TR-98-22

Nonmonotonic generalization bias of Gaussian mixture models

Shotaro Akaho* Hilbert J. Kappen †

Received September 1998

Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba,
Ibaraki, 305, JAPAN

Abstract Most theories of generalization performance of learning tell us that the generalization bias, which is defined as the difference between an training error and an generalization error, increases proportionally to the number of modifiable parameters in average. The present paper, however, reports the case that the generalization bias of a Gaussian mixture model does not increase even if the superficial effective number of parameters increases, where the number of elements in the Gaussian mixture is controlled by a continuous parameter.

*Information Science Division
†University of Nijmegen

1 Introduction

Gaussian mixture models have attracted a lot of attention because they have close relation to several neural network models such as radial basis function (RBF) networks [11] and hierarchical mixture of experts (HME) networks [4, 5].

An important problem of training Gaussian mixture models is to determine the optimal number of Gaussian components[2]. We can fit a model to training samples as precisely as needed by using enough number of components. However, such a model may be different from the 'true model' from which training samples are generated. Therefore, the performance of the model should be evaluated by the generalization error which measures the error for test samples. There is a lot of literatures about the generalization problem from the statistical point of view (Amari[1, 9], Moody[8], Vapnik[14], Rissanen[12]). Most of these results address the generalization bias, which is defined as the difference between a generalization error and a training error, increases as the complexity of the model, usually modifiable parameters, increases.

In the present paper, we consider Radial Basis Boltzmann Machines (RBBM), a special class of Gaussian mixture models proposed by Kappen [13, 6], where the complexity of the model is controlled by a continuous parameter β. When β is small, the 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 phenomena of symmetry breaking, is repeated recursively for increasing β.

For a RBBM with h mixture components we first derive conditions for which the symmetry breaking is 2-way or h-way, respectively.

Next, we show an analytical result that the generalization bias of the model does not increase if the symmetry breaking is 2-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 result suggests that the ML solution just below the critical temperature is expected to realize a smaller generalization error than just above the critical temperature.

2 Radial Basis Boltzmann Machines (RBBM)

Let us consider a Gaussian mixture model in which the variance of all components is identical,

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

where W denotes the set of changeable parameters $\{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_h\}$, h is the number of Gaussian components and β is a control parameter called the 'inverse temperature' in physics β is equal to half of the inverse variance. Only the centers of the individual Gaussians $W = \{\boldsymbol{w}_i\}$ are modifiable, and all temperatures are identical and fixed. The maximal likelihood (ML) solution is derived for each temperature. This model is an unsupervised version of Radial Basis Boltzmann Machines, which are originally proposed by Rose [13] and generalized by Kappen[6, 7, 10]. In the present paper, RBBM indicates the model of (1).

For different temperatures, the ML solution behaves qualitatively as follows: For small β, the ML solution is of the form $\boldsymbol{w}_1 = \boldsymbol{w}_2 = \cdots = \boldsymbol{w}_h$, i.e. 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 don't much care about the number of Gaussian elements h, and h is assumed to sufficiently large in the following sections.

An example of the ML solutions as a function of the temperature, illustrating the symmetry breaking phenomenon is shown in Figure 1. The training data is generated from the mixture distribution: $(u[0.5, 1.5] + N[-1, 0.09])/2$ where $u[a, b]$ is the uniform distribution on $[a, b]$ and $N[\mu, v]$ is the normal distribution with mean μ and variance v. The number of training samples is 100 and the number of kernels is also 100.

3 Symmetry breaking point (SBP) of the RBBM

In this section we study the symmetry breaking process in detail. Although the behavior of symmetry breaking is complicated to analyze in general, we have some beautiful results on the first SBP (symmetry breaking point), the highest temperature in which the phase transition occurs. Those results are expected to be applicable for the other SBPs qualitatively.

3.1 The first SBP

Rose et al[13] has obtained the first SBP analytically as

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

where λ_1 is the maximal eigen value 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.

This result is considered to be applicable to other SBPs as follows. Symmetry breaking in a component occurs when β is equal to the variance of samples in the neighborhood of the component along the first principal component axis. this approximate description is correct in the limit of large distance between component centers.

3.2 Above the first SBP

When the temperature is higher than $1/\beta_c$, all the Gaussian components are degenerated into one Gaussian and the true ML solution $\boldsymbol{w}_i$ is given by[13],

$$\boldsymbol{w}_i = \langle \boldsymbol{x} \rangle, \tag{3}$$

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

$$\langle \cdot \rangle \equiv \int \cdot \, q(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \tag{4}$$

and $q(\boldsymbol{x})$ is the true probabilistic density distribution from which training samples and test samples are generated.

3.3 Below the first SBP

We can characterize behavior of the symmetry breaking under some assumptions.

Assumption 1 The target distribution $q(x)$ is defined on $\mathcal{R}$ (one dimension), and is assumed to be symmetric. Moreover, the number of Gaussian components h 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 the second and fourth order moment of the distribution q. The fourth order cumulant is defined as $\kappa_4 \equiv s_4 - 3(\sigma^2)^2$ (κ_4/s_4 is called 'kurtosis' in statistics).

Theorem 1 Under the Assumption 1, the behavior of the first symmetry breaking is classified into the following two cases:

1. If $\kappa_4 \neq 0$ the symmetry breaking is 2-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 written as

   $$\Delta\beta \simeq \frac{s_4}{6(\sigma^2)^4}(\Delta w_i)^2, \tag{5}$$

   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 Gaussian components are separated into h clusters in the sense of third order approximation. The relation between the ML solution w_i and β in the neighborhood of the first SBP β_c is written as

   $$\Delta\beta \simeq \frac{1}{2(\sigma^2)^2}\sigma_w^2, \tag{6}$$

   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 forth order cumulant. Outline of the proof of Theorem 1 is given in Appendix 1.

Equation (6) is interpreted as follows: Suppose the true distribution is one Gaussian and there are an infinitely large number of Gaussian elements, we obtain the explicit form of the ML solution as the distribution of w, which is a Gaussian distribution with the variance $2(\sigma^2)^2\Delta\beta$. This fact corresponds to (6).

4 Nonmonotonic generalization bias

In this section, we briefly summarize a generalization theory and show our main result about nonmonotonic generalization bias of the RBBM.

4.1 Generalization bias

A lot of theories of generalization of neural network learning have been proposed in recent years. The goal of learning is to obtain the best possible generalization performance. This is defined as the optimum of the true likelihood, and requires knowledge of the complete target distribution. However, neural networks are trained so as to maximize a likelihood calculated just from a finite set of training samples. The difference between the true likelihood and the sample likelihood, which is called the generalization bias, causes overtraining, which results in suboptimal performance on test samples. The optimal model is selected by varying the number of adaptive parameters. 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. Overtraining results in a suboptimal effective number of components and this in a different clustering result from desired one.

Although the generalization bias is stochastic and unknown in general, we can estimate the mean value of the generalization bias. Using this value, we can select the model which minimizes the sum of the training likelihood and the mean generalization bias in order to solve the overtraining.

A well known generalization bias is AIC (Akaike's information criterion), which is given by the number of independent parameters. However, AIC assumes that the target distribution belongs to the model set, and NIC (neural information criterion) proposed by Murata et al[9] or an effective number of parameters by Moody [8] generalizes AIC to apply the case that the target distribution does not belong to the model set.

4.2 Neural information criterion (NIC)

Given a set of P training samples $X^{(P)} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_P\}$ from $q(\boldsymbol{x})$, the ML solution $W = W^{(P)}$ maximizes the expected likelihood over the training samples,

$$R_{\rm emp}^{(P)}(W;\beta) \equiv \frac{1}{P}\sum_{i=1}^{P} \log p(\boldsymbol{x}_i \mid W;\beta). \tag{7}$$

The true likelihood for the parameter W is defined by

$$R_{\rm exp}(W;\beta) \equiv \langle \log p(\boldsymbol{x} \mid W;\beta)\rangle. \tag{8}$$

When P is large enough, the mean generalization bias is asymptotically given by

$$\langle R_{\rm emp}^{(P)}(W^{(P)};\beta)\rangle - R_{\rm exp}(W^{(P)};\beta) \simeq \frac{h_{\rm NIC}}{P}, \tag{9}$$

where the average is taken over $X^{(P)}$ and $h_{\rm NIC}$, the neural information criterion defined by

$$h_{\rm NIC}(\beta) \equiv {\rm Tr}[H(W^*)^{-1}D(W^*)], \tag{10}$$

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

$$H_{ij}(W) \equiv -\langle \partial_i \partial_j \log p(\boldsymbol{x};W,\beta)\rangle, \tag{11}$$

$$D_{ij}(W) \equiv \langle \partial_i \log p(\boldsymbol{x} \mid W,\beta)\, \partial_j \log p(\boldsymbol{x} \mid W,\beta)\rangle, \tag{12}$$

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

In the following subsections, we provide behavior of NIC for RBBM. Our analysis explains that the linear increase of NIC before the symmetry breaking as well as the reason of the decrease.

4.3 Above the first SBP

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

Theorem 2 If $\beta < \beta_c$, NIC is given by

$$h_{\rm NIC}(\beta) = 2\beta {\rm Tr}[V_{\boldsymbol{x}}], \tag{13}$$

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

Therefore, the NIC of the RBBM increases in proportion to β. Outline of the proof of Theorem 2 is given in Appendix 2.

4.4 Below the first SBP

Although the situation below the critical temperature is very complicated, we can analyze the NIC under the same assumption in section 3.3.

Theorem 3 Under the Assumption 1, and also if $\kappa_4 \neq 0$ and $s_4 \neq (\sigma^2)^2$, the right differential coefficient of NIC with respect to β is

$$\frac{\partial}{\partial \beta} h_{\rm NIC}(\beta_c) = -\infty. \tag{14}$$

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

Outline of the proof of Theorem 3 is given in Appendix 3. Theorem 3 states that the NIC of the RBBM decreases even if the superficial number of parameters increases when the symmetry breaking in the first SBP is 2-way.

It is not easy to analyze the case $\kappa_4 = 0$, since the symmetry breaking is h-way and the ML solution is not uniquely determined in the sense of third order approximation as shown in Theorem 1.

5 Experiments

In this section, we show some computer simulation results for the two cases with different forth order cumulants presented in Theorem 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 samples breaks around this value. In the following simulation,

we use the EM algorithm (with acceleration)[3], by which the solution sometimes gets trapped into a local minimum when $\beta > \beta_c$. Both the number of training samples(P) and the number of Gaussian components(h) is set to 100. The initial solution of the EM algorithm is that a center of each Gaussian component is set to each training sample. 100,000 test samples are generated from the same distribution as training samples to calculate the generalization bias.

5.1 The case of $\kappa_4 \neq 0$

The true distribution is a mixture of two Gaussians expressed by

$$q(x) = \frac{1}{2}\sqrt{\frac{C_1}{\pi}}\left[\exp\left\{-C_1(x-C_2)^2\right\} + \exp\left\{-C_1(x+C_2)^2\right\}\right]. \tag{15}$$

where $C_1 = 12.5, C_2 = \sqrt{0.96}$.

A typical solution of the ML solution by changing the temperature around the first SBP is shown in Figure 2.

The generalization bias averaged over 50 experiments with different random numbers are shown in Figure 3, which supports the result of Theorem 3 qualitatively.

5.2 The case of $\kappa_4 = 0$

The true distribution is one Gaussian with a unit variance,

$$q(x) = \frac{1}{\sqrt{2\pi}}\exp(-x^2/2). \tag{16}$$

In this case, the convergence of the EM algorithm was much more unstable than in the case of the previous section, because the lower order terms degenerate.

A typical solution of the ML solution by changing the temperature around the first SBP is shown in Figure 4.

The generalization bias averaged over 50 experiments with different random numbers are shown in Figure 5. We did not observe a clear case among this 50 experiments where the likelihood is maximal around the first SBP as clearly 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 forth 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 effective number of adaptive parameters. We conlude that this relation is

violated around the SBPs. 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. 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.

The analysis of the case that Assumption 1 does not hold and the experiments for more general type of mixture models are future works. In addition, we conjecture that one may avoid instabilities associated with $\kappa_4 \simeq 0$ by using super-Gaussians or sub-Gaussian kernels.

Acknowledgement

The research is partly supported by Real World Computing Program.

References

[1] S. Amari. A universal theorem on learning curves. *Neural networks*, 6:161–166, 1993.

[2] N. Barkai, H. S. Seung, and H. Sompolinsky. Scaling Laws in Learning of Classification Tasks. *Physical Review Letters*, pages 3167–3170, 1993.

[3] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc.* B, 39:1–38, 1977.

[4] R. A. Jacobs and M. I. Jordan. A competitive modular connectionist architecture. In Lippman et al, editor, *Advances in neural information processing systems*, pages 767–773, 1991.

[5] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214, 1994.

[6] B. Kappen. Using Boltzmann Machines for probability estimation: A general framework for neural network learning. In *Proc. of ICANN'93*, 1993.

[7] B. Kappen. Deterministic learning rules for Boltzmann Machines. *Neural Networks*, 8(4):537–548, 1995.

[8] J. Moody. The *effective* number of parameters: an analysis of generalization and regularization in nonlinear learning systems. In J. Moody, S. Hanson, and R. Lippmann, editors, *Advances in Neural Information Processing Systems*. Morgan Kaufmann, 1992.

[9] N. Murata, S. Yoshizawa, and S. Amari. Network information criterions – determining the number of parameters for an artificial neural network model. *IEEE Transactions on Neural Networks*, 5:865–872, 1994.

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

[11] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497, 1990.

[12] J. Rissanen. Stochastic complexity and modeling. *Annals of Statistics*, 14:1080–1100, 1986.

[13] K. Rose, E. Gurewitz, and G. Fox. Statistical mechanisc of phase transitions in clustering. *Physical Review Letters*, 65:945–948, 1990.

[14] V.A. Vapnik. *Estimation of Dependences Based on Empirical Data.* Springer-Verlag, 1984.

Appendix 1 Outline of the proof of Theorem 1

We assume $\langle x \rangle = 0$ without loss of generality. Since we assume the Gaussian components of the RBBM model is even, let $h = 2h'$.

Since the true distribution is symmetric and the number of Gaussians is even, the ML solution can be written as

$$p(x \mid W;\beta) = \frac{1}{2h'} \sum_{i=1}^{h'} \sqrt{\frac{\beta}{\pi}}\, p_i(x \mid w_i, \beta), \tag{17}$$

where

$$p_i(x \mid w_i, \beta) \equiv \exp(-\beta(x - w_i)^2) + \exp(-\beta(x + w_i)^2). \tag{18}$$

The derivative of log likelihood is defined by

$$L_i(x \mid W, \beta) \equiv \frac{\partial}{\partial w_i} \log p(x \mid W, \beta) = \frac{\partial_i p_i(x \mid w_i, \beta)}{p(x \mid W, \beta)}, \tag{19}$$

and the mean value of L_i is equal to zero at the true ML solution,

$$R_{\exp}(W^*, \beta) = \langle L_i(x \mid W^*, \beta) \rangle = 0. \tag{20}$$

Expanding $L_i(x \mid W, \beta)$ by W and β around the first SBP,

$$\begin{aligned} \langle L_i(x \mid W, \beta) \rangle \;=\; & \frac{2}{h'} \Delta\beta \Delta w_i - \frac{1}{2} \sum_{j \neq i} \left(\frac{s_4}{h'^2 (\sigma^2)^2} - 1 \right) \Delta w_i \Delta w_j^2 \\ & + \frac{1}{6h'(\sigma^2)^2} \left\{ \frac{s_4}{(\sigma^2)^2} - 3 - \frac{3}{h'} \left(\frac{s_4}{(\sigma^2)^2} - 1 \right) \right\} \Delta w_i^3 \\ & + \text{higher order terms}, \end{aligned} \tag{21}$$

which is equal to zero if $W = W^*$.

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 Theorem 1.

On the other hand, if $s_4 = 3(\sigma^2)^2$, the simultaneous equations degenerate to the following equation, $\Delta\beta = \sum_i \Delta w_i^2/\{2h'(\sigma^2)^2\}$, which is the second case of Theorem 1 and it means w_i can not be determined uniquely from β when we neglect higher order terms.

Appendix 2 Outline of the proof of Theorem 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.

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

$$D_{ij}(W^*) = 4\beta^2 V_{ij}, \tag{22}$$

$$H_{ij}(W^*) = 2\beta\delta_{ij}, \tag{23}$$

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

$$h_{\mathrm{NIC}}(\beta) = \mathrm{Tr}[H(W^*)^{-1}D(W^*)] = 2\beta\mathrm{Tr}[V_{\boldsymbol{x}}]. \tag{24}$$

Appendix 3 Outline of the proof of Theorem 3

Similar to Appendix 1, we assume $\langle x\rangle = 0$ without loss of generality. The symmetry breaking is 2-way from the assumption of Theorem 3, we analyze NIC of the model of 2 Gaussians.

The model of 2 Gaussians is written by

$$p(x \mid w_1, w_2; \beta) = \frac{1}{2}\sqrt{\frac{\beta}{\pi}}\left[\exp\{-\beta(x-w_1)^2\} + \exp\{-\beta(x+w_2)^2\}\right]. \tag{25}$$

$D(w_1, w_2)$ and $H(w_1, w_2)$ can be calculated by definition, and letting $w_1 = w_2 = w$ because $w_1 = w_2$ at the ML solution, we obtain

$$D(w, w) = \begin{bmatrix} d_0 & d_2 \\ d_2 & d_0 \end{bmatrix}, \tag{26}$$

$$H(w, w) = \begin{bmatrix} d_1 & d_2 \\ d_2 & d_1 \end{bmatrix}, \tag{27}$$

where

$$d_0 = \langle 4\beta^2(x-w)^2\frac{(p_1)^2}{p^2}\rangle, \tag{28}$$

$$d_1 = d_0 + \langle 2\beta\frac{p_1}{p} - 4\beta^2(x-w)^2\frac{p_1}{p}\rangle, \tag{29}$$

$$d_2 = \langle -4\beta^2(x-w)(x+w)\frac{p_1 p_2}{p^2}\rangle, \tag{30}$$

where $p_1 = \exp(-\beta(x-w)^2)$, $p_2 = \exp(-\beta(x+w)^2)$ and $p = p1 + p2$. Let

$$\hat{h}_{\rm NIC}(\beta, w) = {\rm Tr}[H(w,w)^{-1} D(w,w)], \tag{31}$$

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

$$\hat{h}_{\rm NIC}(\beta, w) = {\rm Tr}[H^{-1} D] = 2\frac{d_0 d_1 - d_2^2}{d_1^2 - d_2^2}, \tag{32}$$

we obtain

$$\begin{aligned} \hat{h}_{\rm NIC}(\beta, w) &= \hat{h}_{\rm NIC}(\beta_c, 0) + \left\{ \frac{\partial}{\partial \beta} \hat{h}_{\rm NIC}(\beta_c, 0) \right\} \Delta\beta \\ &\quad + \frac{1}{2} \left\{ \frac{\partial^2}{\partial w^2} \hat{h}_{\rm NIC}(\beta_c, 0) \right\} \Delta w^2 \\ &\quad + \text{higher order terms}, \end{aligned} \tag{33}$$

where both the second and the third terms of the right hand side of the above equation is of order $\Delta\beta$, since $\Delta w^2 \simeq \{6(\sigma^2)^4/s_4\}\Delta\beta$ at the ML solution from (5).

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

$$\frac{\partial}{\partial \beta} \hat{h}_{\rm NIC}(\beta_c, 0) = 2\sigma^2, \tag{34}$$

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

$$\frac{1}{2} \frac{\partial^2}{\partial w^2} \hat{h}_{\rm NIC}(\beta_c, 0) = 4\beta(1 - 2\beta\sigma^2 - \frac{1 - 4\beta^2 s_4}{1 - 2\beta\sigma^2}). \tag{35}$$

Using the fact that $\beta_c = 1/(2\sigma^2)$ and $s_4 \geq (\sigma^2)^2$, (35) diverges to $-\infty$ as β converges to β_c from right, if $s_4 \neq (\sigma^2)^2$.

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, only the problem is the latter case. Without loss of generality, we assume $a = 1$ and the right differential coefficient is derived from a simple calculation,

$$\frac{\partial}{\partial \beta} \hat{h}_{\rm NIC}(\beta_c, 0) = -4. \tag{36}$$

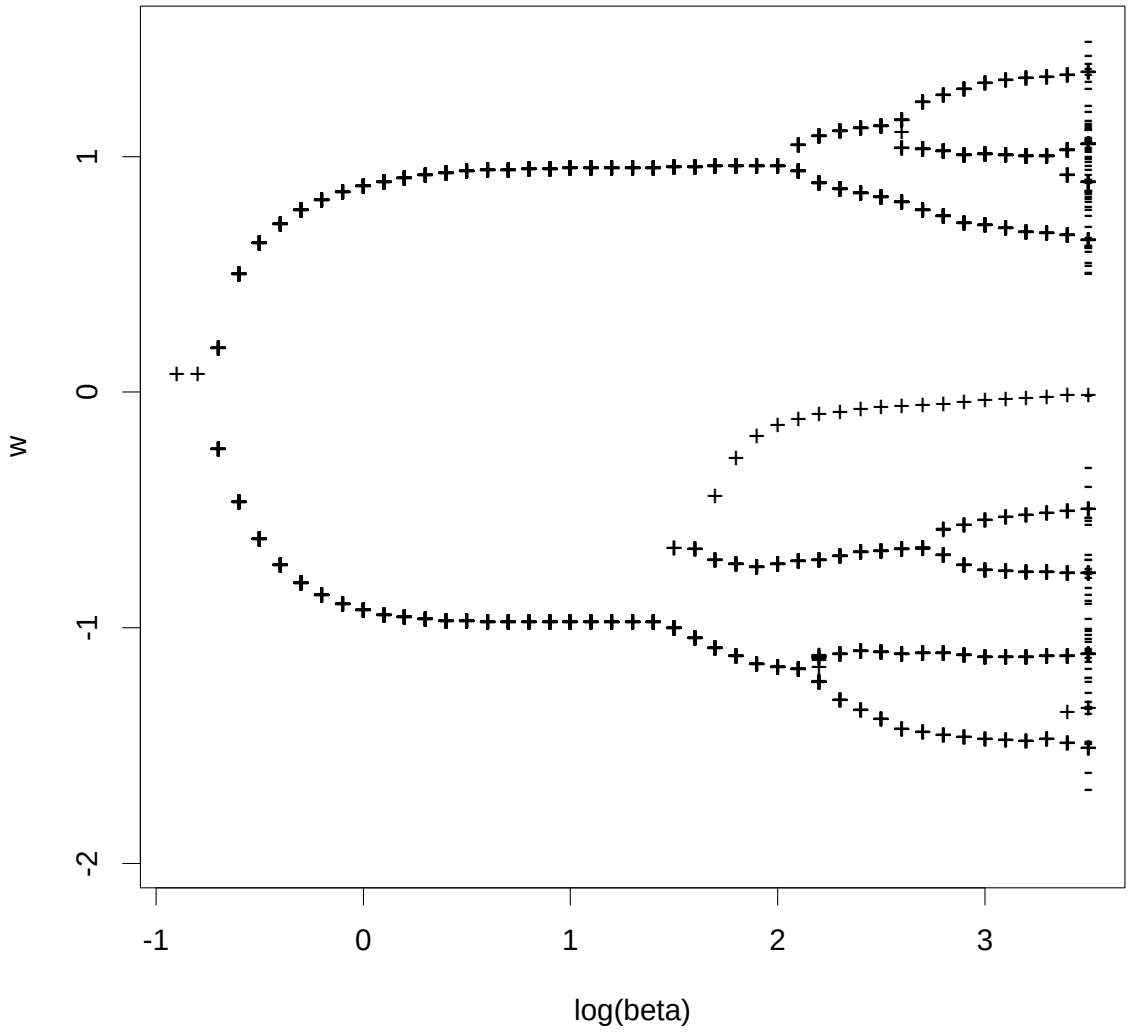

Figure 1: An example of ML solution and symmetry breaking phenomena. horizontal axis : $\log(\beta)$; vertical axis : w or x, $+$ signs show the ML solution for each temperature, $-$ signs in the right end show the training samples.

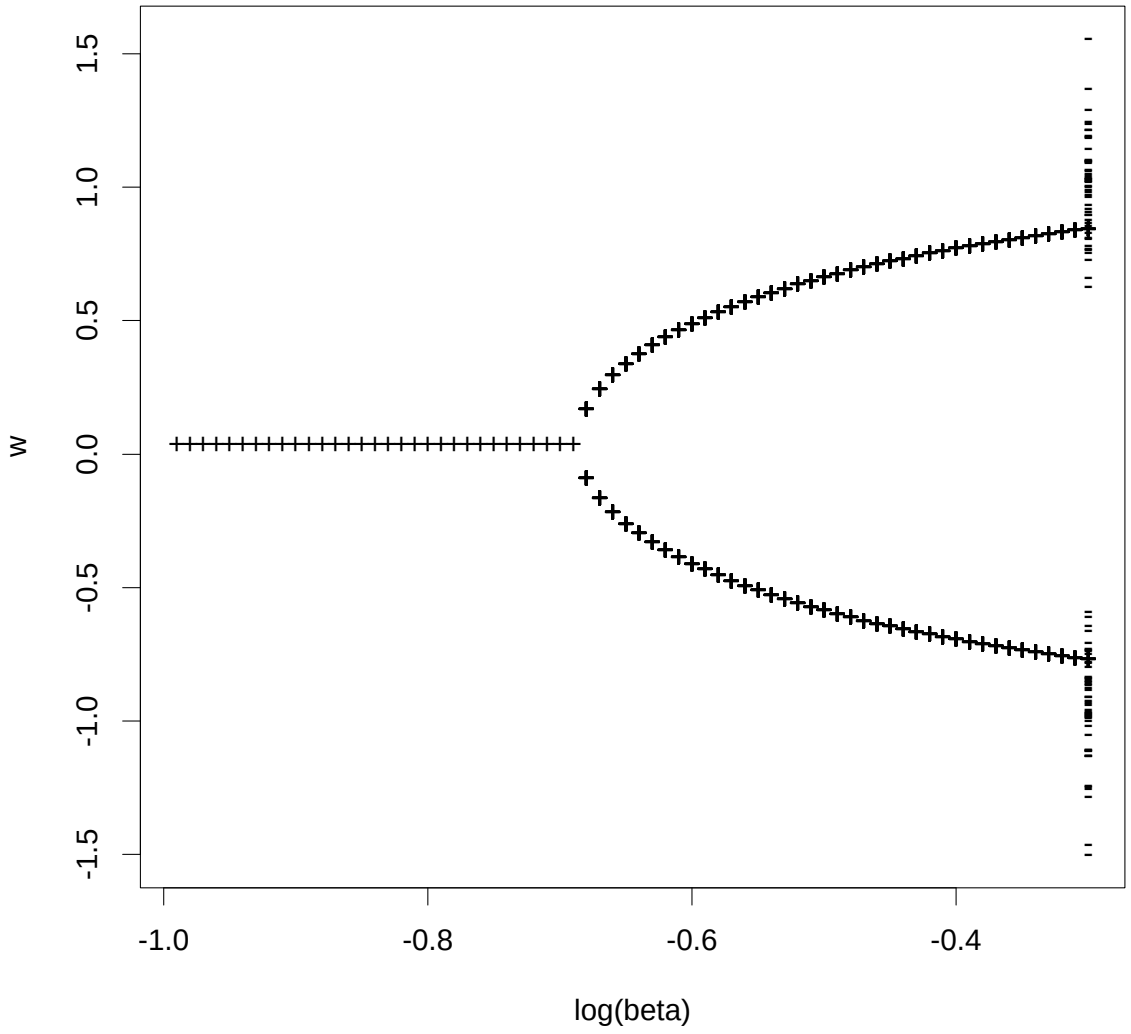

Figure 2: ML solution ($\kappa_4 \neq 0$) horizontal axis : $\log(\beta)$; vertical axis : w or x, $+$ signs show the ML solution for each temperature, $-$ signs in the right end show the training samples.

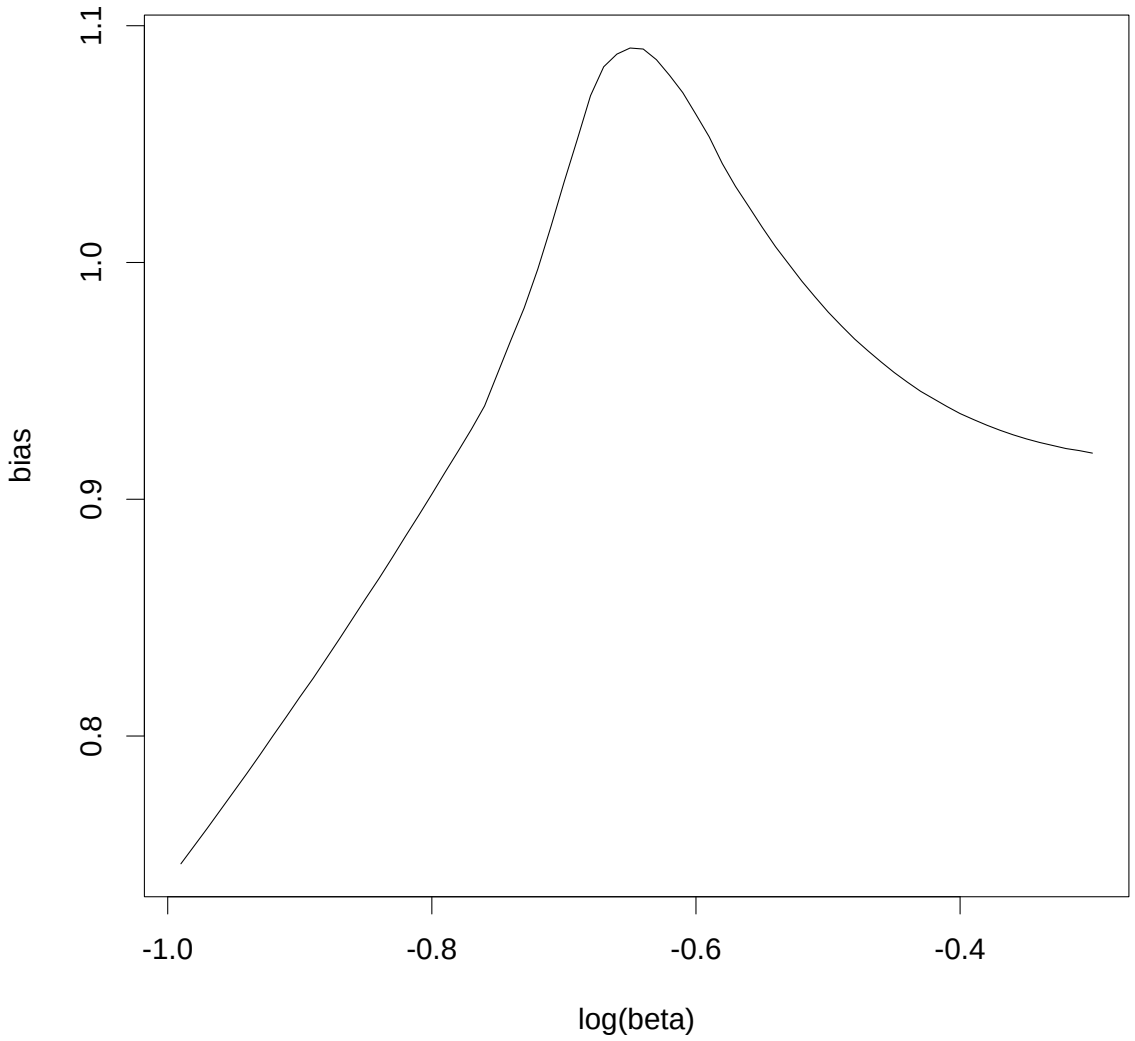

Figure 3: Generalization bias ($\kappa_4 \neq 0$) averaged over 50 experiments. horizontal axis : $\log(\beta)$; vertical axis : generalization bias multiplied by the number of training samples.

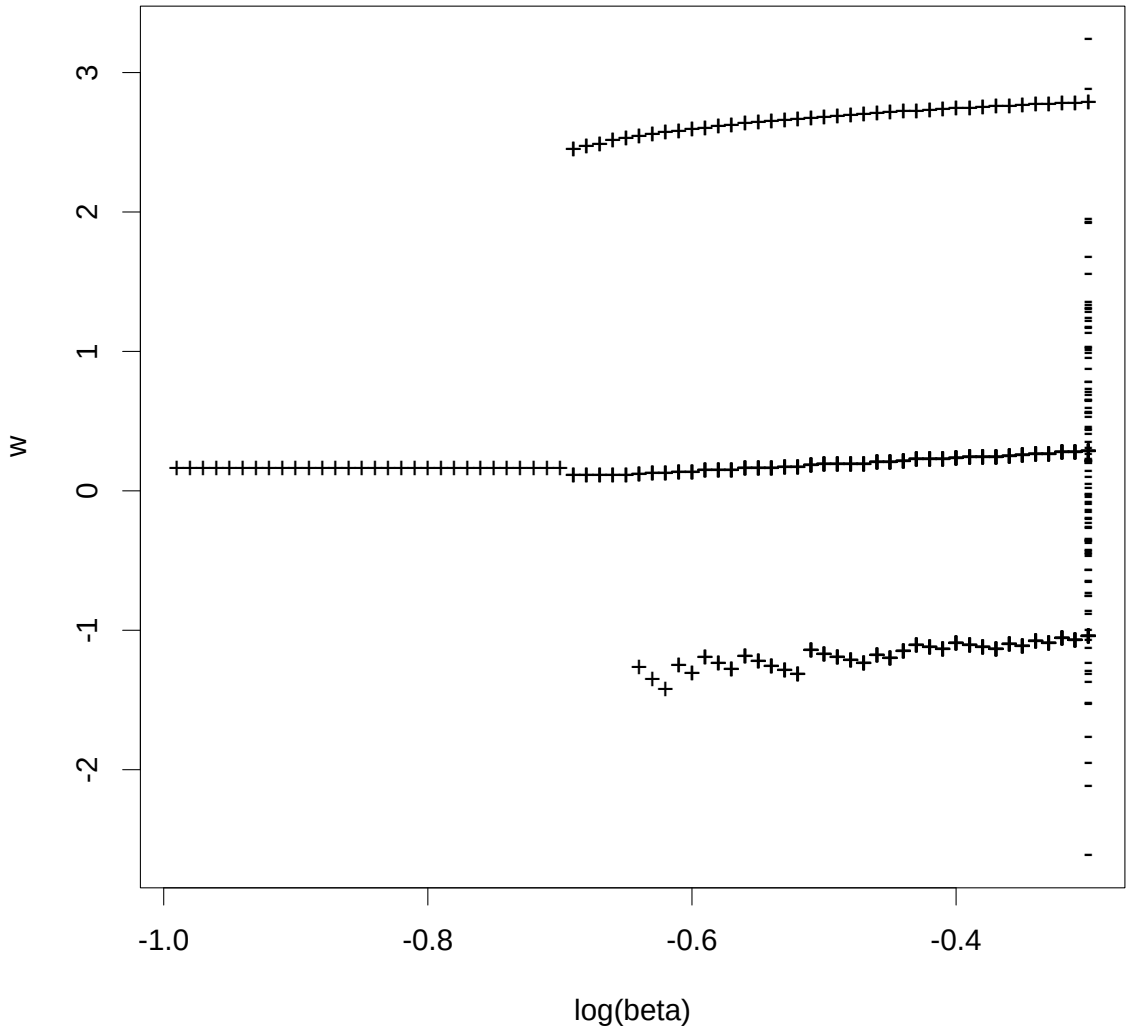

Figure 4: ML solution (Gaussian) horizontal axis : $\log(\beta)$; vertical axis : w or x, $+$ signs show the ML solution for each temperature, $-$ signs in the right end show the training samples.

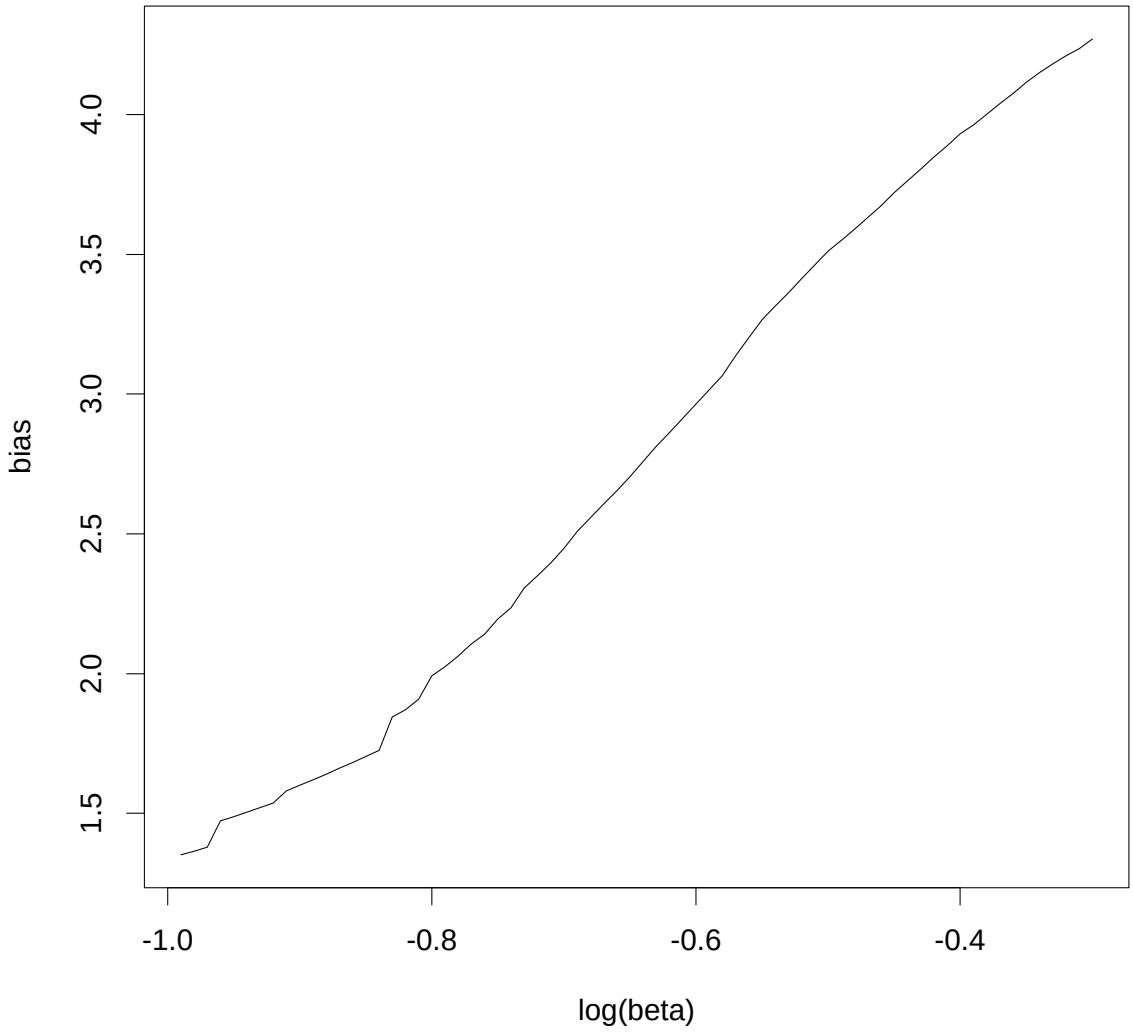

Figure 5: Generalization bias (Gaussian) averaged over 50 experiments. horizontal axis : $\log(\beta)$; vertical axis : generalization bias multiplied by the number of training samples.