GENERALIZATION PERFORMANCE OF REGULARIZED NEURAL NETWORK MODELS

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

Abstract. Architecture optimization is a fundamental problem of neural network modeling. The optimal architecture is defined as the one which minimizes the generalization error. This paper addresses estimation of the generalization performance of regularized, complete neural network models. Regularization normally improves the generalization performance by restricting the model complexity. A formula for the optimal weight decay regularizer is derived. A regularized model may be characterized by an effective number of weights (parameters); however, it is demonstrated that no simple definition is possible. A novel estimator of the average generalization error (called *FPER*) is suggested and compared to the Final Prediction Error (*FPE*) and Generalized Prediction Error (*GPE*) estimators. In addition, comparative numerical studies demonstrate the qualities of the suggested estimator.

INTRODUCTION

One of the fundamental problems involved in design of neural network models is architecture optimization aiming at high generalization performance. In this paper the generalization measure is defined as the *average generalization error*, i.e., the expected squared error averaged over all possible training sets of size N, with N being the number of training samples. The average generalization error, Γ, can be decomposed into three additive components [2], [8]: $\Gamma = \sigma_\varepsilon^2 + MSME + WFP$, viz. the inherent noise variance, the mean square model error, and the weight fluctuation penalty[1]. The inherent noise variance is caused by noise on the data which – per definition – cannot be modeled.

[1]The *MSME* and the *WFP* are related to the *squared bias* and the *variance*, respectively. See [2] for a definition of bias and variance.

Presence of *MSME* reflects the lack of modeling capability of the neural network for modeling the current data, i.e., the network is an *incomplete model* of the data generating system. Finally, the *WFP* reflects the increase in average generalization error caused by fluctuations in the estimated weights, which stem from the fact that the weights are estimated from a given finite training set.

Architecture optimization can be viewed as a bias/variance trade off [2], [11] or equivalently a *MSME/WFP* trade off: The *MSME* is reduced when increasing the network complexity[2] while the *WFP* typically[3] increases. The literature provides a variety of methods for performing this trade off, including architecture pruning and growing schemes, as well as regularization techniques.

TRAINING AND GENERALIZATION

Consider modeling the data generating system:

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

where k is the discrete time index, $y(k)$ is the scalar output signal, $g(\cdot)$ constitutes a nonlinear mapping of the p-dimensional input signal $\boldsymbol{x}(k)$ (column vector), and $\varepsilon(k)$ is an inherent noise signal.

Assumption 1 *The input signal $\boldsymbol{x}(k)$ is assumed to be a strongly mixing*[4] *strictly stationary sequence and the inherent noise $\varepsilon(k)$ is assumed to be a strictly stationary sequence independent on the input, white, with zero mean, and finite variance, σ_ε^2.*

The neural network model of the system in Eq. (1) is given by

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

where $f(\cdot; \boldsymbol{w})$ defines the mapping of the neural network parameterized by the m-dimensional weight vector $\boldsymbol{w}$, and $e(k; \boldsymbol{w})$ is the error signal.

Assumption 2 *The model is assumed* ***complete*** *[8, Def. 6.3], i.e., there exists a true weight vector, $\boldsymbol{w}^\circ$, so as to*

$$\forall\, \boldsymbol{x} : \quad f(\boldsymbol{x}; \boldsymbol{w}^\circ) \equiv g(\boldsymbol{x}) \tag{3}$$

In general, only little a priori knowledge of the data generating system is available, i.e., most neural network models are *incomplete*, which result in non-zero mean square model error. However, a multi-layer perceptron neural

[2]This statement is only true for nested families of network architectures. Moreover, *MSME* may remain unchanged when adding irrelevant complexity.

[3]It should be emphasized that it is possible to give simple examples where the *WFP* actually *decreases* when adding extra complexity [8, Ch. 6.3.4].

[4]Loosely speaking, i.e., the dependence of $\mathbf{x}(k)$ and $\mathbf{x}(k+\tau)$ vanishes as $|\tau| \to \infty$.

network with many hidden neurons is capable of approximating a large class of functions, thus $MSME$ may be small relative to $\sigma_\varepsilon^2 + WFP$, and the model may be regarded as *quasi-complete*. When dealing with cases where the complete model assumption is dubious, it is suggested to estimate the generalization performance by using the GEN estimator [7], [8].

Define the training set of N samples by $\mathcal{T} = \{\boldsymbol{x}(k); y(k)\}$, $k = 1, 2, \cdots, N$. The model is estimated by minimizing a cost function being the sum of the usual mean square cost and a weight decay regularizer[5]:

$$C_N(\boldsymbol{w}) = S_N(\boldsymbol{w}) + \boldsymbol{w}^\top \boldsymbol{R}\boldsymbol{w} \tag{4}$$

where $S_N(\boldsymbol{w}) = N^{-1}\sum_{k=1}^{N} e^2(k;\boldsymbol{w}) = N^{-1}\sum_{k=1}^{N}\left[y(k) - f(\boldsymbol{x}(k);\boldsymbol{w})\right]^2$ is the mean square cost and $\boldsymbol{R}$ is a $m \times m$ symmetric, positive semidefinite regularization matrix. Standard weight decay regularization is obtained by using $\boldsymbol{R} = \kappa \boldsymbol{I}$, where $\kappa \geq 0$ is the weight decay parameter and $\boldsymbol{I}$ the identity matrix. The presented theory is not restricted to the chosen cost function, thus analogous results can be obtained when e.g., using log-likelihood cost functions and more general regularizers, $r(\boldsymbol{w};\boldsymbol{\kappa})$, where $r(\cdot)$ is a regularization function parameterized by $\boldsymbol{\kappa}$.

The weights of the estimated model are denoted the *estimated weights*, i.e.,

$$\widehat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} C_N(\boldsymbol{w}) \tag{5}$$

Also define the *expected cost function*:

$$C(\boldsymbol{w}) = E\left\{C_N(\boldsymbol{w})\right\} = E\left\{e^2(\boldsymbol{w})\right\} + \boldsymbol{w}^\top \boldsymbol{R}\boldsymbol{w} \tag{6}$$

where $E\{\cdot\}$ denotes expectation w.r.t. the joint input-output probability density function. Under mild regularity conditions (see e.g., [8, Ch. 5], [12]) $\lim_{N\to\infty} C_N(\boldsymbol{w}) = C(\boldsymbol{w})$, and the estimated weight vector $\widehat{\boldsymbol{w}}$ becomes a consistent estimate ($N \to \infty$) of the *optimal weight vector*: $\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} C(\boldsymbol{w})$. Since the model is assumed complete $\boldsymbol{w}^*$ is identical to $\boldsymbol{w}^\circ$ when omitting regularization. However, regularization imposes a bias of the optimal weights towards $\mathbf{0}$.

The *generalization error* of the estimated model is defined as the expected squared error on an test sample, $[\boldsymbol{x}; y]$, independent on the training samples, i.e.,

$$G(\widehat{\boldsymbol{w}}) = E\left\{e^2(\widehat{\boldsymbol{w}})\right\} = E\left\{\left[y - f(\boldsymbol{x};\widehat{\boldsymbol{w}})\right]^2\right\} \tag{7}$$

It turns out (see e.g., the discussion in [8, Sec. 6.3.2]) that $G(\widehat{\boldsymbol{w}})$ is not necessarily a reliable measure of the model quality since it depends on the *actual training set* through $\widehat{\boldsymbol{w}}$. In addition, it is not possible to obtain estimates of $G(\widehat{\boldsymbol{w}})$ without perfect knowledge of the joint input-output distribution. Hence, the appropriate model quality measure is the *average generalization error*, e.g., [8], [11]:

$$\Gamma = E_{\mathcal{T}}\left\{G(\widehat{\boldsymbol{w}})\right\} \tag{8}$$

[5] $^\top$ denotes the transpose operator.

where $E_{\mathcal{T}}\{\cdot\}$ denotes expectation over all training sets with N samples. That is, averaging is w.r.t. fluctuation in $\widehat{\boldsymbol{w}}$ due to different training sets. Define $\mathcal{T}_{\boldsymbol{x}} = \{\boldsymbol{x}(k)\}$ and $\mathcal{T}_{\varepsilon} = \{\varepsilon(k)\}$. As the noise and the input are assumed independent, the expectation w.r.t. $\mathcal{T}$ is carried out as[6]:

$$E_{\mathcal{T}}\{G\} = E_{\mathcal{T}_{\boldsymbol{x}}}\left\{E_{\mathcal{T}_{\varepsilon}}\{G \,|\, \mathcal{T}_{\boldsymbol{x}}\}\right\} \tag{9}$$

ESTIMATING THE AVERAGE GENERALIZATION ERROR

The objective of this presentation is to obtain an estimate of Γ defined in Eq. (8) calculated in terms of quantities derived from the estimated model. From a statistical point of view it is possible to set different quality requirements on the estimator. Here the following requirements are made:

Definition 1 *The estimator searched for, $\widehat{\Gamma}$, is required to be consistent, and unbiased to order $1/N$, i.e., $\widehat{\Gamma} \to \Gamma$ as $N \to \infty$, and $E_{\mathcal{T}}\{\widehat{\Gamma}\} = \Gamma + o(1/N)$, where $o(\cdot)$ is the order function.*

The basic tool for deriving an estimator are second order Taylor series expansions of the average training and generalization errors, as follows:

$$\begin{aligned} E_{\mathcal{T}}\{S_N(\widehat{\boldsymbol{w}})\} &\approx E_{\mathcal{T}}\{S_N(\boldsymbol{w}^\circ)\} + E_{\mathcal{T}}\left\{\frac{\partial S_N(\boldsymbol{w}^\circ)}{\partial \boldsymbol{w}^\top}\Delta\boldsymbol{w}\right\} \\ &\quad + E_{\mathcal{T}}\left\{\Delta\boldsymbol{w}^\top \boldsymbol{H}_N(\boldsymbol{w}^\circ)\Delta\boldsymbol{w}\right\} \end{aligned} \tag{10}$$

$$\begin{aligned} E_{\mathcal{T}}\{G(\widehat{\boldsymbol{w}})\} &\approx E_{\mathcal{T}}\{G(\boldsymbol{w}^\circ)\} + E_{\mathcal{T}}\left\{\frac{\partial G(\boldsymbol{w}^\circ)}{\partial \boldsymbol{w}^\top}\Delta\boldsymbol{w}\right\} \\ &\quad + E_{\mathcal{T}}\left\{\Delta\boldsymbol{w}^\top \boldsymbol{H}(\boldsymbol{w}^\circ)\Delta\boldsymbol{w}\right\} \end{aligned} \tag{11}$$

where $\Delta\boldsymbol{w}$ is the weight fluctuation $\Delta\boldsymbol{w} = \widehat{\boldsymbol{w}} - \boldsymbol{w}^\circ$, $\boldsymbol{H}_N(\boldsymbol{w})$ is the Hessian matrix of the mean square cost function, i.e.,

$$\boldsymbol{H}_N(\boldsymbol{w}) = \frac{1}{2}\frac{\partial^2 S_N(\boldsymbol{w})}{\partial \boldsymbol{w}\partial \boldsymbol{w}^\top} = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{\psi}(k;\boldsymbol{w})\boldsymbol{\psi}^\top(k;\boldsymbol{w}) - \boldsymbol{\Psi}(k;\boldsymbol{w})e(k;\boldsymbol{w}) \tag{12}$$

defining $\boldsymbol{\psi}$ as the instantaneous gradient vector of the model output, $\boldsymbol{\psi}(k;\boldsymbol{w}) = \partial f(\boldsymbol{x}(k);\boldsymbol{w}) \,/\, \partial \boldsymbol{w}$. Finally, $\boldsymbol{\Psi}$ is the second derivative matrix of the model output, $\boldsymbol{\Psi}(k;\boldsymbol{w}) = \partial\boldsymbol{\psi}(k;\boldsymbol{w})/\partial\boldsymbol{w}^\top$. Similarly, $\boldsymbol{H}(\boldsymbol{w})$ is the Hessian matrix of the generalization error, given by

$$\boldsymbol{H}(\boldsymbol{w}) = \frac{1}{2}\frac{\partial^2 G(\boldsymbol{w})}{\partial \boldsymbol{w}\partial \boldsymbol{w}^\top} = E\left\{\boldsymbol{\psi}(\boldsymbol{w})\boldsymbol{\psi}^\top(\boldsymbol{w}) - \boldsymbol{\Psi}(\boldsymbol{w})e(\boldsymbol{w})\right\} \tag{13}$$

In order to ensure the validity of the Taylor series approximations it is required that $\Delta\boldsymbol{w}$ is sufficiently small. As mentioned above $\widehat{\boldsymbol{w}}$ is a consistent estimate

[6] Note that expectation over the training set, $\mathcal{T} = \{\mathrm{x}(k); y(k)\}$, equals expectation over input and inherent noise samples, cf. the model definition Eq. (2).

of $\boldsymbol{w}^*$; however, $\boldsymbol{w}^*$ does not collapse onto $\boldsymbol{w}^\circ$ unless $\boldsymbol{R} = \boldsymbol{0}$. Consequently, it is expected that the Taylor series are valid for sufficiently large N and sufficiently small $\boldsymbol{R}$.

The appendix below provides a brief evaluation of the individual terms of Eq. (10), (11). The result is: For $N > 2m_1 - m_2$,

$$E_{\mathcal{T}}\left\{S_N(\widehat{\boldsymbol{w}})\right\} = \sigma_\varepsilon^2\left(1 - \frac{2m_1 - m_2}{N}\right) + M' + o(1/N) \tag{14}$$

$$\Gamma = \sigma_\varepsilon^2\left(1 + \frac{m_2}{N}\right) + M + o(1/N) \tag{15}$$

where m_1, m_2 defines two different *effective number of weights*[7]:

$$m_1 = \operatorname{tr}\left[\boldsymbol{H}(\boldsymbol{w}^\circ)\boldsymbol{J}^{-1}(\boldsymbol{w}^\circ)\right], \; m_2 = \operatorname{tr}\left[\boldsymbol{H}(\boldsymbol{w}^\circ)\boldsymbol{J}^{-1}(\boldsymbol{w}^\circ)\boldsymbol{H}(\boldsymbol{w}^\circ)\boldsymbol{J}^{-1}(\boldsymbol{w}^\circ)\right] \tag{16}$$

$\boldsymbol{J}(\boldsymbol{w}) = \boldsymbol{H}(\boldsymbol{w}) + \boldsymbol{R}$ is the Hessian matrix of the expected cost function which is assumed to be invertible, and $\operatorname{tr}[\cdot]$ is the trace operator.

$$M' = (\boldsymbol{w}^\circ)^\top \boldsymbol{R}\boldsymbol{J}^{-1}(\boldsymbol{w}^\circ)\left(\boldsymbol{H}(\boldsymbol{w}^\circ) + \frac{-2\boldsymbol{K}_1 + \boldsymbol{K}_2}{N}\right)\boldsymbol{J}^{-1}(\boldsymbol{w}^\circ)\boldsymbol{R}\boldsymbol{w}^\circ \tag{17}$$

with $\boldsymbol{K}_1$, $\boldsymbol{K}_2$ being 4th order moments, as shown by[8]:

$$\boldsymbol{K}_1 = E\left\{\left(\boldsymbol{\psi}\boldsymbol{\psi}^\top - \boldsymbol{H}\right)\boldsymbol{J}^{-1}\left(\boldsymbol{\psi}\boldsymbol{\psi}^\top - \boldsymbol{H}\right)\right\} \tag{18}$$

$$\boldsymbol{K}_2 = E\left\{\left(\boldsymbol{\psi}\boldsymbol{\psi}^\top - \boldsymbol{H}\right)\boldsymbol{J}^{-1}\boldsymbol{H}\boldsymbol{J}^{-1}\left(\boldsymbol{\psi}\boldsymbol{\psi}^\top - \boldsymbol{H}\right)\right\} \tag{19}$$

M equals M' except that the term $\boldsymbol{K}_1$ is absent. In general, M and M' are negligible compared to the remaining terms in Eq. (14), (15) when 1) using a regularization matrix close to the optimal setting Eq. (24), and when 2) the signal-to-noise ratio, $V\{g(\boldsymbol{x})\}/\sigma_\varepsilon^2$, is reasonable large.

Neglecting M, M' and eliminating σ_ε^2 in Eq. (14), (15) leads to:

$$\widehat{\Gamma} = \frac{N + m_2}{N - 2m_1 + m_2} E_{\mathcal{T}}\left\{S_N(\widehat{\boldsymbol{w}})\right\}, \quad N > 2m_1 - m_2 \tag{20}$$

which is unbiased to $o(1/N)$. Notice that elimination of σ_ε^2 introduces terms proportional to N^{-j}, $j > 1$. This seems inconsistent; however, for practical purposes the form is convenient since $\widehat{\Gamma}$ typically is an underestimate of Γ on the average. In the case of a complete linear model which is estimated without regularization [3] and [8, Theorem 6.10] support this statement.

The suggested estimator may be viewed as an extension of the classical *FPE* estimator [1], $FPE = E_{\mathcal{T}}\left\{S_N(\widehat{\boldsymbol{w}})\right\}(N + m)/(N - m)$, in which the

[7] It is easily shown that $m_1 \geq m_2 > 0$ thus $2m_1 - m_2 > 0$. Moreover, 1) $m_1 = m_2 = m$ for $\mathbf{R} = \mathbf{0}$ and $\mathbf{H}(\mathbf{w}^\circ)$ non-singular, and 2) $m_1 \to 0$, $m_2 \to 0$ as $\|\mathbf{R}\| \to \infty$.

[8] $\mathbf{K}_1$, $\mathbf{K}_2$ are of order one, and limited by assumption. Further note that all involved quantities are evaluated at $\mathbf{w}^\circ$.

number of weights m is replaced by the different effective number of weights, m_2 and $2m_1 - m_2$. Moreover, the estimator can be interpreted as a special version[9] of the *GPE* estimator [10], [11] where the inherent noise variance is estimated by: $\sigma_\varepsilon^2 = E_{\mathcal{T}} \{S_N(\widehat{\boldsymbol{w}})\} N/(N - 2m_1 + m_2)$. In order to construct a Γ-estimator from observable quantities, estimation of the noise variance is indeed important. This problem is not directly addressed in [10], [11]. The estimator suggested in [10] reads: $\sigma_\varepsilon^2 = E_{\mathcal{T}} \{S_N(\widehat{\boldsymbol{w}})\} N/(N - m_1)$, which obviously differs from the one derived from Eq. (14). In conclusion – as suggested in [9], [11] – it is not possible to define a single quantity m_1 which expresses the effective number of weights in the model, since σ_ε^2 should be estimated from $2m_1 - m_2$ rather than m_1 effective weights.

For practical purposes the quantities in Eq. (20) are estimated from observed quantities. An unbiased $o(1/N)$ estimator within the second order Taylor series expansion Eq. (10), (11) is the the **F**inal **P**rediction **E**rror estimator for **R**egularized models, as shown by:

$$FPER = \frac{N + \widehat{m}_2}{N - 2\widehat{m}_1 + \widehat{m}_2} S_N(\widehat{\boldsymbol{w}}), \quad N > 2\widehat{m}_1 - \widehat{m}_2 \tag{21}$$

where

$$\widehat{m}_1 = \mathrm{tr}\left[\boldsymbol{H}_N(\widehat{\boldsymbol{w}})\boldsymbol{J}_N^{-1}(\widehat{\boldsymbol{w}})\right], \ \widehat{m}_2 = \mathrm{tr}\left[\boldsymbol{H}_N(\widehat{\boldsymbol{w}})\boldsymbol{J}_N^{-1}(\widehat{\boldsymbol{w}})\boldsymbol{H}_N(\widehat{\boldsymbol{w}})\boldsymbol{J}_N^{-1}(\widehat{\boldsymbol{w}})\right] \tag{22}$$

and $\boldsymbol{J}_N(\widehat{\boldsymbol{w}}) = \boldsymbol{H}_N(\widehat{\boldsymbol{w}}) + \boldsymbol{R}$ is the Hessian matrix of the cost function which is assumed to be invertible.

OPTIMIZING THE WEIGHT DECAY REGULARIZATION PARAMETER

For simplicity, consider simple weight decay regularization, i.e., $\boldsymbol{R} = \kappa \boldsymbol{I}$ where κ is the weight decay parameter. As mentioned in the introduction, trading off weight fluctuation penalty (*WFP*) and mean square model error (*MSME*) leads to an optimal setting of κ. In [6] this problem was addressed for linear models and the following may be viewed as an extension of this work.

Inspecting Eq. (15) it turns out that[10] $M = MSME$ and $WFP = \sigma_\varepsilon^2 m_2/N$. The optimal value, κ_{opt}, is found by solving:

$$\frac{\partial WFP}{\partial \kappa} + \frac{\partial MSME}{\partial \kappa} = 0 \tag{23}$$

As expected, $\lim_{N\to\infty} WFP = 0$, since it measures the contribution due to a finite training set. Consequently, in order to reach the minimal average generalization error $\Gamma = \sigma_\varepsilon^2$ the restriction $\lim_{N\to\infty} MSME = 0$ should be met. The κ-dependence of the individual elements of $\boldsymbol{K}_1$ is $(\lambda_i + \kappa)^{-1}$

[9] Notice that this coincidence is based on various important assumptions, e.g., the model being complete and the negligibleness of M.

[10] Notice when determining an optimal κ, M is *not* neglected in Eq. (15).

where λ_i is the i'th eigenvalue of $\boldsymbol{H}(\boldsymbol{w}^\circ)$. For $\boldsymbol{K}_2$ the element dependence is: $\prod_{s=1}^{2}(\lambda_{i_s}+\kappa)^{-1}$. In summary, M is a sum of addends which κ-dependence are given by: $\kappa^2\prod_{r=1}^{\ell}(\lambda_{i_r}+\kappa)^{-1}$, $\ell \in \{2,3,4\}$. That is, to fulfill the requirement $\lim_{N\to\infty} \mathit{MSME} = 0$, $\lim_{N\to\infty}\kappa = 0$ should be imposed. The solution to Eq. (23) can therefore be expressed as: $\kappa_{\mathrm{opt}} = \kappa'_{\mathrm{opt}}/N + o(1/N)$. Expanding the addends of Eq. (23) to first order in κ and $1/N$ and solving for κ gives:

$$\kappa_{\mathrm{opt}} = \frac{\sigma_\varepsilon^2}{N}\cdot\frac{\mathrm{tr}\,\boldsymbol{H}^+(\boldsymbol{w}^\circ)}{(\boldsymbol{w}^\circ)^\top\,\boldsymbol{H}^+(\boldsymbol{w}^\circ)\boldsymbol{w}^\circ} + o(1/N) \tag{24}$$

where $\boldsymbol{H}^+(\boldsymbol{w}^\circ)$ is the Moore-Penrose pseudo inverse. Suppose the eigenvalues of $\boldsymbol{H}(\boldsymbol{w}^\circ)$ obey: $\lambda_1 \geq \cdots \geq \lambda_n > 0$ and $\lambda_i = 0, \forall\, i \in [n{+}1; m]$. The associated eigenvectors are assembled (as column vectors) in the matrix $\boldsymbol{Q}$. The pseudo inverse then reads: $\boldsymbol{H}^+(\boldsymbol{w}^\circ) = \boldsymbol{Q}\,\mathrm{diag}[\lambda_1^{-1},\cdots,\lambda_n^{-1},0,\cdots,0]\,\boldsymbol{Q}^\top$.

Notice two facts concerning κ_{opt}: First, it is proportional to the inherent noise variance. If no noise is present $\mathit{WFP} = 0$, thus one should not introduce *MSME* by employing a non-zero κ. Secondly, κ_{opt} is inversely proportional to the length of the optimal weight vector weighted by the elements of the Moore-Penrose pseudo inverse Hessian matrix. This is due to the fact that we regularize against the zero weight vector.

Since the optimal weights $\boldsymbol{w}^\circ$ are unknown, it is impossible to calculate κ_{opt} directly; however, in [4] *adaptive regularization* is studied for a linear one-dimensional model, and [5] presents an adaptive regularization scheme for the purpose of designing compact time series models. In addition, it is possible to show that the average generalization error is reduced when using $0 < \kappa \leq 2\kappa_{\mathrm{opt}}$.

NUMERICAL EXPERIMENTS

To substantiate the qualities of the suggested *FPER* estimator Eq. (21), numerical comparisons with the *FPE* and *GPE* estimators[11],

$$\mathit{FPE} = \frac{N+m}{N-m}S_N(\widehat{\boldsymbol{w}}) \qquad \mathit{GPE} = \frac{N+\widehat{m}_1}{N-\widehat{m}_1}S_N(\widehat{\boldsymbol{w}}) \tag{25}$$

is – for convenience – performed for a linear model. The linear data generating system (dimension $m = 15$) is given by:

$$y(k) = \boldsymbol{x}^\top(k)\boldsymbol{w}^\circ + \varepsilon(k) \tag{26}$$

where $\boldsymbol{x}(k)$ is an i.i.d. Gaussian distributed sequence with zero mean and, the elements of $\boldsymbol{H} = E\{\boldsymbol{x}\boldsymbol{x}^\top\}$ are selected randomly, resulting in an eigenvalue-spread approx. equal to 900. The optimal weights are drawn independently from a standard Gaussian distribution. The inherent noise is a Gaussian

[11] As regards the *GPE* estimator, the noise variance estimation suggested in [10] is employed.

zero mean, i.i.d. sequence which is independent of the input with variance $\sigma_\varepsilon^2 = 0.25 \cdot E\left\{ \left(\boldsymbol{x}^\top(k) \boldsymbol{w}^\circ \right)^2 \right\} = 0.25 \cdot (\boldsymbol{w}^\circ)^\top \boldsymbol{H} \boldsymbol{w}^\circ$. That is, the signal-to-noise ratio equals approx. 6 dB.

$Q = 2.4 \cdot 10^4$ independent training sets of size N in the interval $[15; 35]$ were randomly generated, and the weights of the associated model were estimated using a simple weight decay regularizer with $\kappa = 2\kappa_{\mathrm{opt}}$.

The "true" average generalization error was estimated by $\widehat{\Gamma}_G = \langle G(\widehat{\boldsymbol{w}}) \rangle$ where $\langle \cdot \rangle$ denotes the average w.r.t. the Q training sets, and

$$G(\widehat{\boldsymbol{w}}) = E\left\{ \left[\varepsilon + \boldsymbol{x}^\top (\boldsymbol{w}^\circ - \widehat{\boldsymbol{w}}) \right]^2 \right\} = \sigma_\varepsilon^2 + (\boldsymbol{w}^\circ - \widehat{\boldsymbol{w}})^\top \boldsymbol{H} (\boldsymbol{w}^\circ - \widehat{\boldsymbol{w}}) \qquad (27)$$

The quality of the estimators[12], $\widehat{\Gamma}(\mathcal{T}) \in \{FPER, FPE, GPE\}$, is quantified by three different measures:

$$NB = \frac{\widehat{\Gamma}(\mathcal{T}) - \widehat{\Gamma}_G}{\widehat{\Gamma}_G} \cdot 100\% \qquad NRMSE = \frac{\sqrt{\left\langle \left[\widehat{\Gamma}(\mathcal{T}) - \widehat{\Gamma}_G \right]^2 \right\rangle}}{\widehat{\Gamma}_G} \cdot 100\% \quad (28)$$

$$\Pi\left(\widehat{\Gamma}\right) = \left\langle \mu\left(\left| \widehat{\Gamma}(\mathcal{T}) - \widehat{\Gamma}_G \right| - \left| FPER(\mathcal{T}) - \widehat{\Gamma}_G \right| \right) \right\rangle \qquad (29)$$

NB is the normalized bias, $NRMSE$ is the normalized root mean square error, and Π is the probability that $FPER$ is closer to the true estimate, $\widehat{\Gamma}_G$, than another estimator, $\widehat{\Gamma}$. Here $\mu(\cdot)$ denotes the step function. Fig. 1 shows plots of the considered measures. NB of $FPER$ is smallest for all training set sizes; however, as the training set size approaches infinity all estimates becomes identical as $\kappa_{\mathrm{opt}} \to 0$. For $N = 35$ $NB(FPER)$ is approx. half the $NB(GPE)$. The $NRMSE$'s of and GPE are approx. identical, thus one could claim that the normalized bias improvement of $FPER$ relative to GPE is lost at increased variance[13]. However, the probability that $FPER$ is closer than GPE to the true Γ is around 0.65; consequently, $FPER$ should be preferred to GPE. FPE shows extremely bad performance in all figures and moreover, FPE is *negative*, possibly infinite when $N \leq 15$.

CONCLUSION

This paper presented an consistent and $o(1/N)$ unbiased estimator of the average generalization error for a complete neural network model, called $FPER$. The network is trained by using a cost function which is the sum of the mean square error and a quadratic regularization term. The estimator may be viewed as an extension of the FPE and GPE estimators [1], [10]. It turns out that the complexity reduction obtained by using regularization is expressed in terms of *two* distinct effective number of weights, unlike defining a single

[12] Notice, the dependence on the particular training set, $\mathcal{T}$, is emphasized.

[13] That is, mean square error the minus squared bias.

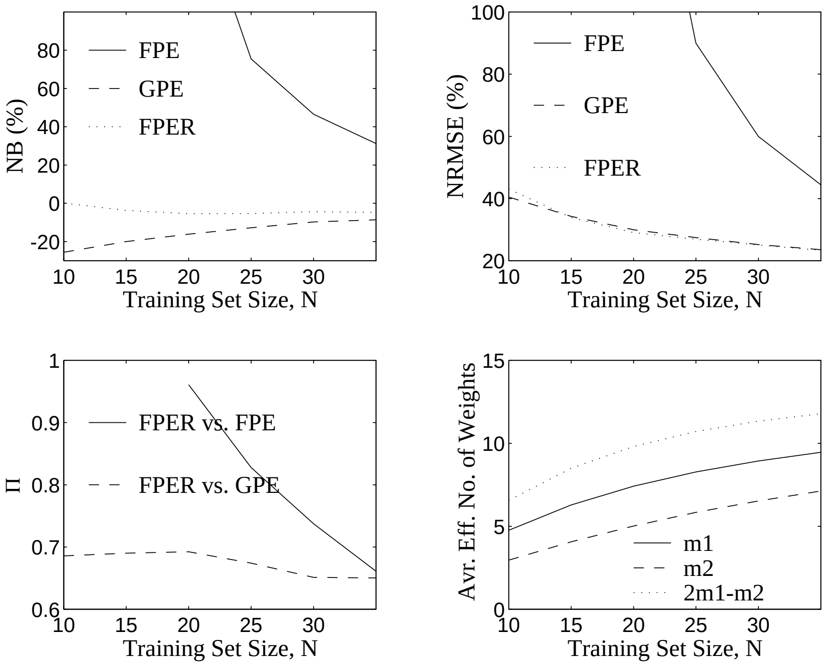

Figure 1: Comparison of FPE, GPE and $FPER$. The FPE curves are not calculated for $N \leq 15$, and the upper panels are cutoff at $+100\%$. The bottom right panel shows the average effective number of weights $\langle \widehat{m}_1 \rangle$, $\langle \widehat{m}_2 \rangle$ as well as $2\langle \widehat{m}_1 \rangle - \langle \widehat{m}_2 \rangle$.

quantity reflecting the effective number of weights, as suggested in [9], [11]. Moreover, an expression for the optimal weight decay parameter is presented and discussed. The potential of the $FPER$ estimator was demonstrated by comparative numerical studies.

ACKNOWLEDGMENTS

This work was supported by the Danish Natural Science and Technical Research Councils through the Computational Neural Network Center.

APPENDIX

Evaluation of the terms in Eq. (10), (11) is based on two observations: First, $\partial C_N(\widehat{\boldsymbol{w}}) \,/\, \partial \boldsymbol{w} = \boldsymbol{0}$ since $\widehat{\boldsymbol{w}}$ minimizes $C_N(\boldsymbol{w})$. A first order Taylor series expansion of $\partial C_N(\widehat{\boldsymbol{w}}) \,/\, \partial \boldsymbol{w}$ reads[14]:

$$\frac{\partial C_N(\boldsymbol{w}^\circ)}{\partial \boldsymbol{w}} + \frac{\partial^2 C_N(\boldsymbol{w}^\circ)}{\partial \boldsymbol{w} \partial \boldsymbol{w}^\top} \Delta \boldsymbol{w} = \boldsymbol{0} \tag{30}$$

[14] Expanding beyond first order result in 3rd and higher order derivatives of the cost function which already are assumed to be negligible.

Subsequently, a few algebraic manipulations result in:

$$\Delta \boldsymbol{w} = \boldsymbol{J}_N^{-1}(\boldsymbol{w}^\circ) \left[\frac{1}{N} \sum_{k=1}^{N} \boldsymbol{\psi}(k; \boldsymbol{w}^\circ) \varepsilon(k) - \boldsymbol{R}\boldsymbol{w}^\circ \right] \tag{31}$$

where $\boldsymbol{J}_N(\boldsymbol{w}^\circ)$ is the non-singular Hessian matrix of the cost function.

The second observation is an expansion of the inverse Hessian obtained by repeatedly using the matrix inversion lemma [8, App. A,B]. The result is: $\boldsymbol{J}_N^{-1}(\boldsymbol{w}^\circ) = \boldsymbol{J}^{-1}(\boldsymbol{w}^\circ) - N^{-1} \cdot \boldsymbol{J}^{-1}(\boldsymbol{w}^\circ)\boldsymbol{\Theta}\boldsymbol{J}^{-1}(\boldsymbol{w}^\circ) + \cdots$, where $\boldsymbol{\Theta} = \boldsymbol{H}_N(\boldsymbol{w}^\circ) - \boldsymbol{H}(\boldsymbol{w}^\circ)$.

REFERENCES

[1] H. Akaike, "Fitting Autoregressive Models for Prediction," Annals of the Institute of Statistical Mathematics, vol. 21, pp. 243–247, 1969.

[2] S. Geman, E. Bienenstock & R. Doursat, "Neural Networks and the Bias/Variance Dilemma," Neural Computation, vol. 4, pp. 1–58, 1992.

[3] L.K. Hansen, "Stochastic Linear Learning: Exact Test and Training Error Averages," Neural Networks, vol. 6, pp. 393–396, 1993.

[4] L.K. Hansen & C.E. Rasmussen, "Pruning from Adaptive Regularization," Preprint Electronics Institute, The Technical University of Denmark, 1993. Accepted for publication in Neural Computation.

[5] L.K. Hansen, C.E. Rasmussen, C. Svarer, & J. Larsen, "Adaptive Regularization," in Proceedings of the 1994 IEEE NNSP Workshop.

[6] A. Krogh & J.A. Hertz, "A Simple Weight Decay Can Improve Generalization," in J.E. Moody, S.J. Hanson, R.P. Lippmann (eds.) Advances in Neural Information Processing Systems 4, Proceedings of the 1991 Conference, San Mateo, California: Morgan Kaufmann Publishers, 1992, pp. 950–957.

[7] J. Larsen, "A Generalization Error Estimate for Nonlinear Systems," in S.Y. Kung, F. Fallside, J. Aa. Sørensen & C.A. Kamm (eds.) Neural Networks for Signal Processing 2: Proceedings of the 1992 IEEE-SP Workshop, Piscataway, New Jersey: IEEE, 1992, pp. 29–38.

[8] J. Larsen, Design of Neural Network Filters, Ph.D. Thesis, Electronics Institute, The Technical University of Denmark, March 1993.

[9] D. MacKay, "A Practical Bayesian Framework for Backprop Networks," Neural Computation, vol. 4, pp. 448-472, 1992.

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

[11] J. Moody, "The Effective Number of Parameters: An Analysis of Generalization and Regularization in Nonlinear Learning Systems," in J.E. Moody, S.J. Hanson, R.P. Lippmann (eds.) Advances in Neural Information Processing Systems 4, Proceedings of the 1991 Conference, San Mateo, California: Morgan Kaufmann Publishers, 1992, pp. 847–854.

[12] H. White, "Consequences and Detection of Misspecified Nonlinear Regression Models," Journal of the American Statistical Association, vol. 76, no. 374, pp. 419–433, June 1981.