Pruning from Adaptive Regularization

Lars Kai Hansen and Carl Edward Rasmussen
CONNECT, Electronics Institute
Technical University of Denmark, B 349
DK-2800 Lyngby. Denmark
Emails: lars@eiffel.ei.dth.dk & carl@cs.toronto.edu

January 12, 1994

Abstract

Inspired by the recent upsurge of interest in Bayesian methods we consider adaptive regularization. A generalization based scheme for adaptation of regularization parameters is introduced and compared to Bayesian regularization. We show that pruning arises naturally within both adaptive regularization schemes. As model example we have chosen the simplest possible: estimating the mean of a random variable with known variance. Marked similarities are found between the two methods in that they both involve a "noise limit", below which they regularize with infinite weight decay, *i.e.*, they *prune*. However, pruning is not always beneficial. We show explicitly that both methods in some cases may increase the generalization error. This corresponds to situations where the underlying assumptions of the regularizer are poorly matched to the environment.

1 Introduction

We believe in Ockham's Razor: the generalization error of a model with estimated parameters is decreased by constraining capacity to a minimum needed for capturing the rule, see *e.g.* (Thodberg, 1991; Solla, 1992). However, this minimum may be hard to define for non-linear noisy systems where the rule is ill-defined.

Pruning is a popular tool for reducing model capacity and pruning schemes have been successfully applied to layered neural networks (Le Cun et al., 1990; Thodberg, 1991; Svarer et al., 1993). While pruning is a discrete decision process, regularization introduces *soft* constraints such as weight decay (Moody, 1991). A common feature of these techniques is the need for control parameters: stop criteria for pruning and weight decays for regularization. In (Svarer et al., 1993) a statistical stop criterion was developed for pruning of networks for regression problems.

Recently, MacKay reviewed a Bayesian approach to adaptive regularization in the context of neural networks, demonstrating that the *evidence* based method can improve the generalization properties and he compared it to cross validation (MacKay, 1992a; MacKay, 1992b). Cross validation is known to be rather noisy; hence methods based on statistical arguments are recommended, see *e.g.* (Akaike, 1969; Moody, 1991; Hansen, 1993). In this presentation we define such an alternative approach for adaptive regularization, and we test it on a case of ultimate simplicity, namely that of estimating the mean of a Gaussian variable of known variance. Detailed insight can be obtained for this case.

In the course of comparing the two schemes, we have discovered a new feature of adaptive regularization. We find that both approaches involve a "noise limit", below which they regularize with infinite weight decay, *i.e.*, they *prune*. This finding unifies the pruning and regularization approaches to capacity control.

2 Specification of the toy problem

The problem is defined as follows: consider a student-teacher setup based on a *teacher* (with parameter $\tilde{w}$) providing N examples of the form

$$y_m = \tilde{w} + \nu_m \qquad \nu_m \sim \mathcal{N}(0, \sigma^2) \qquad m = 1, \ldots, N, \tag{1}$$

where the normal noise contributions are independent, have zero mean and common known variance σ^2. Based on the training data set, $D = \{y_m | m = 1, \ldots, N\}$, the task of the *student* (with parameter w) is to infer the mean: $\tilde{w}$. The measure of success for the student is the generalization error,

$$E_{\mathrm{G}}(w) = \int d\nu P(\nu)(\tilde{w} + \nu - w)^2 = (\tilde{w} - w)^2 + \sigma^2. \tag{2}$$

This is the expected error on a random new test example for the specific student weight as estimated on the given training set. In evaluating an estimation scheme, our measure will be the *average generalization error* obtained by averaging over all possible training sets of the given size N.

In the next three sections we consider three such estimators: the usual "empirical mean" obtained from maximum likelihood, an estimator based on MacKay's maximum evidence, and a novel estimator based on explicit minimization of the expected generalization error. In section 6 we compute the average generalization error of each of the schemes, as a function of the teacher parameter.

3 Maximum likelihood estimation

The likelihood of the student parameter associated with the training set D is

$$P(D|w) = \prod_{m=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_m - w)^2\right) \tag{3}$$

Maximizing this w.r.t. w, produces the standard maximum likelihood estimator, the *empirical mean*

$$w_{\mathrm{ML}}(D) = \frac{1}{N}\sum_{m=1}^{N} y_m. \tag{4}$$

This estimator is *unbiased*, *i.e.*, the average value, averaging over all possible training sets, is the true mean $\tilde{w}$.

For a specific training set, the generalization error is obtained using (2). However, as mentioned above, we are not so much interested in the value for the particular training set. Rather we are interested in the expected error, obtained by averaging over all possible training sets of size N. To compute this quantity, we note that $w_{\mathrm{ML}} \sim \mathcal{N}(\tilde{w}, \sigma^2/N)$, hence the average generalization error is,

$$\langle E_{\mathrm{G}}(w_{\mathrm{ML}})\rangle_N = \langle(\tilde{w} - w_{\mathrm{ML}})^2\rangle + \sigma^2 = \frac{\sigma^2}{N} + \sigma^2. \tag{5}$$

The first term is the *average excess error* made by the student due to the finite training set. The second term is due to the imperfectness of the data; there is noise in the test examples used for grading the student. The minimal error approached for large training sets is simply the noise level.

4 Bayesian regularization

A Bayesian scheme, recently suggested to the neural net community by MacKay, adopts two levels of inference. The first level of inference consists of estimating the teacher parameters from the data, conditioned on a parametrized *prior* distribution of the teacher mean value. The second level of inference, consists of estimating the parameters of the prior, with the purpose of maximizing the *evidence*.

The probability $P(w|D)$ of the parameter conditioned on the data can be obtained using Bayes' rule,

$$P(w|D) = \frac{P(D|w)P(w)}{P(D)}, \tag{6}$$

by specifying a prior distribution $P(w)$ of the parameter . We follow MacKay and employ a parameterized prior: $P(w) \equiv P(w|\alpha)$, with a parameter α playing the role of a "weight decay", to be determined at the second level of inference. The prior takes the form

$$P(w|\alpha) = \sqrt{\frac{\alpha}{2\pi}} \exp\left(-\frac{1}{2}\alpha w^2\right). \tag{7}$$

Using the likehood from eq. (3), we arrive at the *posterior* distribution

$$P(w|D,\alpha) = P(D|\alpha)^{-1}\sqrt{\frac{\alpha}{2\pi}} \exp\left(-\frac{1}{2}\alpha w^2\right) \prod_{m=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_m - w)^2\right). \tag{8}$$

To find the most probable teacher parameter, we maximize $P(w|D,\alpha)$ w.r.t. w, to get

$$w_{\mathrm{MP}}(D,\alpha) = \frac{N}{N+\alpha\sigma^2} w_{\mathrm{ML}}(D). \tag{9}$$

In the following we will suppress the explicit dependence on the training data.

The second level of inference is to estimate the regularization parameter, and we do this by again invoking Bayes' rule

$$P(\alpha|D) = \frac{P(D|\alpha)P(\alpha)}{P(D)}. \tag{10}$$

We assume the prior on α to be flat[1], indicating that we do not know what value α should take. Hence, the most probable regularization is obtained by maximizing the likelihood of it. MacKay dubbed this quantity the *evidence*

$$\begin{aligned} P(D|\alpha) &= \int dw P(D|w,\alpha)P(w|\alpha) & (11) \\ &= \int_{-\infty}^{\infty} dw \sqrt{\frac{\alpha}{2\pi}} \exp\left(-\frac{1}{2}\alpha w^2\right) \prod_{m=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_m - w)^2\right) \\ \Longrightarrow \log P(D|\alpha) &= \frac{1}{2}\log(\alpha/2\pi) - \frac{1}{2}\log\left(\frac{N+\alpha\sigma^2}{2\sigma^2}\right) + \frac{(N w_{\mathrm{ML}})^2}{2\sigma^2(N+\alpha\sigma^2)} + \mathrm{const}, & (12) \end{aligned}$$

[1] By flat we mean flat over $\log(\alpha)$, since α is a scale parameter.

where we have lumped all terms not depending on α in the constant. Maximizing the evidence w.r.t. α,

$$\alpha_{\mathrm{ML}} = \begin{cases} \dfrac{1}{w_{\mathrm{ML}}^2 - \sigma^2/N} & w_{\mathrm{ML}}^2 > \sigma^2/N \\ \infty & w_{\mathrm{ML}}^2 \leq \sigma^2/N \ , \end{cases} \tag{13}$$

which may be substituted into the expression for w_{MP} to find

$$w_{\mathrm{MP}} = \begin{cases} (1 - \dfrac{\sigma^2}{N w_{\mathrm{ML}}^2} w_{\mathrm{ML}} & w_{\mathrm{ML}}^2 > \sigma^2/N \\ 0 & w_{\mathrm{ML}}^2 \leq \sigma^2/N \ . \end{cases} \tag{14}$$

Hence there is a sharp "noise limit", below which the scheme tells us to "prune", that is to regularize with infinite strength. The noise limit has a straightforward interpretation: prune the parameter if the signal to noise ratio is less than unity. Note that the estimator is biased; it's mean value is not $\tilde{w}$. It is *consistent* since $w_{\mathrm{MP}} \rightarrow \tilde{w}$ for large samples, while the noise limit shrinks to zero. To illustrate the power of the method, we compute in section 6 the (training set) ensemble average of the generalization error.

In the above derivation we have used the "flat in $\log(\alpha)$ space" prior. This prior is improper, since it cannot be normalized. However this is not important for the evidence approximation. To see this we introduce limits: $\alpha \in [\alpha_1, \alpha_2]$. The value of w_{MP} will not be affected by these limits as long as $\alpha_{\mathrm{ML}} \in [\alpha_1, \alpha_2]$. Choosing $\alpha_1 \ll N/\sigma^2$ and $\alpha_2 \gg N/\sigma^2$ the effect of the limits will be negligible for all α. This is seen by introducing the limits in eq. (9). In particular we note that qualitative pruning still takes place: $|w_{\mathrm{MP}}| = |w_{\mathrm{ML}}| N/(N + \alpha_2 \sigma^2) \ll |w_{\mathrm{ML}}|$ for $w_{\mathrm{ML}}^2 < \sigma^2/N + 1/\alpha_2$.

The evidence framework is an approximation to the ideal Bayesian approach. Instead of using the maximum likelihood value of α, we should ideally integrate over the distribution of α, *i.e.*, evaluate $P(w) = \int P(w|\alpha)P(\alpha)d\alpha$ and then use the *posterior mean* for our prediction: $w_{\mathrm{PM}} = \int w P(w) dw$. In contrast to the evidence framework the ideal approach is quite sensitive to the upper limit α_2. Indeed if $\alpha_2 = \infty$ then $w_{\mathrm{PM}} = 0$ regardless of the value of w_{ML}. For a range of α_2 we find that w_{PM} shows a qualitative pruning behavior similar to that of w_{MP}.

5 Generalization error minimization

While the evidence is an interesting statistical quantity, it is not obvious what the relation is between maximizing the evidence and minimizing test error (MacKay, 1992a). Since the latter often is the basic modeling objective, we propose here to base the optimization of α on the expected generalization error, cf. eq. (2). A similar approach was mentioned in (Moody, 1991).

In order to be able to compare directly with the Bayesian approach, we use a regularized *least squares* procedure to find the student weight

$$E_{\rm T} = \frac{1}{2\sigma^2} \sum_{m=1}^{N} (y_m - w)^2 + \frac{1}{2}\alpha w^2, \tag{15}$$

Minimizing w.r.t. w, we recover (9)

$$w_{\rm G} = \frac{N}{N + \alpha\sigma^2} w_{\rm ML}. \tag{16}$$

Our aim is to minimize the average generalization error (averaged again w.r.t. training sets). To this end we note that the distribution of $w_{\rm G}(D, \alpha)$ can be computed using simple manipulations of random variables. Noting again that $w_{\rm ML} \sim \mathcal{N}(\tilde{w}, \sigma^2/N)$ we have $w_{\rm G} \sim \mathcal{N}\left(\frac{N}{N+\alpha\sigma^2}\tilde{w}, \frac{N}{(N+\alpha\sigma^2)^2}\sigma^2\right)$[2]. Consequently the averaged generalization error becomes

$$\langle E_{\rm G}(w_{\rm G}) \rangle_N = \left(\frac{\alpha\sigma^2}{N + \alpha\sigma^2} \tilde{w} \right)^2 + \left(1 + \frac{N}{(N + \alpha\sigma^2)^2} \right) \sigma^2. \tag{17}$$

You might be uncomfortable with the appearance of the unknown teacher parameter $\tilde{w}$, however, it will shortly be replaced by an estimate based on the data. Minimizing the generalization error (17) w.r.t. α, we find the simple expression

$$\alpha = \frac{1}{\tilde{w}^2}, \tag{18}$$

which is inserted in eq. (16), to obtain

$$w_{\rm G} = \frac{\tilde{w}^2}{\tilde{w}^2 + \sigma^2/N} w_{\rm ML}. \tag{19}$$

To proceed we need to insert an estimate of the teacher parameter $\tilde{w}$. We will consider two cases, firstly we try setting $\tilde{w} = w_{\rm ML}$ in eq. (19), hence

$$w_{\rm G1} = \frac{w_{\rm ML}^2}{w_{\rm ML}^2 + \sigma^2/N} w_{\rm ML}. \tag{20}$$

This estimator is biased, but consistent and it doesn't call for pruning even if the data are very noisy.

[2] Using $(\xi = a + b\psi) \wedge (\psi \sim \mathcal{N}(c, d^2)) \Rightarrow \xi \sim \mathcal{N}(a + bc, b^2 d^2)$

Secondly, being more brave, we could argue that the best estimate we have of $\tilde{w}$ is w_{G} and accordingly set $\tilde{w} = w_{\mathrm{G2}}$ in eq. (19), to obtain a self consistent equation[3] for w_{G2}

$$w_{\mathrm{G2}} = \frac{w_{\mathrm{G2}}^2}{w_{\mathrm{G2}}^2 + \sigma^2/N} w_{\mathrm{ML}}. \tag{21}$$

By substituting $\tilde{w} = w_{\mathrm{G2}}$, the student is not operating with the true cost function (17), but with a modified, self consistent, cost function depicted in figure 1. We note that by this substitution, a potentially dangerous global minimum is created at $w_{\mathrm{G2}} = 0$.

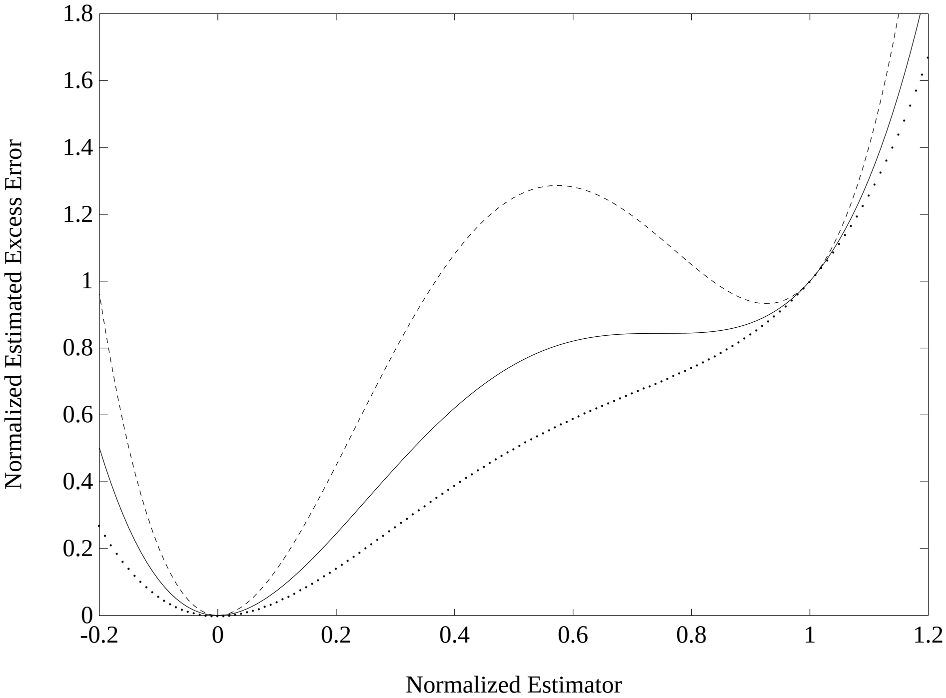

Figure 1: The estimated average excess generalization error as function of normalized estimator. The graphs are obtained by introducing (16) in (17), and setting $\tilde{w} = w_{\mathrm{G}}$. The estimator is scaled by w_{ML}, and the estimated excess error is scaled by σ^2/N. The three graphs correspond to different noise levels (dashed: low, solid: critical & dotted: high). Note that if the noise is low, a local minimum is found, otherwise pruning takes place.

We could envision eq. (21) solved by iteration, *e.g.* starting from $w_{\mathrm{G2}}^{(0)} = w_{\mathrm{ML}}$. Iterating the function on the right hand side we see that, besides the solution $w_{\mathrm{G2}} = 0$, there may be two more fixed points (or solutions) depending on the parameters. Analyzing the iteration scheme for this simple case is straightforward, one of the possible fixed points is

[3] Interestingly, this equation is recovered from the Bayesian approach if we optimize w and α simultaneously (*i.e.*, if we don't separate inference in two distinct levels). In this case Bayes' rule becomes

$$P(w, \alpha|D) = \frac{P(D|w,\alpha)P(w|\alpha)P(\alpha)}{P(D)}$$

stable, the other unstable. If it exists, the stable fixed point (corresponding to the local minimum in figure 1) is found by the iteration scheme, and it is given by

$$w_{\mathrm{G2}} = \begin{cases} \frac{w_{\mathrm{ML}}}{2}\left(1+\sqrt{1-4\sigma^2/Nw_{\mathrm{ML}}^2}\right) & w_{\mathrm{ML}}^2 > 4\sigma^2/N \\ 0 & w_{\mathrm{ML}}^2 \le 4\sigma^2/N \end{cases} . \tag{22}$$

Note that the noise limits are different for the generalization error method and Bayes' method. The generalization error method is more conservative than Bayes' (if we associate conservatism with setting $w = 0$). The noise limit is a factor 4 larger, meaning that pruning or super-regularization will happen more often. Note also that this estimator, like the Bayesian, is biased but consistent.

6 Comparing the regularization schemes

As discussed above, the quantity to compare is the average (w.r.t. training sets) generalization error of the estimators. Since each estimator is a function of the empirical mean, w_{ML}, only, the average involves averaging w.r.t. the distribution of this quantity and we already noted that $w_{\mathrm{ML}} \sim \mathcal{N}(\tilde{w}, \sigma^2/N)$. The generalization error of the maximum likelihood estimator itself was given in (5), it is independent of the teacher magnitude.

Since the regularized estimators only depend on the training set value of w_{ML}, the average excess generalization errors ($\Delta E(\tilde{w}) = E(\tilde{w}) - \sigma^2$) are functions of $\tilde{w}/\sqrt{\sigma^2/N}$

$$\Delta E(\tilde{w}) = \frac{1}{\sqrt{2\pi\sigma^2/N}} \int_{-\infty}^{\infty} dw_{\mathrm{ML}} \exp\left(-\frac{(w_{\mathrm{ML}}-\tilde{w})^2}{2\sigma^2/N}\right) \left(\tilde{w} - w_{\mathrm{E}}(w_{\mathrm{ML}})\right)^2 , \tag{23}$$

where w_{E} is either of the three biased estimators. The integral is easily evaluated numerically. In figure 2 we picture these functions, and we note that adaptation of the regularization parameter leads to a decreased error *only if the teacher is small.* If the teacher is very large the adaptive schemes lead to a negligible increase in error as they indeed should. However, if the teacher is of *intermediate* size all schemes produce seriously increased errors. The reason for this increased error is simply that the adaptive schemes are too conservative, there is a certain probability even for a unit magnitude teacher that the scheme will prune, hence introduce a large error.

We conclude that the benefit of the adaptive schemes depends critically on the distribution of teachers, *i.e.*, the extent to which the domain lives up to the assumptions of the regularizer. In fact distributions of teachers can be given for which each of the estimation schemes would be preferred. We note that for a teacher distribution that has significant mass at small teachers like the $P_0(\tilde{w}) \sim 1/|\tilde{w}|$ distribution corresponding to the prior of the Bayesian scheme (Buntine and Weigend, 1991) the *brave generalization student* is

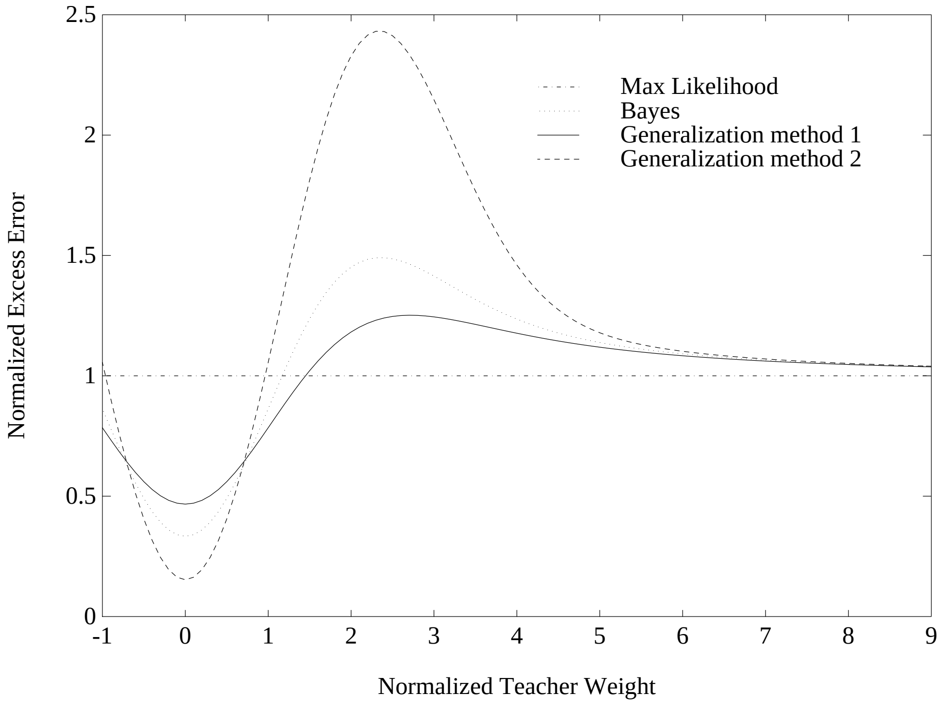

Figure 2: Average excess generalization error $\Delta E = E - \sigma^2$,as function of teacher magnitude for the two adaptive regularization schemes used to estimate the mean of a Gaussian variable. The teacher weight has been scaled by $\sqrt{\sigma^2/N}$, and the average excess error is scaled by the excess error of the maximum likelihood estimator (σ^2/N).

superior for the present problem. This fact is most easily appreciated by noting that integrating the functions in figure 2 w.r.t. $P_0(\tilde{w})$ corresponds to integrating the functions on a logarithmic ordinate axis with a uniform measure, hence stretching the region around $|\tilde{w}| \sim 0$ to infinity. We iterate: this is no big surprise since that student precisely minimizes the generalization error with the correct (implicit) prior.

Our experience with adaptive regularization is generally positive. A detailed account of multivariate applications, where pruning is achieved by utilizing individual weight regularization will be given elsewhere (Rasmussen and Hansen, 1993). The idea of using the *estimated* test error as a cost function for determination of regularization parameters is quite general. We are currently pursuing this in the context of time series processing with feed-forward networks for which the corresponding generalization error estimate is well-known see eg. (Svarer et al., 1993)

In conclusion we have shown that the use of adaptive regularization for the simple task of estimating the mean of a Gaussian variable of known variance, leads to non-trivial decision processes. We have defined a competitive scheme for identification of the regularization parameter, based on minimalization of the expected generalization error. We have shown that pruning can be viewed as infinite regularization, and that this is a natural consequence of adaptive regularization.

7 Acknowledgements

We thank the participants in the 1992 CONNECT Journal Club for discussions on Bayesian methods. We thank Anders Krogh, Jan Larsen, David MacKay, Peter Salamon, Claus Svarer, and Hans Henrik Thodberg for valuable discussions on this work. Also, we are grateful to the anonymous reviewer for constructive criticism. This research is supported by the Danish Research Councils for the Natural and Technical Sciences through the Danish Computational Neural Network Center (CONNECT).

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst. Stat. Mat.*, 21:243–247.

Buntine, W. and Weigend, A. (1991). Bayesian back-propagation. *Complex Systems*, 5:603–643.

Hansen, L. (1993). Stochastic linear learning: Exact test and training error averages. *Neural Networks*, 6:393–396.

Le Cun, Y., Denker, J., and Solla, S. (1990). Optimal brain damage. In Touretzky, D., editor, *Advances in Neural Information Processing Systems*, volume 2, pages 598–605, San Mateo. (Denver 1989), Morgan Kaufmann.

MacKay, D. (1992a). Bayesian interpolation. *Neural Computation*, 4:415–447.

MacKay, D. (1992b). A practical framework for backpropagation networks. *Neural Computation*, 4:448–472.

Moody, J. (1991). Note on generalization, regularization and architecture selection in nonlinear systems. In Kung, S., Juang, B., and Kamm, C., editors, *IEEE International Conference on Neural Networks*, pages 1–10. IEEE.

Rasmussen, C. and Hansen, L. (1993). *In preparation.*

Solla, S. (1992). Capacity control in classsifiers for pattern recognizers. In Kung, S.Y. et al., editor, *IEEE International Conference on Neural Networks*, pages 255–266, Piscataway, NJ. IEEE.

Svarer, C., Hansen, L., and Larsen, J. (1993). On design and evaluation of tapped-delay neural network architectures. In Berenji, H.R. et al., editor, *IEEE International Conference on Neural Networks*, volume 1, pages 46–51, Piscataway, NJ. (San Francisco 1993), IEEE.

Thodberg, H. (1991). Improving generalization on neural networks through pruning. *International Journal of Neural Systems*, 1:317–326.