_2 & d_1 \end{bmatrix}, \tag{C.3}$$

where

$$d_0 = \left\langle 4\beta^2 (x-w)^2 \frac{(p_1)^2}{p^2} \right\rangle, \tag{C.4}$$

$$d_1 = d_0 + \left\langle 2\beta \frac{p_1}{p} - 4\beta^2 (x-w)^2 \frac{p_1}{p} \right\rangle, \tag{C.5}$$

$$d_2 = \left\langle -4\beta^2 (x-w)(x+w) \frac{p_1 p_2}{p^2} \right\rangle, \tag{C.6}$$

where $p_1 = \exp(-\beta(x-w)^2)$, $p_2 = \exp(-\beta(x+w)^2)$ and $p = p_1 + p_2$. Let

$$\hat{h}_{\mathrm{NIC}}(\beta, w) = \mathrm{Tr}[H(w, w)^{-1} D(w, w)], \tag{C.7}$$

which is equal to $h_{\mathrm{NIC}}(\beta)$ for $w = w^*$. Expanding $\hat{h}_{\mathrm{NIC}}(\beta, w)$ around the first SBP ($\beta = \beta_c$, $w^* = 0$) with respect to β and w,

$$\hat{h}_{\mathrm{NIC}}(\beta, w) = \mathrm{Tr}[H^{-1}D] = 2\frac{d_0 d_1 - d_2^2}{d_1^2 - d_2^2}, \tag{C.8}$$

we obtain

$$\begin{aligned} \hat{h}_{\mathrm{NIC}}(\beta, w) = {} & \hat{h}_{\mathrm{NIC}}(\beta_c, 0) + \left\{ \frac{\partial}{\partial \beta} \hat{h}_{\mathrm{NIC}}(\beta_c, 0) \right\} \Delta\beta \\ & + \frac{1}{2} \left\{ \frac{\partial^2}{\partial w^2} \hat{h}_{\mathrm{NIC}}(\beta_c, 0) \right\} \Delta w^2 \\ & + \text{ higher-order terms}, \end{aligned} \tag{C.9}$$

where both the second and the third terms on the right-hand side of equation C.9 are of order $\Delta\beta$, since $\Delta w^2 \simeq \{6(\sigma^2)^4/s_4\}\Delta\beta$ at the ML solution from equation 3.2.

Substituting d_0, d_1, d_2 by their values, the coefficient of the second term is given by

$$\frac{\partial}{\partial\beta}\hat{h}_{\mathrm{NIC}}(\beta_c, 0) = 2\sigma^2, \tag{C.10}$$

and the coefficient of the third term before substituting $\beta = \beta_c$ is given by

$$\frac{1}{2}\frac{\partial^2}{\partial w^2}\hat{h}_{\mathrm{NIC}}(\beta, 0) = 4\beta\left(1 - 2\beta\sigma^2 - \frac{1-4\beta^2 s_4}{1-2\beta\sigma^2}\right). \tag{C.11}$$

When $s_4 \neq (\sigma^2)^2$, and using the fact that $\beta_c = 1/(2\sigma^2)$ and $s_4 \geq (\sigma^2)^2$, we infer that equation C.11 diverges to $-\infty$ as β converges to β_c from right.

The only case that $s_4 = (\sigma^2)^2$ is when $q(x)$ is equal to $\delta(x)$ or $(\delta(x-a) + \delta(x+a))/2$. Since there is no SBP in the former case, we need to consider only the latter case. Without loss of generality, we assume $a = 1$ and the derivative taken from the right is derived from a simple calculation,

$$\frac{\partial}{\partial\beta}\hat{h}_{\mathrm{NIC}}(\beta_c, 0) = -4. \tag{C.12}$$

References

Amari, S. (1993). A universal theorem on learning curves. *Neural Networks, 6*, 161–166.

Barkai, N., Seung H. S., & Sompolinsky, H. (1993). Scaling laws in learning of classification tasks. *Physical Review Letters, 70*, 3167–3170.

Bezdek, J. C. (1980). A convergence theorem for the fuzzy ISODATA clustering algorithms. *IEEE Trans. on PAMI, 2*, 1–8.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Soc. Ser. B, 39*, 1–38.

Jacobs, R. A., & Jordan, M. I. (1991). A competitive modular connectionist architecture. In D. Touretzky & R. Lippmann (Eds.), *Advances in neural information processing systems, 3* (pp. 767–773). San Mateo, CA: Morgan Kaufmann.

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

Kappen, B. (1993). Using Boltzmann machines for probability estimation: A general framework for neural network learning. In S. Gielen et al. (Eds.), *Proc. of ICANN'93* (pp. 521–526). Berlin: Springer-Verlag.

Kappen, B. (1995). Deterministic learning rules for Boltzmann machines. *Neural Networks, 8*, 537–548.

Moody, J. (1992). The *effective* number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In J. E. Moody, S.

J. Hanson, & R. P. Lippman (Eds.), *Advances in neural information processing systems, 4* (pp. 847–854). San Mateo, CA: Morgan Kaufmann.

Murata, N., Yoshizawa, S., & Amari, S. (1991). A criterion for determining the number of parameters in an artificial neural network model. In T. Kohonen et al. (Eds.), *Artificial neural network (ICANN)* (pp. 9–14). Amsterdam: Elsevier.

Murata, N., Yoshizawa, S., & Amari, S. (1994). Network information criterions—determining the number of parameters for an artificial neural network model. *IEEE Trans. on Neural Networks, 5*, 865–872.

Nijman, M. J., & Kappen, H. J. (1997). Symmetry breaking and training from incomplete data with radial basis Boltzmann machines. *International Journal of Neural Systems, 8*, 301–316.

Poggio, T., & Girosi, F. (1990). Networks for approximation and learning. *Proceedings of the IEEE, 78*, 1481–1497.

Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics, 14*, 1080–1100.

Rose, K., Gurewitz, E., & Fox, G. (1990). Statistical mechanics of phase transitions in clustering. *Physical Review Letters, 65*, 945–948.

Titterington, D. M. (1985). *Statistical analysis of finite mixture distribution*. New York: Wiley.

Vapnik, V. A. (1984). *Estimation of dependences based on empirical data*. New York: Springer-Verlag.

Received October 14, 1998; accepted May 5, 1999.