A generalization error estimate for nonlinear systems

Larsen, Jan

Published in:
Proceedings of the IEEE-SP Workshop Neural Networks for Signal Processing

Link to article, DOI:
10.1109/NNSP.1992.253710

Publication date:
1992

Document Version
Publisher's PDF, also known as Version of record

A GENERALIZATION ERROR ESTIMATE FOR NONLINEAR SYSTEMS

Jan Larsen
The Computational Neural Network Center
Electronics Institute, Building 349
Technical University of Denmark
DK-2800 Lyngby, Denmark

INTRODUCTION

Evaluation of the quality of an estimated nonlinear model, e.g. a neural network, is important for the purpose of selecting a proper architecture. In this work the employed quality measure is the generalization error (expected squared prediction error). The topic of the paper is to derive an estimate of the generalization error for *incomplete* models, i.e. models which are not capable of modeling the present nonlinear relationship perfectly.

Consider the following discrete nonlinear system:

$$y(k) = g(\mathbf{x}(k)) + \varepsilon(k) \tag{1}$$

where the scalar output, $y(k)$, (k is the discrete time index) is generated as the sum of a nonlinear mapping, $g(\cdot)$, of the input vector $\mathbf{x}(k)$ and the inherent noise $\varepsilon(k)$. In a signal processing context the input vector may e.g. represent a tapped delay line, i.e. $\mathbf{x}(k) = [x(k), x(k-1), \cdots, x(k-L+1)]^{\mathsf{T}}$ ($^{\mathsf{T}}$ is the transpose operator).

Assumption 1 *The input $\mathbf{x}(k)$ is assumed to be a strictly stationary sequence and $\varepsilon(k)$ a white, strictly stationary sequence with zero mean and variance σ_ε^2. Furthermore, $\mathbf{x}(k)$ is assumed independent of $\varepsilon(k)$, $\forall\, k$.*

Let $\mathcal{F}$ be a set of nonlinear functionals parameterized by an m-dimensional vector $\mathbf{w} = [w_1, w_2, \cdots, w_m]^{\mathsf{T}}$. In general it is assumed that the functionals are nonlinear in $\mathbf{w}$. Feed-forward neural networks with hidden units are examples of $\mathcal{F}$. Let $f(\cdot) \in \mathcal{F}$. The model of Eq. (1) becomes:

$$y(k) = f(\mathbf{x}(k); \mathbf{w}) + e(k; \mathbf{w}) \tag{2}$$

The prediction of $y(k)$, say $\widehat{y}(k)$, is: $\widehat{y}(k) = f(\mathbf{x}(k); \mathbf{w})$. When referring to a nonlinear model $\widehat{y}(k)$ is considered to be nonlinear in $\mathbf{w}$.

0-7803-0557-4/92$03.00 © 1992

Definition 1 *If* $\exists \mathbf{w}^o$*:* $g(\mathbf{x}(k)) \equiv f(\mathbf{x}(k); \mathbf{w}^o)$ *the model is sigified as complete otherwise as incomplete.* $\mathbf{w}^o$ *is denoted the true weights.*

Usually the lack of knowledge concerning the structure of $g(\cdot)$ precludes the possibility of suggesting a complete model with a finite m. Consequently, it is claimed that incomplete models are the common case.

Given a training set: $\mathcal{T} = \{\mathbf{x}(k), y(k)\}$, $k = 1, 2, \cdots, N$, where N is the training set size, the model is estimated by minimizing some cost function, say $S_N(\mathbf{w})$. In this work a least squares (LS) cost is employed:

$$S_N(\mathbf{w}) = \frac{1}{N}\sum_{k=1}^{N} e^2(k; \mathbf{w}) = \frac{1}{N}\sum_{k=1}^{N}(y(k) - f(\mathbf{x}(k); \mathbf{w}))^2 \tag{3}$$

The training performance $S_N(\widehat{\mathbf{w}})$ is usually not a reliable measure of the quality of a model because it depends on the actual training set. A reliable quality measure is the *generalization error*, G, (e.g. [7]) which is defined as the expected, squared prediction error on a test sample, $\{\mathbf{x}_t, y_t\}$ (denoting t for test), which is *independent* of the training set but with identical distribution.

$$G(\mathbf{w}) = E_{\mathbf{x}_t, \varepsilon_t}\left\{[y_t - f(\mathbf{x}_t; \mathbf{w})]^2\right\} \tag{4}$$

$E_{\mathbf{x}_t, \varepsilon_t}\{\cdot\}$ denotes expectation with respect to the joint p.d.f. of $[\mathbf{x}_t, \varepsilon_t]$. Note the dependence on both $f(\cdot)$ and $\mathbf{w}$.

In the litterature several attempts have been made in order to estimate the generalization error of both linear and nonlinear models, for instance [1] and [3] which focus on complete models, while [5] and [7] focus on incomplete models which are claimed to be the most common.

In [5] a generalization error estimator for linear incomplete models is developed. The estimate requires knowledge of the estimated parameters $\widehat{w}_i$, $i = m+1, m+2, \cdots, m^o$ where m^o denotes the dimension for which the model becomes complete. Unfortunately, these estimated parameters are not accessible when fitting with only m parameters. Therefore, the final result of [5] is essentially the *FPE*-criterion [1].

The *GPE* estimator [7] is claimed to estimate the generalization error for both nonlinear and incomplete (in [7] denoted biased) models when using the sum of S_N (LS-term) and a regularizing term as the cost function. However, in the next section, which deals with a new generalization error estimate with validity for both incomplete and nonlinear models, it is established that the error, $e(k; \mathbf{w})$, and the input, $\mathbf{x}(k)$, *are not* independent unless the model is complete. This dependence is not taken into account in [7].

GENERALIZATION ERROR ESTIMATE FOR INCOMPLETE, NONLINEAR MODELS

In this section a new generalization error estimate for incomplete nonlinear models, called *GEN*, is introduced. The aim is to estimate $G(\widehat{\mathbf{w}})$, i.e. how well

the estimated model, $f(\mathbf{x}(k);\widehat{\mathbf{w}})$, generalizes. In order to evaluate Eq. (4) the nonlinear system Eq. (1) must be known. Secondly, knowledge of the input and error distributions is required. However, these assumptions are not met in general; the only knowledge of the actual system is obtained implicitly from the acquired data. For that reason the presented generalization error estimate is based on training data solely.

To ensure the validity of the *GEN*-estimate the following assumptions must be satisfied:

Assumption 2 *Define Ω^m as the compact set which the weights minimizing the cost, S_N, belong to. Assume the existence of a covering of Ω^m in compact subsets, i.e. $\Omega^m = \bigcup_i \Omega^m(i)$, such that the estimator $\widehat{\mathbf{w}}(i) \in \Omega^m(i)$ uniquely minimizes $S_N(\mathbf{w})$ within the partition $\Omega^m(i)$ and further*

$$\frac{\partial S_N(\widehat{\mathbf{w}}(i))}{\partial \mathbf{w}} = \mathbf{0}, \quad \mathbf{a}^\top \frac{\partial^2 S_N(\widehat{\mathbf{w}}(i))}{\partial \mathbf{w} \partial \mathbf{w}^\top} \mathbf{a} > 0, \ \forall \mathbf{a} \neq \mathbf{0}, \ \ \forall i \tag{5}$$

Observe that $\{\widehat{\mathbf{w}}(i)\}$ may contain both local and global minima, even though the global minima are preferred. The occurrence of multiple minima is in evidence among feed-forward neural networks, due to e.g. symmetries which cause multiple minima in the cost function, see e.g. [4].

Assumption 3 *Assume a covering $\Omega^m = \bigcup_i \Omega^m_*(i)$, such that the optimal weights $\mathbf{w}^*(i) \in \Omega^m_*(i)$ uniquely minimze $G(\mathbf{w})$ within the partition $\Omega^m_*(i)$ and further*

$$\frac{\partial G(\mathbf{w}^*(i))}{\partial \mathbf{w}} = \mathbf{0}, \quad \mathbf{a}^\top \frac{\partial^2 G(\mathbf{w}^*(i))}{\partial \mathbf{w} \partial \mathbf{w}^\top} \mathbf{a} > 0, \ \forall \mathbf{a} \neq \mathbf{0}, \ \ \forall i \tag{6}$$

Note that the optimal weight vectors reflect the "best" models within the actual set $\mathcal{F}$. That is, the models obtained by training on an infinite training set corresponding to minimal generalization error as $\lim_{N\to\infty} S_N(\mathbf{w}_m) = G(\mathbf{w})$ (provided that $e^2(k;\mathbf{w})$ is mean-ergodic).

Assumption 4 *Let the minimization of S_N on the training set result in the estimate: $\widehat{\mathbf{w}}$[1]. Assume the existence of an optimal weight vector $\mathbf{w}^*$ such that the remainder of the following second order Taylor series expansion is negligible.*

$$G(\widehat{\mathbf{w}}) \approx G(\mathbf{w}^*) + \Delta\mathbf{w}^\top \mathbf{H}(\mathbf{w}^*) \Delta\mathbf{w} \tag{7}$$

where $\Delta\mathbf{w} = \widehat{\mathbf{w}} - \mathbf{w}^*$, $\mathbf{H}(\mathbf{w}^*)$ is the nonsingular (by Eq. (6)) Hessian matrix

$$\mathbf{H}(\mathbf{w}^*) = \frac{1}{2} \frac{\partial^2 G(\mathbf{w}^*)}{\partial \mathbf{w} \partial \mathbf{w}^\top} = E_{\mathbf{x}_t, e_t} \left\{ \boldsymbol{\psi}_t(\mathbf{w}^*) \boldsymbol{\psi}_t^\top(\mathbf{w}^*) - \boldsymbol{\Psi}_t(\mathbf{w}^*) e_t(\mathbf{w}^*) \right\}, \tag{8}$$

[1]Note that the weight estimate is highly dependent on the chosen weight estimation algorithm due to local optimization, initial conditions, etc. An alternative algorithm used on the same training set may therefore result in a different weight estimate.

$\boldsymbol{\psi}_t(\mathbf{w}^*) = \partial f(\mathbf{x}_t;\mathbf{w}^*)/\partial \mathbf{w}$ *and* $\boldsymbol{\Psi}_t(\mathbf{w}^*) = \partial \boldsymbol{\psi}(\mathbf{x}_t;\mathbf{w}^*)/\partial \mathbf{w}^\mathsf{T}$. *Note that* $\partial G(\mathbf{w}^*)/\partial \mathbf{w} = \mathbf{0}$ *according to Eq. (6).*

Further assume that the remainder of expanding S_N *around* $\widehat{\mathbf{w}}(i)$ *to the second order is negligible, i.e.*

$$S_N(\mathbf{w}^*) \approx S_N(\widehat{\mathbf{w}}) + \Delta\mathbf{w}^\mathsf{T}\mathbf{H}_N(\widehat{\mathbf{w}})\Delta\mathbf{w} \tag{9}$$

where $H_N(\widehat{\mathbf{w}})$ *is the nonsingular Hessian given by*

$$\mathbf{H}_N(\widehat{\mathbf{w}}) = \frac{1}{2}\frac{\partial^2 S_N(\widehat{\mathbf{w}})}{\partial \mathbf{w}\partial \mathbf{w}^\mathsf{T}} = \frac{1}{N}\sum_{k=1}^{N}\boldsymbol{\psi}(k;\widehat{\mathbf{w}})\boldsymbol{\psi}^\mathsf{T}(k;\widehat{\mathbf{w}}) - \boldsymbol{\Psi}(k;\widehat{\mathbf{w}})e(k;\widehat{\mathbf{w}}), \tag{10}$$

$\boldsymbol{\psi}(k;\widehat{\mathbf{w}}) = \partial f(\mathbf{x}(k);\widehat{\mathbf{w}})/\partial \mathbf{w}$ *and* $\boldsymbol{\Psi}(k;\widehat{\mathbf{w}}) = \partial \boldsymbol{\psi}(\mathbf{x}(k);\widehat{\mathbf{w}})/\partial \mathbf{w}^\mathsf{T}$. *Note that* $\partial S_N(\widehat{\mathbf{w}})/\partial \mathbf{w} = \mathbf{0}$ *according to Eq. (5).*

Assumption 5 $\mathbf{x}(k)$ *is an M-dependent stationary sequence, i.e.* $\mathbf{x}(k)$, $\mathbf{x}(k+\tau)$ *are independent* $\forall \tau > M$ *(A weaker assumption aims at* $\mathbf{x}(k)$ *being a strongly mixing sequence [8, p. 62]).*

Assumption 6 *Assume large training sets, i.e.* $N \to \infty$. *In particular:* $N > M$. *Further, assume that* m *is finite.*

Definition 2 *The generalization error estimate for nonlinear systems, GEN, is defined as a consistent* $(N \to \infty)$ *estimator of* Γ *(the expectation of the generalization error w.r.t. the training set),*

$$\Gamma = E_{\mathcal{T}}\{G(\widehat{\mathbf{w}})\} \tag{11}$$

where $\widehat{\mathbf{w}}$ *is the actual weight estimate and* $\mathcal{T}$ *is the training set.*

Theorem 1 *Assume that the nonlinear system is described by Eq. (1). If assumptions 1 - 6 hold and the model in Eq. (2) is incomplete then the GEN-estimate is given by:*

$$GEN = S_N(\widehat{\mathbf{w}}) + \frac{2}{N}\cdot \mathrm{tr}\left[\left(\mathbf{R}(0) + \sum_{\tau=1}^{M}\frac{N-\tau}{N}\left(\mathbf{R}(\tau) + \mathbf{R}^\mathsf{T}(\tau)\right)\right)\mathbf{H}_N^{-1}(\widehat{\mathbf{w}})\right] \tag{12}$$

where the correlation matrices $\mathbf{R}(\tau)$, $\tau = 0, 1, \cdots, M$, *are calculated as:*

$$\mathbf{R}(\tau) = \frac{1}{N}\sum_{k=1}^{N-\tau}\boldsymbol{\psi}(k;\widehat{\mathbf{w}})e(k;\widehat{\mathbf{w}})\boldsymbol{\psi}^\mathsf{T}(k+\tau;\widehat{\mathbf{w}})e(k+\tau;\widehat{\mathbf{w}}) \tag{13}$$

Sketch of Proof The basis of the proof is the Taylor series expansions in Eq. (7), (9). Taking the expectation, $E_{\mathcal{T}}\{\cdot\}$ (i.e. w.r.t. the training set) of these equations it is possible to substitute Eq. (9) into (7) and thus express $E_{\mathcal{T}}\{G(\widehat{\mathbf{w}})\}$ in terms of training data. This is due to the relation:

$E_{\mathcal{T}}\{S_N(\mathbf{w}^*)\} = E_{\mathcal{T}}\{G(\mathbf{w}^*)\}$. When evaluating the expectations it is important to notice that the error (cf. Eq. (1) and (2))

$$e(k;\mathbf{w}) = \varepsilon(k) + g(\mathbf{x}(k)) - f(\mathbf{x}(k);\mathbf{w}) \tag{14}$$

depends both on $\mathbf{x}(k)$ and $\varepsilon(k)$ unless the model is complete and $\mathbf{w}^*$ is the global optimum since $g(\mathbf{x}) \equiv f(\mathbf{x}(k);\mathbf{w}^*)$ in this case[2]. In [6] the details of the proof are given and the estimate is further extended to treat other cost functions, for instance the LS-cost with inclusion of a weight decay term as in [7]. Note, that the derivation is valid even when dealing with noise free systems, i.e. $\sigma_\varepsilon^2 = 0$.

Theorem 2 *If the system in Eq. (1) is linear, the model Eq. (2) is linear and complete, $\mathbf{w}^*$ in Assumption 4 is the global minimum, and $\sigma_\varepsilon^2 \neq 0$ then the GEN-estimate coincides with the FPE-Criterion [1]:*

$$GEN = FPE = \frac{N+m}{N-m} S_N(\hat{\mathbf{w}}) \tag{15}$$

Proof See the sketch above and [6].

NUMERICAL EXPERIMENTS

In this section the validity of the proposed generalization error estimate is tested by comparison with the *FPE*-estimate and the leave-one-out cross-validation technique. A linear system and a simple neural network is under consideration.

Linear System

The linear system is given by:

$$y(k) = y^o(k) + \varepsilon(k) = [x(k), x^2(k)]\mathbf{w}^o + \varepsilon(k) \tag{16}$$

where $\mathbf{w}^o = [1,1]^\mathsf{T}$. The input $x(k) = \sum_{n=0}^{15} b_n u(k-n)$ where $u(k)$ is an i.i.d. Gaussian sequence with zero mean and unit variance. b_n is designed to implement a low-pass filter[3] with normalized cutoff frequency 0.01. $x(k)$ is consequently colored and M-dependent (see Ass. 5 above) with $M = 15$. $\varepsilon(k)$ is an i.i.d. Gaussian noise sequence with zero mean, $\sigma_\varepsilon^2 = 0.2 \cdot E_{x(k)}\{(y^o)^2(k)\}$, and independent of $u(k)$. The model used is incomplete and given by:

$$y(k) = wx(k) + e(k;w) \tag{17}$$

[2]Note that $g - f$ may be equal to a constant which is independent of $\mathbf{x}$. However, this case never occurs if the model contains a bias term.

[3]The design is performed by the MATLAB (The Math Works, Inc.) M-file "fir1" which uses a Hamming windowed ideal impulse response (i.e. sinc(x)).

GEN is compared to two different methods for estimating the generalization error. First, a comparison with the *FPE*-estimate which is much less computationally complex according to Eq. (12) and (15).

Secondly, comparison with the leave-one-out cross-validation method [2], [4] is performed. Within this method training is replicated N times. The j'th training is performed on the data: $\{\mathbf{x}(k), y(k)\}$, $k = 1, 2, \cdots, j-1, j+1, \cdots, N$, $j = 1, 2, \cdots, N$[4] resulting in the estimate $\widehat{\mathbf{w}}^{(j)}$. $e^2(j, \widehat{\mathbf{w}}^{(j)})$ is a qualified estimate of the generalization error and consequently G is estimated as:

$$L = \frac{1}{N} \sum_{j=1}^{N} e^2(j, \widehat{\mathbf{w}}^{(j)}) \tag{18}$$

Knowing the details of the system Eq. (16) it is possible to compute the true generalization error $G(\widehat{w})$ according to Eq. (4). Let $E\{\cdot\}$ denote expectation w.r.t. x_t and ε_t. Now, noting that x_t is Gaussian:

$$\begin{aligned} G(\widehat{w}) &= E\left\{\left[w_1^o x_t + w_2^o x_t^2 + \varepsilon_t - \widehat{w} x_t\right]^2\right\} \\ &= (w_1^o - \widehat{w})^2 E\{x_t^2\} + 3(w_2^o E\{x_t^2\})^2 + \sigma_\varepsilon^2 \end{aligned} \tag{19}$$

In order to simulate the statistical variations of the training sets Q independent training sets: $\{x^{(q)}(k), y^{(q)}(k)\}$, $q = 1, 2, \cdots, Q$, is generated for every specific training set size, N. Next, Q models are estimated by (the LS-estimator):

$$\widehat{w}^{(q)} = \left[\sum_{k=1}^{N} (x^{(q)}(k))^2\right]^{-1} \cdot \sum_{k=1}^{N} x^{(q)}(k) y^{(q)}(k) \tag{20}$$

Let $\widehat{G}$ be a specific generalization error estimator, i.e. *GEN*, *FPE*, *C* or *L*. For the purpose of comparison the *relative average deviation* (*RAD*) is defined as:

$$RAD = \frac{\langle G(\widehat{\mathbf{w}}^{(q)}) \rangle - \langle \widehat{G}(\widehat{\mathbf{w}}^{(q)}) \rangle}{\langle G(\widehat{\mathbf{w}}^{(q)}) \rangle} \tag{21}$$

where $\langle \cdot \rangle$ denotes the average with respect to the Q realizations.

The result of comparing the *RAD*'s of *GEN* and *FPE* is shown in Fig. 1. Averaging was done with:

$$Q = \begin{cases} 30000 & 5 \leq N \leq 9 \\ 20000 & 10 \leq N \leq 170 \end{cases}$$

When N is small compared to M $\mathbf{R}(\tau)$ (cf. Eq. (13)) will be rather noisy. Therefore $\tau = 1, 2, \cdots, \min\{M, [N/10]\}$ was used, where $[\cdot]$ denotes rounding to the nearest integer. Using a standard Gaussian test (details omitted) it

[4] Note that $\min(k) = 2$ for $j = 1$ and $\max(k) = N - 1$ for $j = N$.

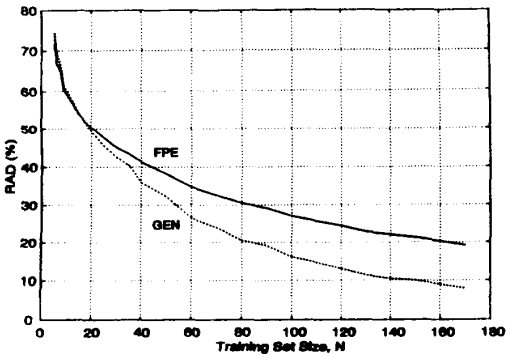

Figure 1: Comparison of the *RAD*'s of *GEN* and *FPE* for the linear model Eq. (17).

is seen that the *RAD* of the *GEN*-estimate is significantly[5] better than the *RAD* of the *FPE*-estimate for all $N > 15$ and roughly equal as $N \leq 15$. However, the *RAD* only shows the average performance. The estimator with the best *RAD* performance may still not be the preferred estimator. In order to elucidate the variations in the estimates the probability that *GEN* is closer than *FPE* to the average of the true generalization, $\langle G \rangle$, was estimated. That is,

$$\gamma = P\{\,|\langle G\rangle - GEN| < |\langle G\rangle - FPE|\,\} \tag{22}$$

It was found that $\gamma > 0.5\ \forall\, N \geq 25$ and $\gamma \approx 0.75$ when $N \geq 40$. Consequently, one may prefer the *GEN*-estimator when $N > 25$.

Next, *GEN* is tested against leave-one-out cross-validation and averaging was done with:

$$Q = \begin{cases} 10000 & 5 \leq N \leq 9 \\ 5000 & 10 \leq N \leq 100 \end{cases}$$

The result is shown in Fig. 2. As $N > 15$ the *GEN*-estimate is significantly better than leave-one-out cross-validation as the *RAD* is lower. Further[6], $\gamma > 0.5$ as $N \geq 30$ and $\gamma \approx 0.75$ for $N \geq 40$. This is in spite of the fact that the computationally complexity of the *L*-estimate normally is greater than that of the *GEN*-estimate. The number of multiplications involved in the computation of the *GEN*-estimate is approximately MNm^2 whereas the *L*-estimate requires in the order of N^2m^2 multiplications (this is due to the fact that training is replicated N times).

Simple Neural Network

Consider a simple nonlinear system which consists of a single neuron:

$$y(k) \;=\; y^{\circ}(k) + \varepsilon(k) = h(\mathbf{x}^{\mathsf{T}}(k)\mathbf{w}^{\circ}) + \varepsilon(k)\,, \tag{23}$$

[5] Here and in the following a 0.5% significance level is employed.
[6] *FPE* is replaced by *L* in Eq. (22).

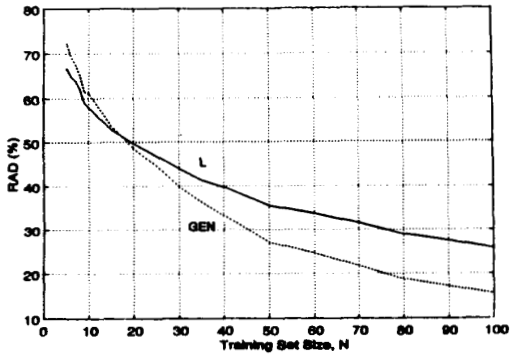

Figure 2: Comparison of the *RAD*'s of *GEN* and leave-one-out cross-validation, L, for the linear model Eq. (17).

$$h(z) = \exp\left(-\left(\frac{z-\nu}{\eta}\right)^2\right) - \exp\left(-\left(\frac{z+\nu}{\eta}\right)^2\right) \tag{24}$$

where $\mathbf{w}^o = [3,3]^T$. Let $\mathbf{u}(k)$ be a two-dimensional i.i.d. Gaussian sequence with zero mean and $E\{u_i^2(k)\} = 1$, $E\{u_1(k)u_2(k)\} = 0.5$. b_n is given as in the preceding subsection and $x_i(k) = \sum_{n=0}^{15} b_n u_i(k-n)$, $i = \{1,2\}$. $\varepsilon(k)$ is an i.i.d. Gaussian noise sequence with zero mean, $\sigma_\varepsilon^2 = 0.1 \cdot E_{\mathbf{x}(k)}\{(y^o)^2(k)\}$, and independent of $u_i(k)$,. The activation function $h(z)$ is chosen to be a sum of two Gausssian functions in order to enable the evaluation of the true generalization error Eq. (4). In this simulation: $\nu = 2$ and $\eta = 1$. The employed incomplete nonlinear model of Eq. (23) is:

$$y(k) = h(wx_1(k)) + e(k;w) \tag{25}$$

According to Eq. (4), (23), and (25) ($E\{\cdot\}$ w.r.t. $[\mathbf{x}_t, \varepsilon_t]$):

$$\begin{aligned} G(\widehat{w}) &= E\left\{\left[\varepsilon_t + h(\mathbf{x}^T(k)\mathbf{w}^o) - h(\widehat{w}x_1(k))\right]^2\right\} \\ &= E\left\{\left[h(\mathbf{x}^T(k)\mathbf{w}^o) - h(\widehat{w}x_1(k))\right]^2\right\} + \sigma_\varepsilon^2 \end{aligned} \tag{26}$$

Evaluation of the first term in Eq. (26) is possible, however, due to the extent of the derivation it is omitted, see [6] for further details.

The parameter w in Eq. (25) is estimated using a modified Gauss-Newton algorithm [9, Ch. 14]. That is, for each training set $\{x_1^{(q)}(k), y^{(q)}(k)\}$, $q = 1,2,\ldots,Q$ (below the q index is omitted for simplicity):

$$w_{(i+1)} = w_{(i)} + \mu\widetilde{\mathbf{H}}_N^{-1}(w_{(i)})\nabla(w_{(i)}), \tag{27}$$

$$\widetilde{\mathbf{H}}_N(w_{(i)}) = \sum_{k=1}^{N}\left[h'(w_{(i)}x_1(k)) \cdot x_1(k)\right]^2, \tag{28}$$

$$\nabla(w_{(i)}) = \sum_{k=1}^{N} h'(w_{(i)}x_1(k)) \cdot x_1(k) \cdot e(k; w_{(i)}) \tag{29}$$

where $0 \leq \mu \leq 1$ is the step-size and $'$ denotes the derivative. For each iteration, i, μ is adjusted in order to ensure: $S_N(w_{(i+1)}) < S_N(w_{(i)})$. The employed stopping criterion [9, Sec. 14.4] was: $(S_N(w_{(i+1)}) - S_N(w_{(i)}))/S_N(w_{(i)}) < 10^{-12}$.

The result of comparing *GEN* to *FPE* is shown in Fig. 3 ($Q = 5000$). It is

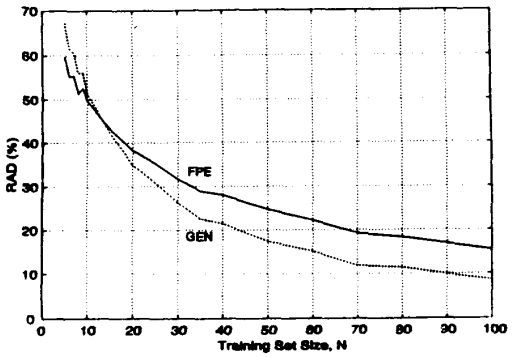

Figure 3: Comparison of the *RAD*'s of *GEN* and *FPE* for the nonlinear model Eq. (25).

observed that the *RAD* of the *GEN*-estimate is significantly better than that of the *FPE*-estimate for all $N > 10$ and that (cf. Eq. (22)) $\gamma > 0.5$ as $N \geq 15$ and $\gamma \approx 0.6$ for $N > 15$.

CONCLUSION

In this paper a new estimate (*GEN*) of the generalization error is presented. The estimator is valid for both ***incomplete*** and ***nonlinear*** models. An incomplete model is characterized in that it does not model the actual nonlinear relationship perfectly. The estimator can be viewed as an extension of the *FPE* and *GPE* estimators [1], [7]. The *GEN*-estimator has been evaluated by simulating incomplete models of linear and simple neural network systems respectively. Within the linear system *GEN* is compared to the Final Prediction Error (*FPE*) criterion and the leave-one-out cross-validation technique. It was found that the *GEN*-estimate of the true generalization error is less biased on the average. Further the probability, γ, of *GEN* being closer to the true generalization error than the other estimators was estimated, and it was found that $\gamma > 0.5$ within a large range of training set sizes. Comparing the *GEN*-estimate to *FPE* when simulating a simple neural network shows that *GEN* is less biased on the average and that $\gamma \approx 0.6$ when using trainig sets of sizes greater than 15. In summary it is concluded that *GEN* is an

applicable alternative in estimating the generalization at the expense of an increased complexity. However, the leave-one-out cross-validation estimate which possess a higher complexity was not able to outperform *GEN* in the chosen example.

ACKNOWLEDGEMENTS

The author would like to thank Lars Kai Hansen, Nils Hoffmann, Peter Koefoed Møller and John Aasted Sørensen for helpfull comments on this paper. This work is partly supported by the Danish Natural Science and Technical Research Councils through the Computational Neural Network Center.

REFERENCES

[1] H. Akaike, "Fitting Autoregressive Models for Prediction," Ann. Inst. Stat. Math., vol. 21, pp. 243–247, 1969.

[2] N.R. Draper & H. Smith, Applied Regression Analysis, New York, New York: JOHN WILEY & SONS, 1981.

[3] D.B. Fogel, "An Information Criterion for Optimal Neural Metwork Selection," IEEE Transactions on Neural Networks, vol. 2, no. 5, Sept. 1991.

[4] L.K. Hansen & P. Salomon, "Neural Network Ensembles," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 10, Oct. 1990.

[5] R. Kannurpatti & G.W. Hart, "Memoryless Nonlinear System Identification With Unknown Model Order," IEEE Transaction on Information Theory, vol. 37, no 5, Sept. 1991.

[6] J. Larsen, Design of Neural Network Filters, Ph.D. Thesis, Electronics Institute, Technical University of Denmark. In Preparation.

[7] J. Moody, "Note on Generalization, Regularization, and Architecture Selection in Nonlinear Learning Systems", in Proceedings of the first IEEE Workshop on Neural Networks for Signal Processing, Eds. B.H. Juang, S.Y. Kung & C.A. Kamm, Piscataway, New Jersey: IEEE, 1991, pp. 1–10.

[8] M. Rosenblatt, Stationary Sequences and Random Fields, Boston, Massachusetts: BIRKHÄUSER, 1985.

[9] G.A.F. Seber & C.J. Wild, Nonlinear Regression, New York, New York: JOHN WILEY & SONS, 1989.