OVERTRAINING, REGULARIZATION, AND SEARCHING FOR MINIMUM WITH APPLICATION TO NEURAL NETWORKS

J. Sjöberg, L. Ljung
Department of Electrical Engineering
Linköping University
S-581 83 Linköping, Sweden
Phone: +46 13 281890
E-mail: `sjoberg@isy.liu.se`

May 25, 1994

Abstract

In this paper we discuss the role of criterion minimization as a means for parameter estimation. Most traditional methods, such as maximum likelihood and prediction error identification are based on these principles. However, somewhat surprisingly, it turns out that it is not always "optimal" to try to find the absolute minimum point of the criterion. The reason is that "stopped minimization" (where the iterations have been terminated before the absolute minimum has been reached) has more or less identical properties as using regularization (adding a parametric penalty term). Regularization is known to have beneficial effects on the variance of the parameter estimates and it reduces the "variance contribution" of the misfit. This also explains the concept of "overtraining" in neural nets.

How does one know when to terminate the iterations then? A useful criterion would be to stop iterations when the criterion function applied to a validation data set no longer decreases. However, we show in this paper, that applying this technique extensively may lead to the fact that the resulting estimate is an unregularized estimate for the total data set: Estimation + validation data.

1 Introduction

A general nonlinear regression model can be expressed as

$$y(t) = g(\varphi(t), \theta) \tag{1.1}$$

When used in system identification $y(t)$ would be the output of the system and $\varphi(t)$ would contain past inputs and outputs. θ is the parameter vector which has to be fitted to data so that the model resembles the the input-output behavior of the system as well as possible. If $g(\cdot, \cdot)$ is a nonlinear black-box model it typically must contain quite a few parameters to possess the flexibility to approximate almost “any” function. The results in this paper apply to all ill-conditioned, parameterized models and we will especially address adaptive Neural Networks (NN) which typically belongs to this group, (Saarinen et al., 1991).

A characteristic feature of NN models is that the dimension of θ is quite high, often several hundreds. From an estimation point of view this should raise some worries, since it is bound to give a large “variance error” (*i.e.*, that modeling errors that originate from the noise disturbances in the estimation (“training”) data set cause misfits when the model is applied to a validation (“generalization”) data set). Nevertheless NN’s have shown good abilities for modeling dynamical systems.

In this paper we explain why this is so. The key is the concept of *regularization*. In Section 2 we show how regularization reduces the variance error and in Section 3 we show how typically used iterative estimation (“training”) procedures, such as backpropagation, implicitly employs regularization if the search is terminated before the absolute minimum is reached. We also explain the phenomenon of “overtraining” in NN in this way. Section 4 contains an example that illustrates this analysis.

It thus becomes a key issue is to determine when to terminate the search, *i.e.*, how many iterations of the numerical algorithm should be applied. In the general case the number of iterations could be chosen by cross-validation, *i.e.*, a second data set, not used to estimate the parameters in the model, is used to determine when the search should be terminated and in Section 5 we show that this can lead to some paradoxes, if too much trust is given to the cross-validation.

2 Regularization and variance reduction

Regularization is a well known concept in statistical parameter estimation. See, *e.g.*, (Wahba, 1990; Vapnik, 1982; Draper and Nostrand, 1979). Recently the concept has also been brought up in connection with neural networks; *e.g.*, (Poggio and Girosi, 1990; Moody, 1992; MacKay, 1991; Sjöberg and Ljung, 1992; Ljung and Sjöberg, 1992). We shall in this section describe, in a tutorial fashion, the basic aspects and benefits of regularization.

Let us first stress that the calculations to come are carried out under typical and “conventional” assumptions in parameter estimation. The set-up is the same as in, *e.g.*, (Ljung, 1987). This means that the analysis is to its character is *asymptotic* in the number of observations and *local* around “the true value”

of the parameters. “Global” issues, like the existence of local minima of the criterion function will thus not be addressed. It should however also be said, that the only property that we require below from the “true value” θ_0 is that it is a local minimum of the expected criterion function, such that its associated prediction errors are uncorrelated with the first and second derivatives of the function g with respect to θ (evaluated at θ_0). We do not require that such a parameter value is unique.

Consider the following general nonlinear regression problem: We observe $y(t)$ and $\varphi(t)$ for $t = 1, \dots, N$ and introduce the notation

$$\begin{aligned} Z^N &= [y^N, u^N], \\ y^N &= [y(1), \dots, y(N)]^T \\ u^N &= [u(1), \dots, u(N)]^T \end{aligned} \tag{2.1}$$

We want to estimate the parameter θ in a relationship

$$y(t) = g(\theta, \varphi(t)) \tag{2.2}$$

Assume that the observed data actually can be described by

$$y(t) = g(\theta_0, \varphi(t)) + e(t) \tag{2.3}$$

for a realization of a white noise sequence $\{e(t)\}$. Let

$$\lambda_0 = Ee^2(t) \tag{2.4}$$

We estimate θ by a common prediction error method:

$$\hat{\theta}_N = \arg \min_{\theta} V_N(\theta, Z^N) \tag{2.5}$$

where

$$V_N(\theta, Z^N) = \frac{1}{N} \sum_{t=1}^N (y(t) - g(\theta, \varphi(t)))^2 \tag{2.6}$$

which is the *maximum likelihood* (ML) criterion if $e(t)$ Gaussian noise. We do not assume that the value $\hat{\theta}_N$ is unique. The only thing that matters is that, as N grows larger, $\hat{\theta}_N$ becomes close to a value θ_0 with the properties mentioned in the beginning of this section.

We thus have the following model for the relationship between $y(t)$ and $\varphi(t)$:

$$\hat{y}(t) = g(\hat{\theta}_N, \varphi(t)) \tag{2.7}$$

As a quality measure for the model we could choose

$$\bar{V}_N = E\bar{V}(\hat{\theta}_N) \tag{2.8}$$

where

$$\bar{V}(\theta) = E(y(t) - g(\theta, \varphi(t))^2 \quad (2.9)$$

Note that in (2.9) expectation is over $\varphi(t)$ and $e(t)$, while in (2.8) expectation is over the random variable $\hat{\theta}_N$ (which depends on the random variables $\varphi(t)$, $e(t)$ $t \leq N$)

It is a well-known, general result that

$$\bar{V}_N \approx \lambda_0(1 + \frac{d}{N}) \quad (2.10)$$

where

$$d = \dim \theta \quad (2.11)$$

See, *e.g.*, (Ljung, 1987), p 418.

This says that when the parametrization is flexible enough to contain a true description of the system (2.3), each estimated parameter gives a contribution to the error (2.10) that does not depend on how important the parameter actually is for the fit. In other words, regardless of how sensitively $V_N(\theta)$ depends on a certain parameter θ_i , its contribution to the variance model error (2.10) still is λ_0/N . We can thus have parameters that do no good in improving the fit in (2.6) but just makes the model (2.7) worse when applied to a new data set. This is known as *overfit*.

There could also be parameters that do contribute marginally to the fit, but whose contribution to the error (2.10) is larger. That is to say that if a parameter improves the fit in $\bar{V}_N$, defined in (2.8), by less than λ_0/N , it is better to leave it out from the parametrization, because its overall effect on the model will be negative. Let us call such parameters *superfluous*.

In practice, we may have a situation where we suspect that there are too many parameters in a given parametrization, but we cannot *a priori* point to those that are superfluous. This situation apparently is at hand for neural network models. The traditional way to deal with this problem in statistics is to use *regularization*.

In our context this means that we seek to minimize

$$W_N(\theta) = V_N(\theta) + \delta \|\theta - \theta_0\|^2 \quad (2.12)$$

instead of (2.6). (Of course, in practice we do not know θ_0 , so that the “attraction term” in (2.12) will be around some nominal guess $\theta^\#$. We shall discuss that modification later). This criterion corresponds to a *maximum a posteriori* (MAP) estimate of the parameters with the prior assumption that they belong to a normal distribution centered at $\theta^\#$ and with variance $1/\delta I$. Let $\hat{\theta}_N^\delta$ be the value of θ that minimizes $W_N(\theta)$ so that

$$W'_N(\hat{\theta}_N^\delta) = 0 \quad (2.13)$$

Now, $\hat{\theta}_N^\delta$ will be arbitrarily close to θ_0 for N sufficiently large. (Recall the properties of θ_0 that we listed above. Then, by Taylor's expansion around θ_0 , we obtain

$$0 = W'_N(\hat{\theta}_N^\delta) \approx W'_N(\theta_0) + (\hat{\theta}_N^\delta - \theta_0)W''_N(\theta_0) \quad (2.14)$$

so that

$$(\hat{\theta}_N^\delta - \theta_0) \approx -(W''_N(\theta_0))^{-1}W'_N(\theta_0) \quad (2.15)$$

Note that, for large N , due to the law of large numbers

$$W''_N(\theta_0) = V''_N(\theta_0) + 2\delta I \approx 2(Q + \delta I)$$

where

$$Q = E\psi(t)\psi^T(t) \quad (2.16)$$

$$\psi(t) = \frac{d}{d\theta}g(\theta, \varphi(t)) \big|_{\theta=\theta_0} \quad (2.17)$$

Moreover

$$W'_N(\theta_0) = \frac{2}{N} \sum_{t=1}^N \psi(t)e(t) \quad (2.18)$$

so

$$E \, N W'_N(\theta_0)(W'_N(\theta_0))^T = 4\lambda_0 Q \quad (2.19)$$

which shows that

$$P \doteq E(\hat{\theta}_N^\delta - \theta_0)(\hat{\theta}_N^\delta - \theta_0)^T \approx \frac{\lambda_0}{N}(Q + \delta I)^{-1}Q(Q + \delta I)^{-1} \quad (2.20)$$

(These calculations follow exactly the corresponding traditional calculations, *e.g.*, in (Ljung, 1987), Chapter 9. All assumptions and formal verifications are entirely analogous).

Now, what is the model quality measure (2.8) in this case? Let us introduce

$$\bar{V}_N^\delta = E\bar{V}(\hat{\theta}_N^\delta)$$

We have, by Taylor's expansion,

$$\begin{aligned} E\bar{V}(\hat{\theta}_N^\delta) &\approx E \left\{ \bar{V}(\theta_0) + (\hat{\theta}_N^\delta - \theta_0)\bar{V}'(\theta_0) + \frac{1}{2}(\hat{\theta}_N^\delta - \theta_0)^T \bar{V}''(\theta_0)(\hat{\theta}_N^\delta - \theta_0) \right\} = \\ &= \lambda_0 + 0 + \frac{1}{2}E \operatorname{tr} \left\{ 2Q(\hat{\theta}_N^\delta - \theta_0)(\hat{\theta}_N^\delta - \theta_0)^T \right\} = \lambda_0 + \operatorname{tr} Q P \end{aligned} \quad (2.21)$$

Here we used that $\bar{V}'(\theta_0) = 0$ and $\bar{V}''(\theta_0) = 2Q$. Now plugging in (2.20) gives

$$\bar{V}_N^\delta = \lambda_0 \left(1 + \frac{1}{N} \operatorname{tr} Q(Q + \delta I)^{-1}Q(Q + \delta I)^{-1} \right)$$

Since all the matrices written within *trace* can be diagonalized simultaneously we find that

$$\bar{V}_N^\delta = \lambda_0 \left(1 + \frac{\tilde{d}}{N}\right) \quad (2.22)$$

$$\tilde{d} = \sum_{i=1}^d \frac{\sigma_i^2}{(\sigma_i + \delta)^2} \quad \sigma_i : \text{ eigenvalues of } Q \quad (2.23)$$

First, we notice that for $\delta = 0$ we reobtain (2.10). Second, suppose that the spread of the eigenvalues σ_i is quite substantial (which usually is the case for neural nets, see (Saarinen et al., 1991)), so that σ_i is often either significantly larger or significantly smaller than δ . In that case, we can interpret $\tilde{d}$ as follows:

$$\tilde{d} \approx \begin{array}{l} \text{number of eigenvalues of } Q \\ \text{that are larger than } \delta \end{array} \quad (2.24)$$

Here we see the important benefit of *regularization*: *The superfluous parameters no longer have a bad influence on the model performance.*

We must however also address the problem that in practice we cannot use θ_0 in (2.12) but must use

$$W_N(\theta) = V_N(\theta) + \delta \|\theta - \theta^\#\|^2$$

for some nominal guess $\theta^\#$. Then the estimate $\hat{\theta}_N^\delta$ will not converge to the true value θ_0 but to some other value θ^* as $N \rightarrow \infty$. This gives an extra contribution to (2.22) which is

$$\bar{V}(\theta^*) - \bar{V}(\theta_0) \approx (\theta^* - \theta_0)^T Q (\theta^* - \theta_0) \quad (2.25)$$

Now, in general θ^* and θ_0 need not be close. We stress that the analysis to follow is based on (2.25), which will hold if a second order approximation of the criterion function is reasonable in the region “between” θ^* and θ_0 . This may be true even if θ^* and θ_0 are not close.

We know that, by definition of θ^* and θ_0

$$\bar{V}'(\theta^*) + 2\delta(\theta^* - \theta^\#) = 0 \quad (2.26)$$

$$\bar{V}'(\theta^*) = \bar{V}'(\theta_0) + 2Q(\theta^* - \theta_0) = 0 + 2Q(\theta^* - \theta_0) \quad (2.27)$$

This gives

$$Q(\theta^* - \theta_0) = -\delta(\theta^* - \theta^\#) = -\delta(\theta^* - \theta_0 + \theta_0 - \theta^\#)$$

or

$$(Q + \delta I)(\theta^* - \theta_0) = -\delta(\theta_0 - \theta^\#) \quad (2.28)$$

which, inserted into (2.25) gives

$$\bar{V}(\theta^*) - \bar{V}(\theta_0) = \delta^2(\theta_0 - \theta^\#)^T (Q + \delta I)^{-1} Q (Q + \delta I)^{-1} (\theta_0 - \theta^\#)$$

Now,

$$\| (Q + \delta I)^{-1} Q (Q + \delta I)^{-1} \| \leq \max_i \frac{\sigma_i}{(\sigma_i + \delta)^2} \leq \frac{1}{4\delta}$$

so

$$| \bar{V}(\theta^*) - \bar{V}(\theta_0) | \leq \frac{\delta}{4} | \theta_0 - \theta^\# |^2 \quad (2.29)$$

What all this means is the following: Suppose that we do have some knowledge about the order of magnitude of θ_0 (so that $| \theta_0 - \theta^\# |$ is not exceedingly large). Suppose also that there are $d^\#(\delta)$ parameters that contribute with eigenvalues of Q that are less than δ . Then introducing regularization - with parameter δ - decreases the model error (2.22) by $\lambda_0 d^\# / N$, at the same time as the increase (2.25) due to bias is less than $\delta/4 \cdot | \theta_0 - \theta^\# |^2$. Regularization thus leads to a better mean square error of the model if

$$\frac{\lambda_0 d^\#(\delta)}{N} - \frac{\delta}{4} | \theta_0 - \theta^\# |^2 > 0 \quad (2.30)$$

This expression gives a precise indication of when regularization is good and what levels of δ are reasonable. Note that

$$Q = E\psi(t)\psi^T(t)$$

is a matrix of which we have a good estimate from data (assuming that the current estimate of θ is close to θ_0). This means that $d^\#(\delta)$ is known to us with good approximation. Likewise, we will have a good estimate of λ_0 . N is a fixed, known number, and really the only unknown quantity in (2.30) is θ_0 . With some measure of the size of $| \theta_0 - \theta^\# |$ we should thus choose the regularization parameter δ such that

$$\delta^* = \arg \max_\delta \left\{ \frac{\lambda_0 d^\#(\delta)}{N} - \frac{\delta}{4} | \theta_0 - \theta^\# |^2 \right\} \quad (2.31)$$

Let us conclude this section by deriving how much the estimate $\hat{\theta}_N^\delta$ differs from $\hat{\theta}_N$. We will find use of it in the next section.

We have

$$0 = W'_N(\hat{\theta}_N^\delta) = V'_N(\hat{\theta}_N^\delta) + 2\delta(\hat{\theta}_N^\delta - \theta^\#)$$

Also

$$V'_N(\hat{\theta}_N^\delta) = V'_N(\hat{\theta}_N) + V''_N(\hat{\theta}_N)(\hat{\theta}_N^\delta - \hat{\theta}_N) \approx 2Q(\hat{\theta}_N^\delta - \hat{\theta}_N)$$

We thus have

$$Q(\hat{\theta}_N^\delta - \hat{\theta}_N) = -\delta(\hat{\theta}_N^\delta - \hat{\theta}_N + \hat{\theta}_N - \theta^\#)$$

or

$$(\hat{\theta}_N^\delta - \hat{\theta}_N) = \delta(Q + \delta I)^{-1}(\theta^\# - \hat{\theta}_N) \quad (2.32)$$

This can also be written as

$$\hat{\theta}_N^\delta = (I - M_\delta)\hat{\theta}_N + M_\delta\theta^\# \quad (2.33)$$

$$M_\delta = \delta(\delta I + Q)^{-1} \quad (2.34)$$

so the regularized estimate is a weighted mean between the unregularized one and the nominal value $\theta^\#$.

3 Terminating iterative search for the minimum is regularization

The unregularized estimate $\hat{\theta}_N$, defined by (2.5) is usually computed by iterative search of the following kind:

$$\begin{aligned} \hat{\theta}_N^{(i+1)} &= \hat{\theta}_N^{(i)} - \mu_N^{(i)} H^{(i)} V'_N(\hat{\theta}_N^{(i)}) \\ \hat{\theta}_N^{(0)} &= \theta^\# \end{aligned} \quad (3.1)$$

Here superscript (i) indicates the i :th iterate. The step size $\mu_N^{(i)}$ is determined by some search along the indicated line. $H^{(i)}$ is a positive definite matrix that may modify the search direction from the negative gradient one, V'_N , to some other one. The initial value $\theta^\#$ is natural to use, viewing $\theta^\#$ as some prior guess of θ_0 .

Now, note that for any reasonable choices of $\mu_N^{(i)}$ and $H^{(i)}$ (neglecting local minima) we have

$$\lim_{i \rightarrow \infty} \hat{\theta}_N^{(i)} = \hat{\theta}_N \quad (3.2)$$

The question we are going to address here is what estimate we have after a finite number of iterations.

Now, by Taylor's expansion

$$V'_N(\hat{\theta}_N^{(i)}) \approx Q(\hat{\theta}_N^{(i)} - \hat{\theta}_N) \quad (3.3)$$

where Q is defined by (2.16). Introduce

$$\tilde{\theta}_N^{(i)} = \hat{\theta}_N^{(i)} - \hat{\theta}_N \quad (3.4)$$

Then (3.1) gives

$$\tilde{\theta}_N^{(i)} = \prod_{j=1}^i (I - \mu_N^{(j)} H^{(j)} Q) (\theta^\# - \hat{\theta}_N) \quad (3.5)$$

This can be written as

$$\hat{\theta}_N^{(i)} = (I - M_i) \hat{\theta}_N + M_i \theta^\# \quad (3.6)$$

$$M_i = \prod_{j=1}^i (I - \mu_N^{(j)} H^{(j)} Q) \quad (3.7)$$

If $H^{(i)}$ and $\mu_N^{(i)}$ are held constant during the estimation (3.7) becomes

$$M_i = (I - \mu_N H Q)^i \quad (3.8)$$

Compare (3.6), (3.8) with (2.33), (2.34)! The point now is that - except in the true Newton case $H = Q^{-1}$ - the matrices M_i and M_δ behave similarly. For simplicity, take $H = I$ (making (3.1) a gradient search scheme), and assume that Q is diagonal with decreasing values along the diagonal (this can always be achieved by a change of variables). Then the j :th diagonal element of M_δ will be

$$M_\delta^{(jj)} = \frac{\delta}{\sigma_j + \delta} \quad (3.9)$$

while the corresponding element of M_i is

$$M_i^{(jj)} = (1 - \mu\sigma_j)^i \quad (3.10)$$

By necessity, we must choose $\mu < 2/\sigma_1$ to keep the scheme (3.1) convergent. In Figure 1 we plot these two expressions as functions of $1/\sigma_j$. We see that the effect is pretty much the same. The amount of regularization that is inflicted depends as the size of δ (large δ corresponds to small i).

We may formalize this relationship between i and δ further. Asymptotically, for small σ_i we have that

$$M_\delta^{(jj)} = M_i^{(jj)}$$

implies

$$i \log(1 - \mu\sigma_j) = -\log(1 + \sigma_j/\delta)$$

or

$$i \approx \frac{1}{\mu\delta} \quad (3.11)$$

Hence the number of iterations is directly linked to the regularization parameter. The number, $\tilde{d}$, of efficient parameters in the parametrization depends on the regularization parameter δ :

$$\tilde{d}(\delta) \quad (3.12)$$

as described by (2.23). Via (3.11) we thus see that the efficient number of parameters used in the parametrization actually increases with the number of iterations ("training cycles") :

$$\tilde{d}(i) \quad (3.13)$$

In summary we have established here that "unfinished" search for the minimization of a criterion function has the same effect as regularization towards the initial value $\theta^\#$. Only schemes that allow exact convergence in a finite number of steps do not show this feature. If the regularization is implemented by terminated search we will call it *implicit regularization* to distinguish from the *explicit regularization* when W_N is being minimized. A related discussion with similar ideas can also be found in (Wahba, 1987).

Figure 1: Solid line: $M_i^{(jj)}$; dashed line: $M_\delta^{(jj)}$. $i = 8$, $\mu = 0.1$, and $\delta = 1$. X-axis: $1/\sigma_j$

4 An Illustrative Example

Let us now illustrate what has been said by modeling an hydraulic robot arm with a NN model.

The goal is to model the dynamics of a hydraulically controlled robot arm. The position of the robot arm is controlled by the oil pressure in the cylinder which in turn is controlled by the size of the valve through which the oil flows. The valve size, $u(t)$, and the oil pressure, $y(t)$ are input and output signals, respectively. They are shown in Figure 3. As seen in the oil pressure, we have a very oscillative settling period after a step change of the valve size. These oscillations are caused by mechanical resonances in the robot arm.

If the oil pressure had been linearly dependent on the valve position then, of course, there would be no reason to use a NN to model this behavior, but a linear black-box model would have been sufficient. The oscillations of the oil pressure, however, decay in a nonlinear way, and therefore we tried a network model.

The model is a feed forward neural network with one hidden layer, shown in Figure 2. (see *e.g.*, (Hecht-Nielsen, 1990), or any book about NN for an explanation of the NN terminology). The model can be mathematically described as

$$g(\theta, \varphi(t)) = \sum_{j=1}^H c_j \sigma(A_j \varphi(t) + a_j) + c_0 \quad (4.1)$$

where θ contains all parameters, c_j , A_j and a_j . H is the number of *hidden units* and A_j are vectors of parameters with the same dimension as $\varphi(t)$. $\sigma(\cdot)$ is given

Figure 2: A feedforward network with one hidden layer and one output unit.

Figure 3: Measured values of valve position (top) and oil pressure (bottom).

by

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (4.2)$$

which is called the *sigmoid* function.

We had 1024 samples available which were divided into two equally sized sets; the estimation and validation sets.

The input vector to the network, $\varphi(t)$ consists of values of $y(s)$ and $u(s)$ with $s < t$, and the output is $\hat{y}(t) = g(\varphi(t), \theta)$, the predicted value of $y(t)$. The dimension of $\varphi(t)$ determines the number of inputs in the network. After having tried several alternatives, we chose the following regression vector

$$\varphi(t) = [u(t-1) \ u(t-2) \ u(t-3) \ y(t-1) \ y(t-2)] \quad (4.3)$$

To finally settle the structure of the NN model the number of hidden units had to be chosen. This number will decide the networks approximation abilities

and we obtained sufficient flexibility for a good fit with ten hidden units. The model did not improve if the number of hidden units was increased further.

With ten hidden units, five inputs, and one output unit we have $10 * (5 + 1) + (10 + 1) = 71$ parameters.

We then estimated the parameters in this network with our 524 samples in the estimation set. In Figure 4 the criterion computed on the validation set after each iteration of the minimization algorithm is shown for two different sizes of regularization. In the first case (solid line) the regularization was set to zero, and we can clearly see that overtraining occurs. First the criterion decreases and the fit improves. After approximately 10 iterations the minimum is reached which corresponds to the model that minimizes the fit for the validation data. Further iterations give an increasing criterion and the model becomes worse again. The theory from the preceding sections explains this behavior. At the beginning of the search the most important parameters are adapted and the fit improves. After some ten iterations the important parameters are fitted and just the superfluous parameters remain unfitted. When they also converge they give rise to overfit which is seen by the increase of the criterion. This can also be seen as a regularization which is slowly switched off during the search. First, before the search starts, there is nothing but bias towards the initial parameter value, which corresponds to a very large regularization (recall from (3.11) that the regularization parameter is linked to the number of iterations!). Then gradually during the search as the parameters are fitted to the data, the bias diminish. The minimum of the criterion corresponds to an optimal choice of regularization.

In the second case (dashed line in Figure 4) the regularization was chosen to $10^{-2} | \theta - \theta^\# |$, with $\theta^\# = 0$, and we see that no overtraining occurs, there is no increase of the criterion. In this case the regularization is not switched off and, hence, the overfit is prevented.

Regularization towards other values than $\theta^\# = 0$, but still within a reasonable neighborhood, gives similar results. Hence, the most important feature is to prevent the superfluous parameters to adapt to noise, and not to attract the parameters towards particular values.

How are the singular values of the Hessian affected by the regularization? In Figure 5 the singular values for these two cases are shown. Without regularization the singular values can get arbitrarily small, and in this example the smallest is of the order 10^{-7} . This means that the corresponding parameters, or parameter combinations, are very fragile to disturbances.

The singular values cannot be smaller than the regularization currently used, and we see this when a regularization of 10^{-2} is being used. The former small singular values are set equal the regularization, which then means that they cannot be disturbed so easily any longer. These singular values corresponds to superfluous parameters, virtually switched off by introduction of the regularization.

In respect to these arguments we can regard the singular values larger than

Figure 4: The criterion of fit evaluated for validation data after each iteration of the numerical search. Solid line: no regularization. The criterion reaches a minimum and increases again, *i.e.*, overtraining. Dashed line: regularized with 10^{-2} . No overtraining occurs

the regularization as corresponding to the useful parameters. From Figure 5 we conclude that one can make use of roughly 30 parameters when identifying this plant with the given data, and with this particular model.

An interesting feature is that if the number of hidden units was increased and the regularization was kept at 10^{-2} , then all additional parameters became superfluous, *i.e.*, the virtual number of parameters stayed at 30. This means that the data does not contain any more information which can be incorporated by an increase of the number of hidden units.

In Figure 6 the NN model is used for simulation on the validation set and compared to the true output of the plant and the simulation of the best linear model.

An alternative to regularization would be to reduce the number of hidden units and this way get rid of some of the parameters so that no overtraining occurs, *i.e.*, the number of parameters is reduced until no superfluous parameters remain. This is not trivial because when we speak of superfluous parameters we actually mean superfluous directions in the parameter space. These directions are given by the eigenvectors of the Hessian, Q , (2.16). Hence, it is not certain that the small eigenvalues correspond to distinct parameters. However, we tried this in the framework of our NN model, *i.e.*, we applied the same NN model and reduced the number of hidden units until no regularization was necessary to prevent overtraining. We landed on five hidden units, *i.e.*, $5 \cdot (5+1) + (5+1) = 36$ parameters, however the remaining model did not possess enough flexibility of adaptation, and the overall result was worse then with the larger model in connection with regularization. The conclusion is that by removing superfluous

Figure 5: Logarithm of the singular values of the Hessian of the criterion of fit; [x] - no regularization; [o] - regularized with 10^{-2} .

Figure 6: True output (solid line), simulated by NN model (dashed line), and simulated by the best linear model (dotted).

parameters we had also removed some important ones too.

5 Criterion Minimization Using Estimation Data and Validation Data

The obvious thought when faced with an estimation problem like (2.5) is of course to carry out the minimization until the absolute minimum has been reached. As we have found in the previous sections this might however lead to a worse estimate than when the iterations are “prematurely” stopped. This is due to the implicit, and beneficial, regularization effect.

We are thus left with the problem to decide when to stop the iteration - or, equivalently according to (3.11), to select the size of the regularization parameter δ .

A reasonable and general approach is to apply cross-validation, *i.e.*, to use a second data set - the validation data, that have not been utilized for the parameter estimation - to decide when to stop the iterations. This is natural, since the criterion that we really want to minimize is the expected value of the criterion function, $\bar{V}_N$, in (2.8). Pragmatically, this also makes sense, since we are looking for a model that is good at reproducing data it has not been adjusted to.

In practical use, this idea is easily applied as follows: Let

$$V_N^E(\theta, Z_E^N) \quad (5.4)$$

be the criterion function (like (2.6)) evaluated for the estimation data and let

$$V_M^V(\theta, Z_V^M) \quad (5.5)$$

be the corresponding function evaluated for the validation data. Run the minimization routine of your choice to minimize $V_N^E(\theta, Z_E^N)$, like (3.1) and use

$$V_M^V(\hat{\theta}_N^{(i+1)}, Z_V^M) > V_M^V(\hat{\theta}_N^{(i)}, Z_V^M) \quad (5.6)$$

as the stopping criterion. This indeed is a commonly used method, not the least in connection with neural network estimation.

To carry this idea somewhat further, it is natural to extend the iterative search routine for the iterative search routine for the minimization of V_N^E to look in all descent directions for a new estimate. That leads to the following stopping criterion:

$$\begin{aligned} &\text{Stop the minimization when no parameter} \\ &\text{value can be found that decreases the values} \\ &\text{of both } V_N^E \text{ and } V_M^V \end{aligned} \quad (5.7)$$

We are thus lead to a multicriterion minimization problem, and it is of interest to find out the parameters that obey (5.7), *i.e.*, the Pareto-optimal points. We confine this analysis to the case of quadratic criteria or, equivalently, local properties.

In the quadratic case the criteria can be written

$$V_N^E(\theta, Z_N^E) = (\theta - \bar{\theta}_N^E)R_N^E(\theta - \bar{\theta}_N^E) + \lambda_0^E \quad (5.8)$$

$$V_M^V(\theta, Z_M^V) = (\theta - \bar{\theta}_M^V)R_M^V(\theta - \bar{\theta}_M^V) + \lambda_0^V \quad (5.9)$$

where $\bar{\theta}_N^E$ and $\bar{\theta}_M^V$ are the minima of V_N^E and V_M^V , R_N^E and R_M^V are the Hessians of V_N^E and V_M^V , and λ_0^E and λ_0^V are the noise variances of the data sets Z_N^E and Z_M^V (recall (2.4)), respectively.

A point θ^* , such that (5.8) and (5.9) cannot both be decreased by picking another value, must have the property that the gradient of V_N^E and V_M^V have opposite directions at the this value, *i.e.*,

$$R_N^E(\theta^* - \bar{\theta}_N^E) = -kR_M^V(\theta^* - \bar{\theta}_M^V) \quad (5.10)$$

for some $k > 0$

The possible resulting estimates $\hat{\theta}_N$ using the stopping rule (5.7) are thus given by

$$\hat{\theta}_N = (R_N^E + kR_M^V)^{-1}(kR_M^V\bar{\theta}_M^V + R_N^E\bar{\theta}_N^E) \quad (5.11)$$

for some positive number k . Which values we converge to depends on the initial conditions.

Now, clearly, $\hat{\theta}_N$ is also the value that minimizes $V_N^E + kV_M^V$, *i.e.*,

$$\hat{\theta}_N = \arg\min_{\theta}(V_N^E(\theta, Z_N^E) + kV_M^V(\theta, Z_M^V)) \quad (5.12)$$

The implications of this analysis are somewhat paradoxical: Applying thoughtful cross-validation in order to obtain well balanced regularization properties is the same as calculating the absolute minimum - without regularization - of a criterion (5.12) which weights the estimation and validation data together. If such a combination of data is to be made, it is natural to let k reflect the relative signal-to-noise properties in the respective sets, and thus take $k = 1$ if the E and V data have been collected under similar circumstances.

Notice that this actually entirely removes the “*validation*” aspects of the data Z_M^V : we could as well have joined Z_N^E and Z_M^V to begin with into one bigger data set and minimize the corresponding (unregularized) criterion.

The validation data must thus be used with caution, and actually just be utilized to check the validity of one given model.

Note also that due to the links (3.11), (3.12), and (3.13) between number of iterations i , regularization parameter δ , and efficient number of estimated parameters $\tilde{d}$ (“efficient model order”) the same remarks apply also to the model structure selection in general.

6 Conclusions

This contribution has pointed out the following results:

1. Regularization is a necessary feature to keep the variance error reasonably small in models employing many parameters.
2. Performing a limited amount of iterations when searching for the criterion minimum is a way to achieve regularization.

One consequence of this is that performing more (gradient-type) iterations of a non-regularized criterion will make the model worse. This phenomenon is well-known in neural network modeling and is called “overtraining”, or “overlearning”. It is also interesting to link the concept of “overtraining” to the classical statistical problem of “overfit”. By overfit is then meant that an *unnecessary large number of parameters* have been employed by the parametrization, so that the model has excessive variance. In the NN context “overtraining” typically relates to the *number of training cycles*. By the discussion in section 3 - in particular (3.11) - the true connection between these concepts have been clarified.

3. A natural approach to use this result, would be to iteratively try to minimize the criterion based on estimation data, and stop the iterations when the criterion applied to validation data no longer decreases. In NN modeling this is common practice.
4. However, if all descent directions for the criterion based on estimation data are tried out using this method, it turns out that this is equivalent to minimize a criterion based on both the estimation and the validation data sets, using no regularization.
5. There are two important consequences of the last observation:
 - a Using the validation data in this way, actually annihilates the regularization effect. We obtain an unregularized estimate for the total data set.
 - b It also shows that one has to be careful to use a “validation” data set when deciding about degree of regularization. It actually turns out that the validation data set starts to play a role of estimation data.
6. Regularization or early stopped iterations have the same basic features as a choice of model order: Many parameters in a model structure have a similar effect as using a low degree regularization and as using many iterations in the minimization. The above point thus indicates that using validation data for model structure selection may show a similar feature in that the validation data no longer becomes a good independent test of the resulting models qualities.

References

- Draper, N. and Nostrand, R. V. (1979). Ridge regression and James-stein estimation: Review and comments. *Technometrics*, 21(4):451–466.
- Hecht-Nielsen, R. (1990). *Neurocomputing*. Addison-Wesley.
- Ljung, L. (1987). *System Identification: Theory for the User*. Prentice-Hall, Englewood Cliffs, NJ.
- Ljung, L. and Sjöberg, J. (1992). A system identification perspective on neural nets. In *1992 IEEE Workshop on Neural Networks for Signal Processing*, IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854-4150.
- MacKay, D. (1991). *Bayesian Methods for Adaptive Models*. PhD thesis, Caltech, Pasadena, CA 91125.
- Moody, J. (1992). The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In Moody, J., Hanson, S., and Lippmann, R., editors, *Advances in Neural Information Processing Systems 4*. Morgan Kaufmann Publishers, San Mateo, CA.
- Poggio, T. and Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982.
- Saarinen, S., Bramley, R., and Cybenko, G. (1991). Ill-conditioning in neural network training problems. Technical report, CSRDRReport 1089, Center for Supercomputing Research and Development, University of Illinois at Urbana-Champaign, 305 Talbot-104 South Wright Street, Urban, IL 61801-2932.
- Sjöberg, J. and Ljung, L. (1992). Overtraining, regularization, and searching for minimum in neural networks. In *Preprint IFAC Symposium on Adaptive Systems in Control and Signal Processing*, pages 669–674, Grenoble, France.
- Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York.
- Wahba, G. (1987). Three topics in ill-posed problems. In Engl, H. and Groetsch, C., editors, *Inverse and Ill-posed Problems*. Academic Press.
- Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, 3600 University City Science Center, Philadelphia, PA 19104-2688.