Regularization learning and early stopping in linear networks

KATSUYUKI HAGIWARA KAZUHIRO KUNO

Faculty of Physics Engineering, Mie University
1515 Kamihama, Tsu, 514-8507, Japan, {hagi,kuno}@phen.mie-u.ac.jp

1 Introduction

Generally, learning is performed so as to minimize the sum of squared errors between network outputs and training data. Unfortunately, this procedure does not necessarily give us a network with good generalization ability when the number of connection weights are relatively large. In such situation, overfitting to the training data occurs. To overcome this problem, there are several approaches such as regularization learning[6][11][12][16] and early stopping[2][15]. It has been suggested that these two methods are closely related[4][5][8][14]. In this article, we firstly give an unified interpretation for the relationship between two methods through the analysis of linear networks in the context of statistical regression ; i.e. linear regression model. On the other hand, several theoretical works have been done on the optimal regularization parameter[6][11][12][16] and the optimal stopping time[2][15]. Here, we also consider the problem from the unified viewpoint mentioned above. This analysis enables us to understand the statistical meaning of the optimality. Then, the estimates of the optimal regularization parameter and the optimal stopping time are present and those are examined by simple numerical simulations. Moreover, for the choice of regularization parameter, the relationship between the Bayesian framework and the generalization error minimization framework is discussed.

2 Linear least squares estimation

2.1 Linear networks

Consider a linear network with K inputs and one output, whose output for input $\boldsymbol{x} = (x_1, \dots, x_K) \in \mathbf{R}^K$ is

$$f_{\boldsymbol{w}}(\boldsymbol{x}) = \sum_{k=1}^{K} w_k x_k, \tag{1}$$

where $\boldsymbol{w} = (w_1, \dots, w_K) \in \mathbf{R}^K$ is a K-dimensional weight vector. The training data is denoted by $D = \{(\boldsymbol{x}_m, y_m) : \boldsymbol{x}_m = (x_{m,1}, \dots, x_{m,K}) \in \mathbf{R}^K, y_m \in \mathbf{R}, 1 \le m \le M\}$, in which each output data y_m is assumed to be generated according to $y_m = f_{\boldsymbol{w}^*}(\boldsymbol{x}_m) + \xi_m$, $f_{\boldsymbol{w}^*}(\boldsymbol{x}_m) = \sum_{k=1}^{K^*} w_k^* x_{m,k}$, where $\{\xi_m : 1 \le m \le M\}$ is an i.i.d. noise sequence from a probability distribution with mean 0 and variance σ^2, $\boldsymbol{w}^* = (w_1^*, \dots, w_{K^*}^*) \in \mathbf{R}^{K^*}$ is a true weight vector and K^* is the true number of weights ; i.e. $w_k^* = 0$ for $k > K^*$. Here, we assume that $\boldsymbol{x}_m$, $m = 1, \dots, M$ are not stochastic. For simplicity, we also assume that $K \ge K^*$; i.e. realizable or overrealizable scenario. As well known, the above setting of problem is that of linear regression. In this framework, we usually use a set of fixed functions $\{G_k : k = 1, \dots, K\}$ and set $x_k = G_k(\boldsymbol{u})$ for an input vector $\boldsymbol{u}$ in (1).

In the following, we use the matrix notion such as $\boldsymbol{y} := (y_1 \cdots y_M)'$, $\boldsymbol{w} := (w_1 \cdots w_K)'$, $\boldsymbol{f_w} := (f_{\boldsymbol{w}}(\boldsymbol{x}_1) \cdots f_{\boldsymbol{w}}(\boldsymbol{x}_M))' = X\boldsymbol{w}$, where X is a $M \times K$ matrix whose (m, k) element is given by $x_{m,k}$ and $'$ denotes the transpose of a matrix. Let us define $\|\boldsymbol{a}\|^2 = \sum_{m=1}^{M} a_m^2$ for an M-dimensional vector $\boldsymbol{a} = (a_1 \cdots a_M)'$. We assume that $\|\boldsymbol{f}_{\boldsymbol{w}^*}\|^2 = O(M)$ in our article. Hereafter, we also assume that $(X'X)$, which is called the design matrix, is not singular. The design matrix plays an important role in our analysis.

2.2 Linear least squares estimator

For a linear network, we usually estimate $\boldsymbol{w}$ so as to minimize the sum of squared errors :

$$E(\boldsymbol{w}) \;\; := \;\; \|\boldsymbol{y} - \boldsymbol{f_w}\|^2. \tag{2}$$

0-7695-0619-4/00 $10.00 © 2000 IEEE

The minimizing weight vector of $E(\boldsymbol{w})$ is the least squares estimator and is given by

$$\widehat{\boldsymbol{w}}_{\text{lse}} = (X'X)^{-1}X'\boldsymbol{y}. \tag{3}$$

For the linear least squares estimator in realizable/overrealizable scenario, it is well known that

$$\begin{aligned} E_{\boldsymbol{y}}\{\widehat{\boldsymbol{w}}_{\text{lse}}\} &= \boldsymbol{w}^*, &(4)\\ V_{\boldsymbol{y}}\{\widehat{\boldsymbol{w}}_{\text{lse}}\} &= \sigma^2(X'X)^{-1}, &(5)\end{aligned}$$

where $E_{\boldsymbol{y}}$ and $V_{\boldsymbol{y}}$ stand for the expectation and covariance matrix with respect to the joint probability distribution of $y_1, \ldots, y_M$. The equation (4) means that the least squares estimator is the unbiased estimator.

For an estimator $\widehat{\boldsymbol{w}}$, the expected generalization error is defined as

$$E_{\text{gen}}(K, M) := E_{\boldsymbol{y}}\left\{E_{\boldsymbol{z}}\left\{\frac{1}{M}\|\boldsymbol{z} - \boldsymbol{f}_{\widehat{\boldsymbol{w}}}\|^2\right\}\right\}, \tag{6}$$

where $\boldsymbol{z} = (z_1 \ \cdots \ z_M)'$, in which each element z_m has the same probability distribution with y_m and $z_1, \ldots, z_M, y_1, \ldots, y_M$ are independent. For the least squares estimator, we can easily obtain

$$E_{\text{gen}}^{\text{lse}}(K, M) = \sigma^2 + \frac{K\sigma^2}{M}, \tag{7}$$

by using (4) and (5) (e.g. [3]).

3 Regularization learning

3.1 Regularization learning

The regularized estimator is obtained by minimizing the cost function defined by

$$C(\boldsymbol{w}) = E(\boldsymbol{w}) + \lambda R(\boldsymbol{w}), \tag{8}$$

where $E(\boldsymbol{w})$ is defined in (2), $R(\boldsymbol{w})$ is a regularization term and $\lambda \in [0, \infty)$ is a regularization parameter. For the analysis, we introduce the following regularization term.

$$R(\boldsymbol{w}) := \boldsymbol{w}'(X'X)\boldsymbol{w} = \|\boldsymbol{f}_{\boldsymbol{w}}\|^2. \tag{9}$$

3.2 Statistical properties of regularized estimator

In our case, the regularized estimator $\widehat{\boldsymbol{w}}_{\text{reg}}$, which is the minimizing weight vector of $C(\boldsymbol{w})$, is given by

$$\widehat{\boldsymbol{w}}_{\text{reg}} = \gamma(X'X)^{-1}X'\boldsymbol{y}, \tag{10}$$

where $\gamma = 1/(1+\lambda)$. Because we have

$$\widehat{\boldsymbol{w}}_{\text{reg}} = \gamma\widehat{\boldsymbol{w}}_{\text{lse}}$$

by (3) and $0 < \gamma \leq 1$, the regularized estimator $\widehat{\boldsymbol{w}}_{\text{reg}}$ is a shrinkage estimator[7]. For the regularized estimator, it is easily shown that

$$\begin{aligned} \overline{\boldsymbol{w}} &:= E_{\boldsymbol{y}}\{\widehat{\boldsymbol{w}}_{\text{reg}}\} = \gamma\boldsymbol{w}^*, &(11)\\ V_{\boldsymbol{y}}\{\widehat{\boldsymbol{w}}_{\text{reg}}\} &= \sigma^2\gamma^2(X'X)^{-1}. &(12)\end{aligned}$$

(11) is rewritten as

$$E_{\boldsymbol{y}}\{\widehat{\boldsymbol{w}}_{\text{reg}}\} = \boldsymbol{w}^* - \frac{1}{1+1/\lambda}\boldsymbol{w}^*.$$

Hence, the regularized estimator is biased even in the realizable/overrealizable scenario. When $\lambda = 0$, $\widehat{\boldsymbol{w}}_{\text{reg}} = \widehat{\boldsymbol{w}}_{\text{lse}}$ and the bias is 0.

3.3 Bias/variance dilemma in the expected generalization error

For the regularized estimator $\widehat{\boldsymbol{w}}_{\text{reg}}$, the expected generalization error is shown to be given by

$$E_{\text{gen}}^{\text{reg}}(\lambda, K, M) = \sigma^2 + B(\lambda) + V(\lambda), \tag{13}$$

where

$$B(\lambda) = \frac{1}{M}\|\boldsymbol{w}^* - \overline{\boldsymbol{w}}\|_{X'X}^2, \quad V(\lambda) = \frac{1}{M}E_{\boldsymbol{y}}\left\{\|\widehat{\boldsymbol{w}}_{\text{reg}} - \overline{\boldsymbol{w}}\|_{X'X}^2\right\},$$

where $\|\boldsymbol{a}\|_{X'X}^2 = \boldsymbol{a}'X'X\boldsymbol{a}$ for some vector $\boldsymbol{a}$. Because $B(\lambda)$ is caused by deviation between the true weight vector and the expectation of the regularized estimator, it represents the bias. On the other hand, because $V(\lambda)$ arises from fluctuation of the regularized estimator around its expectation, it represents the variance. Therefore, the expected generalization error is decomposed into the bias and variance even in the realizable/overrealizable scenario. Here, it is easy to show that

$$B(\lambda) = \frac{1}{M}(\gamma - 1)^2 \|\boldsymbol{f}_{\boldsymbol{w}^*}\|^2, \quad V(\lambda) = \frac{1}{M}\gamma^2\sigma^2 K,$$

by using (11) and (12). In case of $\lambda = 0$, $B(0) = 0$ and $V(0) = \sigma^2 K/M$ hold, which corresponds to the case of the least squares estimator. When $\lambda \to \infty$, $B(\infty) = \|\boldsymbol{f}_{\boldsymbol{w}^*}\|^2/M$ and $V(\infty) = 0$ hold, which implies that the network output is 0 regardless of the true function. These facts say that the regularization parameter controls the balance between the bias and variance in the expected generalization error. Thus, the optimal regularization parameter, which minimize the expected generalization error, is given by solving the bias/variance dilemma. Note that this bias/variance dilemma is additional one to the well-known (intrinsic) bias/variance dilemma argued in [5][9]. The bias in our analysis arises from estimation procedures and exists even in the realizable/overrealizable scenario. The optimal regularization parameter is shown to be given by

$$\lambda_{\text{opt}} = \frac{\sigma^2 K}{\|\boldsymbol{f}_{\boldsymbol{w}^*}\|^2}. \tag{14}$$

The difference between the expected generalization error for the least squares estimator and for the optimal regularized estimator is given by $d_{\text{reg}}(K, M) = \frac{1}{M}\sigma^2 K \frac{\lambda_{\text{opt}}}{1+\lambda_{\text{opt}}}$. Because $\|\boldsymbol{f}_{\boldsymbol{w}^*}\|^2 = O(M)$ by the definition, $\lambda_{\text{opt}} = O(1/M)$ for fixed K. Thus, $d_{\text{reg}}(K, M) = O(1/M^2)$ for fixed K. This order is consistent with the result in [12], which solved the problem in more general case. Thus, the optimal regularized estimator dominates the least squares estimator at the order of $1/M^2$. But, this implies that the effect of the regularization learning on the generalization error is negligible when M is large and K is small. On the other hand, it is easily found that the effect of the regularized estimator will be essential in the situation where K is large because $d_{\text{reg}}(K, M) = O(K)$ for fixed M. The results suggest that the regularization technique may improve the generalization ability when the number of data is small or the number of weights is large relative to the number of data.

4 Overtraining and early stopping

4.1 Learning rule

Here, we consider the following update rule.

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \eta(X'X)^{-1}\frac{\partial E(\boldsymbol{w}(t))}{\partial \boldsymbol{w}(t)}, \tag{15}$$

where $\eta(> 0)$ is a learning rate and $E(\boldsymbol{w}(t))$ is the error function defined in (2) with $\boldsymbol{w}$ at t. It is easily found that this is a special case of natural gradient method[1]. If we assume that η is very small, the approximation by continuous dynamical system yields

$$\boldsymbol{w}(t) = (1 - \alpha)\widehat{\boldsymbol{w}}_{\text{lse}} + \alpha\boldsymbol{w}(0), \tag{16}$$

where $\alpha = \alpha(t) := e^{-2\eta t}$. Therefore, $\boldsymbol{w}(t)$ approaches $\widehat{\boldsymbol{w}}_{\text{lse}}$ linearly from $\boldsymbol{w}(0)$, where $\boldsymbol{w}(\infty) = \widehat{\boldsymbol{w}}_{\text{lse}}$. Note that if we set $\boldsymbol{w}(0) = \boldsymbol{0}$ then it is easily found that $\boldsymbol{w}(t)$ is a shrinkage estimator because $0 < \alpha \leq 1$.

4.2 Statistical properties of early stopping estimator

For simplicity, $\boldsymbol{w}(0)$ is assumed to be fixed at any sampling. Because $\widehat{\boldsymbol{w}}_{\text{lse}}$ is the unbiased estimator of $\boldsymbol{w}^*$, we have

$$\overline{\boldsymbol{w}}(t) := E_{\boldsymbol{y}}\{\boldsymbol{w}(t)\} = (1 - \alpha)\boldsymbol{w}^* + \alpha\boldsymbol{w}(0), \tag{17}$$

$$V_{\boldsymbol{y}}\{\boldsymbol{w}(t)\} = (1 - \alpha)^2\sigma^2(X'X)^{-1} + \alpha^2\boldsymbol{w}(0)\boldsymbol{w}(0)', \tag{18}$$

by using (4), (5) and (16). (17) tells us that the obtained weight vector $\boldsymbol{w}(t)$ is biased if we stop the learning before convergence. The bias is vanished if the learning is not stopped early.

4.3 Bias/variance dilemma in the expected generalization error

The expected generalization error at each t is easily shown to be

$$E^{\text{es}}_{\text{gen}}(t) = \sigma^2 + B(t) + V(t), \tag{19}$$

where

$$B(t) = \frac{1}{M}\|\overline{\boldsymbol{w}}(t) - \boldsymbol{w}^*\|^2_{X'X}, \quad V(t) = \frac{1}{M}E_{\boldsymbol{y}}\left\{\|\boldsymbol{w}(t) - \overline{\boldsymbol{w}}(t)\|^2_{X'X}\right\}.$$

As in the regularization learning, again, the expected generalization error is decomposed into the bias $B(t)$ and variance $V(t)$. By using (16), (17) and (18), those are calculated as

$$B(t) = \frac{\alpha^2}{M}\|\boldsymbol{f}(0) - \boldsymbol{f}_*\|^2, \quad V(t) = (1-\alpha)^2\frac{K\sigma^2}{M},$$

where $\boldsymbol{f}_* := \boldsymbol{f}_{\boldsymbol{w}^*}$ and $\boldsymbol{f}(0) := \boldsymbol{f}_{\boldsymbol{w}(0)}$. It is easily found that the regularization parameter and the stopping time are linked by the relation $\alpha(t) = 1 - \gamma$. In the above, $B(\infty) = 0$ and $V(\infty) = K\sigma^2$, which is the case of without stopping and corresponds to the case of the least squares estimator. On the other hand, $V(0) = 0$ while it leads to the large bias $B(0) = \|\boldsymbol{f}(0) - \boldsymbol{f}_*\|^2$. Thus, again, the optimal stopping time is determined by solving the bias/variance dilemma and we obtain

$$\alpha_{\text{opt}} = \frac{K\sigma^2}{K\sigma^2 + \|\boldsymbol{f}(0) - \boldsymbol{f}_*\|^2}. \tag{20}$$

Thus, by the definition of $\alpha = \alpha(t)$, we have

$$t_{\text{opt}} = -\frac{2}{\eta}\log\alpha_{\text{opt}} \tag{21}$$

as the optimal stopping time according to the expected generalization error. Inserting (20) into (19), we have

$$E^{\text{es}}_{\text{gen}}(\alpha_{\text{opt}}, K, M) = \sigma^2 + (1 - \alpha_{\text{opt}})\frac{K\sigma^2}{M}. \tag{22}$$

Therefore, the optimally stopped weight vector reduces the generalization error by the factor of $d_{\text{es}}(K, M) = \alpha_{\text{opt}}K\sigma^2/M$ compared with the least squares estimator. By (20), $d_{\text{es}}(K, M) = O(1/M^2)$ for fixed K if $\boldsymbol{f}(0) \neq \boldsymbol{f}_*$. The effectiveness at order $1/M^2$ is consistent with the results in [2][15]. As in the regularization learning, again, the results tell us that the advantage of early stopping is negligible when M is large and K is small, but it may essentially improve the generalization error when K is large relative to M.

5 Discussions

5.1 On the meaning of the regularization learning and early stopping

The importance of the regularization learning and early stopping is explained from a statistical point of view. It is well known that the least squares estimator is the best linear unbiased estimator, which is known as the Gauss-Markov theorem ; e.g. [13]. Moreover, it can be shown that the least squares estimator gives the minimum expected generalization error among linear unbiased estimators. Thus, in linear unbiased estimators, there is no good estimator according to the expected generalization error other than the least squares estimator. However, this does not hold for biased estimators. In this meaning, the introduction of the regularizer or early stopping can be regarded as one of the attempts for searching a good estimator in a collection of biased estimators. More specifically, as seen in the preceding sections, both of the regularization learning and early stopping yield shrinkage estimators. Although [7] defined the amount of shrinkage as the least squares slope of validation sample on the predictor, the shrinkage naturally arises as a solution to the estimation procedures. As described in the above, the best biased estimator or the best amount of shrinkage is determined by minimizing the expected generalization error, which have a meaning of the solution to the bias/variance dilemma in the expected generalization error.

On the other hand, in the Bayesian framework, our regularizer corresponds to the introduction of the Gaussian prior with covariance matrix $(X'X)^{-1}$. For our regularizer (9), the effective choice of the regularization parameter by means of the method of integrating over hyperparameters[6][16] is given by

$$\lambda(\boldsymbol{w}) = \frac{K\sigma^2(\boldsymbol{w})}{\|\boldsymbol{f}_{\boldsymbol{w}}\|^2}, \tag{23}$$

where $\sigma^2(\boldsymbol{w})$ is defined as

$$\sigma^2(\boldsymbol{w}) := \frac{1}{M}\sum_{m=1}^{M}(y_m - f_{\boldsymbol{w}}(\boldsymbol{x}_m))^2. \tag{24}$$

Therefore, if we set $\boldsymbol{w} = \boldsymbol{w}^*$ in (23) then $\lambda(\boldsymbol{w}^*) = \lambda_{\rm opt}$, which is the optimal regularization parameter obtained by minizing the expected generalization error.

5.2 On the estimation of the optimal regularization parameter

Because the optimal regularization parameter obtained in (14) can not be calculated in real world problems, we must estimate that based on the given training data. Here, we give a natural estimate of the optimal regularization parameter and examine the estimate through a simple numerical simulation.

The input-output data is generated by the following manner. For $z_m \in I \subset \mathbf{R}$,

$$x_{k,m} = \exp\left(\frac{1}{2\tau}(z_m - \mu_k)^2\right), \quad y_m = \sum_{k=1}^{K^*} w_k^* x_{k,m} + \xi_m,$$

where $\xi_1, \ldots, \xi_M$ are independent samples according to $N(0, \sigma^2)$ and we fixed τ and μ_k, $k = 1, \ldots, K$. Here, we set $I = [-5, 5]$, $\sigma^2 = 1.0$, $M = 50$ and $K^* = 1$. For the given training data, we train linear networks with $K = 1 \sim 7$. Thus, we consider the realizable/overrealizable scenario. In our simulation, we first generate 1000 sets of training data. For each set of training data, we train a linear network under the regularized cost function (8), (2) and (9) with $\lambda = \widehat{\lambda}_{\rm opt}$ defined below.

$$\widehat{\lambda}_{\rm opt} \quad = \quad \frac{\sigma^2(\widehat{\boldsymbol{w}}_{\rm lse})K}{\|\boldsymbol{f}_{\widehat{\boldsymbol{w}}_{\rm lse}}\|^2}, \tag{25}$$

where $\sigma^2(\widehat{\boldsymbol{w}}_{\rm lse})$ is defined by (24) with $\boldsymbol{w} = \widehat{\boldsymbol{w}}_{\rm lse}$. This is closely related to the proposed estimate in [12] for weight decay. The generalization error is estimated by using 1000 sets of M new samples. Thus, we have the estimate of the generalization error for each trained network. Then , the expected generalization error is estimated by the average of the estimated generalization error.

Figure 1 shows the result on the generalization error at each number of weights K. In the figure, we show the estimate of the expected generalization error with $\widehat{\lambda}_{\rm opt}$ (open circle), the expected generalization errors with the least squares estimator (gray line) and with $\lambda = \lambda_{\rm opt}$ (solid line). In the figure, the generalization error with the estimated regularization parameter is larger than one with the optimal, but, smaller than one with the least squares estimator. These results suggest that there is a chance to construct a network with better generalization ability by using the regularization technique even in practical situations. Note that if we set $\boldsymbol{w} = \widehat{\boldsymbol{w}}_{\rm lse}$ in (23), $\lambda(\widehat{\boldsymbol{w}}_{\rm lse})$ is equal to $\widehat{\lambda}_{\rm opt}$; i.e. the effective Bayesian choice of the empirical regularization parameter is consistent with a natural estimate of the regularization parameter which minimizes the expected generalization error.

5.3 On the estimation of the optimal stopping time

For early stopping technique in practical applications, we again encounter the problem to estimate the optimal stopping time $t_{\rm opt}$. Here, we employ

$$\widehat{\alpha}_{\rm opt} = \frac{K\sigma^2(\widehat{\boldsymbol{w}}_{\rm lse})}{K\sigma^2(\widehat{\boldsymbol{w}}_{\rm lse}) + \|\boldsymbol{f}(0) - \boldsymbol{f}_{\widehat{\boldsymbol{w}}_{\rm lse}}\|^2}, \quad \text{thus} \quad \widehat{t}_{\rm opt} = -\frac{2}{\eta}\log\widehat{\alpha}_{\rm opt} \tag{26}$$

as a natural estimate of $t_{\rm opt}$. To examine this estimate, we have done a numerical simulation. The setting of the simulation is the same one as in the previous simulation of regularization learning. The simulation results are summarized in figure 2. In the figure, we show the expected generalization error for the least squares estimator given by (7) (gray line) ; i.e. without stopping, the optimal time (21) (sold line) and the estimate of the expected generalization error for the estimator obtained by stopping at the estimated optimal time (26)(open circle). The results tell us that the estimated stopping time works better than the least squares estimator. Although [2] proved the cross-validation early stopping works worse than the least squares estimator, the result shows that the early stopping may be effective if we choose an appropriate estimate of stopping time, especially in small sample situation.

6 Conclusions and future works

In this article, we present a unified statistical interpretation of regularization learning and early stopping for linear networks in the context of statistical regression ; i.e. linear regression model. Here, the regularization learning and the early stopping are shown to be equivalent with the use of biased estimator, or more

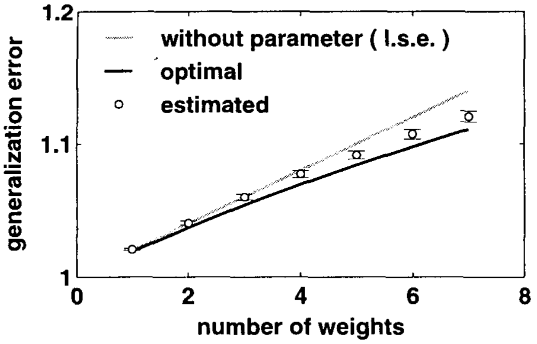

Figure 1. The result of a numerical simulation on regularization learning.

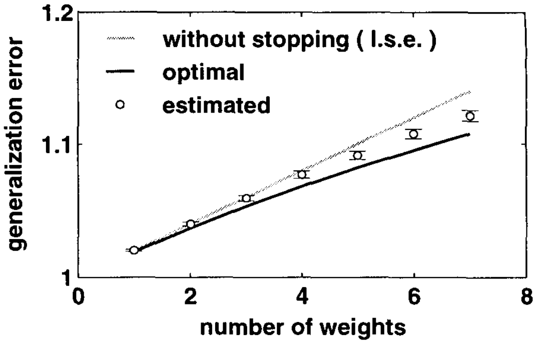

Figure 2. The result of a numerical simulation on early stopping.

specifically shrinkage estimator. We also showed that the optimal regularization parameter or the optimal stopping time according to the expected generalization error are obtained by solving the bias/variance dilemma, which is additional one to the well-known intrinsic bias/variance dilemma argued in [5][9]. Moreover, we gave estimates of the optimal regularization parameter and stopping time. The effectiveness of those estimates is shown by the numerical simulations. The theoretical analysis as in [7] on the estimates is left as a future work. Finally, the effective regularization parameter obtained in the Bayesian framework is linked to the optimal and empirically estimated regularization parameter in the generalization error minimization framework.

References

[1] Amari S. (1998). "Natural gradient works efficiently in learning", *Neural Computation*, **10**, 251–276.

[2] Amari S., Murata N., Müller K., Finke M., Yang H. H. (1997). "Asymptotic statistical theory of overtraining and cross-validation", *IEEE Trans. on Neural Networks*, **8**, 5, 985–996.

[3] Barron A. R. (1984). "Predicted squared error : a criterion for automatic model selection", In *Self-Organizing Methods in Modeling*, S. Farlow, ed., Marcel Dekker, New York, 87–103.

[4] Bishop C. M. (1995). "Neural networks for pattern recognition", Oxford University Press.

[5] Breiman L. (1998). "Bias-variance, regularization, instability and stabilization", In *Neural Networks and Machine Learning*, Bishop C. M. ed., Springer, 27–55.

[6] Buntine W.L., Weigend A.S. (1991). "Bayesian Back-Propagation", *Complex Syst.*, **5**, 877–922.

[7] Copas J. B. (1983). "Regression, prediction and shrinkage", *J. R. Statist. Soc. B*, **45**, 3, 311–354.

[8] Cataltepe Z., Abu-Mostafa Y. S., Magdon-Ismail M. (1999). "No free lunch for early stopping", *Neural Computation*, **11**, 995–1009.

[9] Geman S., Bienenstock E., Doursat R. (1992). "Neural Networks and the Bias/Variance Dilemma", *Neural Computation*, **4**, 1–58.

[10] Goutte C., Hansen L. K. (1997). "Regularization with a Pruning Prior ", *Neural Networks*, **10**, 6, 1053–1059.

[11] MacKay D. J. C. (1992). "Bayesian Interpolation", *Neural Computation*, **4**, 415–447.

[12] Murata N. (1998). "Bias of Estimators and Regularization Terms", In *1998 Workshop on Information-Based Induction Sciences*, 87–94.

[13] Rao C. R (1973). "Linear Statistical Inference and Its Applications", *John Wiley & Sons.*

[14] Ripley B. D. (1998). "Statistical theories of model fitting", In *Neural Networks and Machine Learning*, Bishop C. M. ed., Springer, 3–22.

[15] Wang C., Venkatesh S. S., Judd J. S. (1995). "Optimal stopping and effective machine complexity in learning", *Advances in Neural Information Processing Systems 6*, 303–310.

[16] Williams P. M. (1995). "Bayesian Regularization and Pruning Using a Laplace Prior", *Neural Computation*, **7**, 117–143.