Statistical Analysis of Regularization Constant

— From Bayes, MDL and NIC Points of View

Shun-ichi Amari and Noboru Murata

RIKEN Frontier Research Program

Wako-shi, Hirosawa 2-1, Saitama 351-01, JAPAN

fax: +81-48-462-9881

amari@zoo.riken.go.jp; mura@zoo.riken.go.jp

Abstract

In order to avoid overfitting in neural learning, a regularization term is added to the loss function to be minimized. It is naturally derived from the Bayesian standpoint. The present paper studies how to determine the regularization constant from the points of view of the empirical Bayes approach, the maximum description length (MDL) approach, and the network information criterion (NIC) approach. The asymptotic statistical analysis is given to elucidate their differences. These approaches are tightly connected with the method of model selection. The superiority of the NIC is shown from this analysis.

1 Introduction

Overfitting is a serious problem in neural learning. Model selection is an important method to avoid overfitting. It chooses a model of an adequate size depending on the number of examples. To this end, a number of criteria for fitness of models are proposed. One criterion is the MDL (Rissanen, 1989) which minimizes the coding or description length for a given set of examples. This is closely related to the empirical Bayes criterion of model selection (McKay, 1992 ; Rissanen, 1989).

Another criterion for model selection is the NIC (Murata et al., 1994) which is a generalization of the AIC (Akaike, 1974). This is the same as the method of effective degrees of freedom (Moody, 1992). This criterion minimizes an unbiased estimate of the generalization error.

Brake et al. (1995) compared these criteria and reported that no general superiority is found among these two. In other words, one is better in some examples and the other is better in other examples. See also Ripley (1995).

Introduction of a regularization term is another commonly used technique for avoiding overfitting (Poggio and Girosi, 1990). In this case, it is necessary to determine the regularization constant adequately. The empirical Bayes method (McKay, 1992) and crossvalidation method are widely used for this purpose.

The present paper uses the above-mentioned criteria of fitness of models to determine the regularization constant. We give an asymptotic statistical analysis of the optimal choice of the regularization constant under various criteria. This analysis elucidates the relation between the model selection and the regularization theory. It is shown that the NIC is useful for this purpose (Murata et al., 1994). The NIC is expected to give asymptotically a better result than the Bayesian method.

2 Stochastic Framework of Neural Learning

Let us consider a parametric model of stochastic input-output systems including neural networks. Let x and y be an input vector and an output scalar, respectively, where the conditional distribution of y conditioned on x is given by $p(y|x;w)$ specified by w. Here, w is a d-dimensional vector to parameterize this model. The system is given by the set of conditional probability distributions,

$$M = \{p(y|x;w),\ w \in \boldsymbol{R}^d\}. \tag{2.1}$$

A typical example is a multilayer perceptron with an additive Gaussian noise n, where output y is given by

$$y = f(x;w) + n. \tag{2.2}$$

Here, $f(x;w)$ is the deterministic (noiseless) output from the multilayer perceptron specified by the vector w consisting of all the modifiable parameters (synaptic weights and thresholds). Then,

$$p(y|x;w) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}\{y - f(x;w)\}^2\right] \tag{2.3}$$

where σ^2 is the variance of the noise, that is, $n \sim N(0, \sigma^2)$.

Let $D_T = \{(x_1, y_1), \cdots, (x_T, y_T)\}$ be a set of T input-output training examples generated independently from an unknown but fixed probability distribution $q(x)q(y|x)$, where $q(x)$ is the probability distribution of input x and $q(y|x)$ represents the stochastic mechanism of the teacher that generates the output y from the input x.

Let us consider a loss function given by

$$l(x, y; w) = -\log q(x)p(y|x;w). \tag{2.4}$$

The empirical loss evaluated by the examples D_T is called the training error, and is given by

$$L_{\text{train}}(w) = \sum_{t=1}^{T} l(x_t, y_t; w). \tag{2.5}$$

The maximum likelihood estimator $\hat{w}$ is the minimizer of $L_{\text{train}}(w)$,

$$\hat{w} = \operatorname{argmin} L_{\text{train}}(w). \tag{2.6}$$

In order to avoid overfitting, we add a regularization term $r(w)$ which penalizes large values w_i of the components of w. A typical example is

$$r(w) = \frac{1}{2}\sum_{i=1}^{d}(w_i)^2. \tag{2.7}$$

The loss function to be minimized is then written as

$$l(x, y; w, \lambda) = l(x, y; w) + \lambda r(w), \tag{2.8}$$

where λ is called the regularization constant.

Given λ, the minimizer of

$$L_{\text{train}}(w, \lambda) = L_{\text{train}}(w) + \lambda r(w) \tag{2.9}$$

is denoted by

$$\hat{w}_\lambda = \operatorname{argmin} L_{\text{train}}(w, \lambda). \tag{2.10}$$

The loss and the estimator without the regularization term correspond to the case with $\lambda = 0$.

The estimator $\hat{w}_\lambda$ should be evaluated by the generalization error without λ. It is given by

$$L_{\text{gen}}(w) = E[l(x, y; w)], \tag{2.11}$$

which is the expectation of the loss with respect to future input-output pairs (x, y). This is equivalent to the Kullback-Leibler divergence loss given by

$$L^M_{\text{gen}}(w) = D[q(x)q(y|x) : q(x)p(y|x; w)]. \tag{2.12}$$

Let w^* be the optimal parameter such that

$$w^* = \operatorname*{argmin}_w L^M_{\text{gen}}(w).$$

3 Bayesian, MDL and NIC Criteria for Model Selection

We first explain the Bayesian standpoint (McKay, 1992). The Bayesian statistics assumes that the unknown parameter w is subject to the prior probability $\pi(w)$. However, we do not know $\pi(\omega)$ in many cases. In such a case, there is a method of determing $\pi(\omega)$ from the observed data. This is called the empirical Bayes method. It assumes a parametric family $\pi(w; \lambda)$ of the prior distributions specified by an extra parameter λ called the hyper parameter.

In the case of neural learning, it is typically written as

$$\pi(w; \lambda) = \exp\{-\lambda r(w) + c(\lambda)\}, \tag{3.1}$$

where

$$c(\lambda) = -\log \int \exp\{-\lambda r(w)\} dw. \tag{3.2}$$

When $r(w)$ is given by

$$\begin{aligned} r(w) &= \frac{1}{2} \sum_{i=1}^{d} (w_i)^2, \\ c(\lambda) &= \frac{1}{2} \log \lambda - \frac{1}{2} \log(2\pi). \end{aligned} \tag{3.3}$$

The present paper assumes that $\pi(w;\lambda)$ is differentiable with respect to w. The non-differentiable case is also important, and will be studied in a forthcoming paper.

Given λ, the joint probability of D_T and w is given by

$$p(D_T, w, \lambda) = \pi(w;\lambda) \prod_{t=1}^{T} q(x_t) p(y_t|x_t;w). \tag{3.4}$$

The parameter $\hat{w}_\lambda = \hat{w}_\lambda(D_T)$ that maximizes $p(D_T, w, \lambda)$, or equivalently that maximizes the posterior distribution of w under the condition that D_T is observed,

$$p(w|D_T) = \frac{p(D_T, w, \lambda)}{p(D_T, \lambda)}, \tag{3.5}$$

is called the maximum posterior estimator. The maximum likelihood estimator is given by $\hat{w}_0$ with $\lambda = 0$. We denote the negative logarithm of the joint probability by

$$L_{\text{train}}(D_T, w, \lambda) = -\log \pi(w;\lambda) - \sum_{t=1}^{T} l(y_t|x_t;w) - c \tag{3.6}$$

where

$$c = \sum \log q(x_t)$$

does not depend on w and λ. This term is neglected hereafter. In the case of the multilayer perceptron,

$$L_{\text{train}}(D_T, w, \lambda) = \frac{1}{2\sigma^2} \sum_{t=1}^{T} \{y_t - f(x_t;w)\}^2 + \lambda r(w) + c(\lambda). \tag{3.7}$$

Hence, the maximum posterior estimator $\hat{w}_\lambda$ is the parameter that minimizes the sum of the squared error and the regularization term $\lambda r(w)$.

Now let us define the marginal distribution of D_T,

$$p(D_T, \lambda) = \int p(D_T, w, \lambda) dw. \tag{3.8}$$

In order to evaluate the likelihood of the observed data D_T, we consider a number of statistical models $M_1, M_2, M_3, \cdots$, where each M_i specifies a conditional probability $p_{M_i}(y|x;w^{(i)})$. The log likelihood of data D_T under the Bayesian model M with prior $\pi(w;\lambda)$ is given by

$$L_M(D_T, \lambda) = -\log p_{M_i}(D_T, \lambda). \tag{3.9}$$

From the Bayesian standpoint, the model M_{i_0} that maximizes $L_{M_i}(D_T, \lambda)$ is regarded as the best model that explains the observed data D_T, where λ is a given constant. Hence, the

Bayesian criterion of model selection is given by $L_M(D_T, \lambda)$. When T is large, we have the following asymptotic expansion.

Theorem 1.

$$L_M(D_T, \lambda) = L_{\text{train}}(D_T, \hat{w}_\lambda, \lambda) + d \log T + O_p(1). \tag{3.10}$$

Proof is given in Appendix.

This shows the Bayesian criterion of model selection.

The $L_M(D_T, \lambda)$ depends on the hyperparameter λ. In order to explain data D_T, it is wise to choose such λ that minimizes the negative of the likelihood. So we define

$$L_M(D_T) = \min_\lambda L_M(D_T, \lambda). \tag{3.11}$$

This gives us the criterion of model selection which does not depend on λ.

We now describe the three criteria of the model selection. For various models M and given data D_T, the model M^* that minimizes the following criterion should be selected.

I. Bayesian criterion of model selection: $L^{\text{Bayes}}_{\text{train}}(D_T, \lambda)$ defined by (3.6).

II. MDL criterion :

$$L^{\text{MDL}}_M(D_T, 0) = L_{\text{M, train}}(\hat{w}) + d \log T.$$

III. NIC criterion :

$$L^{\text{NIC}}_M(D_T) = L_{M,\text{train}}(\hat{w}) + 2\text{tr}(K^{-1}G),$$

where K and G are matrices defined by

$$\begin{aligned} K &= E[\nabla_w \nabla_w l(x, y; \hat{w})], \\ G &= K[\nabla_w l(x, y; \hat{w}) \nabla_w l(x, y; \hat{w})], \end{aligned}$$

∇_w denoting the gradient with respect to w.

When the optimal model includes the true distribution, $G = K$ and is the Fisher information matrix. In this case, we have

$$L^{\text{NIC}}_M(D_T) = L_{M,\text{train}}(D_T) + 2d.$$

The MDL and NIC criteria do not directly include the regularization term, whereas the Bayesian criterion depends on λ. When $\lambda = 0$, it coincides with the MDL criterion that minimizes the description length. It is possible to generalize the NIC such that it depends on λ, because the NIC is defined for a general loss function including the regularization term other than the log loss.

We analyze in the following the effect of the regularization term. Here the value of λ is determined such that it minimizes the model selection criteria.

4 Bayesian Method of Determining λ

The optimal λ is determined form the Bayesian point of view to minimize $L_M(D_T, \lambda)$. We have the following theorem.

Theorem 2. The Bayesian optimal λ is given by the solution of

$$\frac{d}{d\lambda} \log \pi(\hat{w}_\lambda, \lambda) = 0. \tag{4.1}$$

In particular, for the multilayer perceptron with the Gaussian prior (1.7),

$$\lambda \text{opt} = \frac{d}{2r(\hat{w}_0)} + O_p\left(\frac{1}{T}\right) = \frac{d}{\sum(\hat{w}_i^0)^2} + O_p\left(\frac{1}{T}\right). \tag{4.2}$$

Proof. The Bayesian optimal λ maximizes $l(D_T, \hat{w}_\lambda, \lambda)$. Therefore, we have

$$\frac{d}{d\lambda} l(D_T, \hat{w}_\lambda, \lambda) = \frac{\partial}{\partial w} l(D_T, \hat{w}_\lambda, \lambda) \cdot \frac{d}{d\lambda} \hat{w}_\lambda + \frac{\partial}{\partial \lambda} l(D_T, \hat{w}_\lambda, \lambda) = 0. \tag{4.3}$$

Since $\nabla_w l(D_T, \hat{w}_\lambda, \lambda) = 0$, from (3.6), we have (4.1). In the case of the multilayer parceptron, we have (4.2) from (4.1)and (3.3).

Now we give an asymptotic relation which shows how the regularization term modifies the maximum likelihood estimator $\hat{w}_0$ into $\hat{w}_\lambda$. From (3.6), by putting

$$\hat{w}_\lambda = \hat{w}_0 + \varepsilon, \tag{4.4}$$

we have

$$\nabla_w \log \pi(\hat{w}_0 + \varepsilon, \lambda) + \sum \nabla_w l(y_t | x_t; \hat{w}_0 + \varepsilon) = 0.$$

By expansion,

$$\sum \nabla_w l(y_t|x_t;\hat{w}_0+\varepsilon) = \sum \nabla_w \nabla_w l(y_t|x_t;\hat{w}_0)\varepsilon.$$

Let us put

$$A = \frac{1}{T}\sum \nabla_w \nabla_w l(y_t|x_t;\hat{w}_0),$$

and we have

$$\varepsilon = \frac{A^{-1}}{T}\nabla_w \log \pi(\hat{w}_0).$$

It should be noted that A converges to $K = E[\nabla_w \nabla_w l(y|x;\hat{w}_0)]$. In the case of the multilayer perceptron, we have

$$\hat{w}_\lambda = \hat{w}_0 - \frac{\lambda}{T}K^{-1}\nabla_w r(\hat{w}_0), \tag{4.5}$$

which is a shrinkage estimator of the form

$$\hat{w}_\lambda = \left(I - \frac{\lambda}{T}K^{-1}\right)\hat{w}_0. \tag{4.6}$$

It should be remarked that K is not the identity matrix in general. This is the Fisher information matrix in the realizable case of

$$q(y|x) = p(y|x;w^*).$$

5 Evaluation of the optimal λ by NIC

We have obtained the model selection criterion and adaptive selection of λ based on data from the Bayesian standpoint at the same time. However, it is still not certain how good they are. So we search for the method of minimizing the generalization error directly, and compare the results. This has already been obtained by Murata et al. (1997) which we recapitulate shortly.

The generalization error $E_{\rm gen}(w)$ of the network belonging to M with parameter w is given by

$$E^M_{\rm gen}(w) = -E[\log p(y|x;w)]. \tag{5.1}$$

This does not include the regularization term. This can be written as

$$E^M_{\rm gen}(w) = \frac{1}{2}E[|y - f(x;w)|^2] \tag{5.2}$$

for (1.3). The purpose of introducing the regularization term is to decrease the generalization error preventing from overfitting. In order to see its effect, we calculate the generalization error of the trained network with parameter λ.

Let w^* be the minimizer of $E^M_{\rm gen}(w)$,

$$w^* = \arg\min E^M_{\rm gen}(w). \tag{5.3}$$

We then have

$$E^M_{\rm gen}(\hat{w}_\lambda) = E^M_{\rm gen}(w^*) + \frac{1}{2}(\hat{w}_\lambda - w^*)^T K(\hat{w}_\lambda - w^*) \tag{5.4}$$

where higher order terms are neglected. Let e be the deviation error of the maximum likelihood estimator $\hat{w}_0$ from the optimal one,

$$\hat{w}_0 = w^* + e. \tag{5.5}$$

The asymptotic behavior of the maximum likelihood estimator is well known in statistics. It is asymptotically subject to the normal distribution. It has a bias of order $1/T$, and its variance is given by matrix $(1/T)K^{-1}GK^{-1}$, where

$$G = E[\nabla_w \log p(y|x;w^*)\nabla_w \log p(y|x;w^*)].$$

The bias term b is given by

$$E[\hat{w}_0] = w^* + \frac{b}{T} + O\left(\frac{1}{T^2}\right), \tag{5.6}$$

where the bias vector $b = (b^i)$ is calculated as

$$b^i = \sum_{k,l,m} k^{ik}k^{lm}s_{klm} + \frac{1}{2}\sum_{k,l,r,m,s} k^{ik}k^{lm}k^{rs}t_{klr}g_{ms}, \tag{5.7}$$

where (k^{ik}) is the inverse of matrix K, $G = (g_{ms})$, and

$$s_{klm} = E\left[\frac{\partial^2}{\partial w^k \partial w^l}\log p(y|x;w)\frac{\partial}{\partial w^m}\log p(y|x;w)\right] \tag{5.8}$$

$$t_{klr} = E\left[\frac{\partial^3}{\partial w^k \partial w^l \partial w^r}\log p(y|x;w)\right] \tag{5.9}$$

Taking account of (4.6), we have

$$E^M_{\rm gen}(\hat{w}_\lambda) = E^M_{\rm gen}(w^*) + \frac{1}{2}\left\{e - \frac{\lambda}{T}K^{-1}\nabla_w r(\hat{w}_\lambda)\right\}^T K\left\{e - \frac{\lambda}{T}K^{-1}\nabla_w r(\hat{w}_\lambda)\right\}.$$

Now we expand

$$\nabla_w r(\hat{w}_\lambda) = \nabla_w r(w^*) + Re - \frac{\lambda}{T} RK^{-1} \nabla_w r(w^*) \tag{5.10}$$

where

$$R = \nabla_w \nabla_w r(w^*). \tag{5.11}$$

We also decompose e as

$$e = \tilde{e} + \frac{b}{T} \tag{5.12}$$

where $\tilde{e}$ is non-biased. Then, the term

$$\hat{w}_\lambda - w^* = e - \frac{\lambda}{T} K^{-1} \nabla_w r(\hat{w}_\lambda) \tag{5.13}$$

is decomposed as the sum of the fluctuating term and bias term,

$$\hat{w}_\lambda - w^* = \left(I - \frac{\lambda}{T} K^{-1} R \right) \tilde{e} - \frac{\lambda}{T} K^{-1} \nabla_w r + \frac{b}{T}, \tag{5.14}$$

where higher order terms are neglected. Neglecting higher order terms again, the expectation of $E^M_{\rm gen}(\hat{w}_\lambda)$ is written as

$$E[E^M_{\rm gen}(\hat{w}_\lambda)] = E^M_{\rm gen}(w^*) + \frac{1}{2} E\left[\tilde{e}^T \left(I - \frac{\lambda}{T} K^{-1} R \right)^T K \left(I - \frac{\lambda}{T} K^{-1} R \right) \tilde{e} \right] + \frac{\lambda^2}{2T^2} (\nabla_w r)^T K^{-1} \nabla_w r + \frac{\lambda}{T^2} b^T \nabla_w r.$$

Taking account of

$$E[\tilde{e}\tilde{e}^T] = \frac{1}{T} G,$$

we have the following theorem.

Theorem 3. The optimal λ that minimizes the expected generalization error (NIC) is given by

$$\lambda_{\rm opt} = \frac{{\rm tr}(RK^{-1}GK^{-1}) + b^T \nabla_w r}{(\nabla_w r)^T K^{-1} (\nabla_w r)}. \tag{5.15}$$

When r is Gaussian,

$$\lambda_{\rm opt} \simeq \frac{{\rm tr}(K^{-1}GK^{-1} + b^T w^*)}{(w^*)^T K^{-1} w^*} \simeq \frac{{\rm tr}(K^{-1}GK^{-1}) + \hat{b}^T \hat{w}_0}{(\hat{w}_0)^T K^{-1} \hat{w}_0}. \tag{5.16}$$

This is the formula obtained by Murata.

Except for the case of $K = G =$ identity matrix, this is different from the Bayesian principle. The Bayesian $\lambda_{\rm opt}$ does not depend on the structure of the underlying neural network, whereas $\lambda_{\rm opt}$ derived from the least generalization error is sensitive to the underlying structures of K and G. This is the λ that directly minimizes the generalization error. So it is plausible that this λ gives a better performance of minimizing the generalization error.

6 Appendix A

Proof of Theorem 1.

We expand (3.6) at $\hat{w}_\lambda$, giving

$$\begin{aligned} l(D_T, w, \lambda) &= l(D_T, \hat{w}_\lambda, \lambda) + \frac{1}{2}\{\sum \nabla_w \nabla_w l(y_t|x_t; \hat{w}_\lambda) \\ &\quad + \nabla_w \nabla_w \log \pi(w; \lambda)\}(w - \hat{w}_\lambda)^2 \\ &\quad + \text{higher order terms}, \end{aligned}$$

because

$$\nabla_w l(D_T, \hat{w}_\lambda, \lambda) = 0.$$

Here, we used the abbreviated notation such as $\nabla_w \nabla_w l(w - w_\lambda)^2$ which implies the quadratic form

$$(w - \hat{w}_\lambda)^T \nabla_w \nabla_w l(w - \hat{w}_\lambda)$$

since $\nabla_w \nabla_w l$ is a matrix and $w - \hat{w}_\lambda$ is a vector. Therefore,

$$p(D_T, \lambda) = \exp\{l(D_T, \hat{w}_\lambda, \lambda)\} \int \exp\left\{-\frac{T}{2} A (w - \hat{w}_\lambda)^2\right\} dw,$$

where

$$-A = \frac{1}{T} \sum \nabla_w \nabla_w l(y_t|x_t; \hat{w}_\lambda) + \frac{1}{T} \nabla_w \nabla_w \log \pi.$$

The matrix A converges to

$$K = E[\nabla_w \nabla_w l(y|x; w^*)].$$

Therefore, by integration, we have (3.10) by neglecting higher order terms.

7 Conclusions

The present paper studies the effect of the regularization constant in neural learning by asymptotic statistical analysis. The Bayesian method is used to determine the optimal constant, and its relation to MDL is elucidated. The generalization error is analyzed when learning takes place from the Bayesian standpoint. It is then shown that the NIC criterion is useful for obtaining the optimal regularization constant, given a better generalization error.

References

[1] H. Akaike (1974) A new look at statistical model identification, *IEEE Transactions on Automatic Control*, **19**, 716–723.

[2] G. Brake, J.N. Kok and P.M.B. Vitányi (1995) Model Selection for Neural Networks: Comparing MDL and NIC,

[3] D.J.C. McKay (1992a) Bayesian interpolation, *Neural Computation*, **4**, 415–447.

[4] J.E. Moody (1992) The effective number of parameters: an analysis of generalization and regularization in nonlinear learning systems, in *NIPS4*, pp.847–854.

[5] N. Murata, S. Yoshizawa and S. Amari (1994) Network information criterion — determining the number of hidden units for artificial nerual network models, *IEEE Transactions on Neural Networks*, **5**, 865–872.

[6] T. Poggio and F. Girosi (1990) Regularization algorithms for learning that are equivalent to multilayer networks, *Science*, **247**, 978–982.

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

[8] J.Rissanen (1989) *Stochastic Complexity in Statistical Inquiry*, Singapore: World Scientific Publishing Co.