USING VARIABILITY TO ANALYSE THE EFFECTS OF ADDING JITTER

JÜRGEN VAN GORP
Department ELEC, Vrije Universiteit Brussel
Pleinlaan 2, B-1050 Brussel, Belgium
Email: Jurgen.Van.Gorp@vub.ac.be

Abstract: This paper starts with the definition of the variability concept, next to the variance and bias of a model. The variability is then used to analyse the much used method of adding *jitter*. Within the domain of Neural Network (NN) identification, there is a folklore that increasing the number of measurements by adding noise to the existing measurements (generally called jitter), will help to obtain a better generalization of the NN model. This is however not guaranteed. Usually the NN models are trained using a simple Least Squares (LS) cost function. In that case it can be shown that adding jitter indeed leads to a reduction of the variability of a model. The variance of the model, however, is still the same, while bias errors are introduced.

I. INTRODUCTION

Despite of the many Black Box (BB) models used in the literature, a model is basically a compression algorithm. The goal is the exact description of a nonlinear surface or hyperplane with a small parameter vector θ.

As a rule of thumb it is said that at least ten to twenty times more measurements must be taken than the number of parameters used in the model. Assume the particular case where an experimenter expects his BB model to perform bad in a certain region, e.g. caused by a lack of measurements available in that region. Additional measurements are not possible but still the experimenter wants his model to perform well. Usually it is demanded that the model should have a good generalization, but in practise the demand is reduced to the need that the model should have a smooth behaviour in between measurement points.

Definition 1: Generalization is the ability to predict future observations, using new data, even when the model was optimized using noisy or sparse data. □

Bad generalization can also be cause by biasing effects, caused by mapping a nonlinear model with noisy input measurements. The best way to guarantee a good generalization, therefore, is to obtain enough noiseless measurements. The modelled system must be measured in the whole domain of interest with a sufficient density such that detailed information is extracted from the hyperplane to be modelled.

To define this domain of interest, consider a robot setup where a gripper is positioned by a collection of individually controlled motors. The combination of all possible motor positions results in various positions and angles for the grippers. Call the whole set of all possible gripper positions and angles and the set of all according motor positions the domain of interest, and let $\psi \subset \mathbb{R}^D$ denote this domain of interest with dimension D.

To explain some of the concepts in the sequel of this paper, it is needed to make a distinction between variance and variability of a model. Consider a plant with a true linear or nonlinear transfer function $\boldsymbol{y}^* = \boldsymbol{g}(\boldsymbol{u}^*)$ with exact inputs $\boldsymbol{u}^* \in \mathbb{R}^m$ and outputs $\boldsymbol{y}^* \in \mathbb{R}^n$. The plant is measured N times with noisy measurements $\boldsymbol{u}_k = \boldsymbol{u}_k^* + \boldsymbol{n}_{u,k}$ and $\boldsymbol{y}_k = \boldsymbol{y}_k^* + \boldsymbol{n}_{y,k}$. A large number of BB models $\boldsymbol{f}_{BB}(\boldsymbol{u}, \theta_i)$, with θ_i the BB parameters and $i = 1, 2, ..., K$, is mapped on the measurement data. The different model parameters are obtained using different initial values for the optimization vector.

Remark: In this paper a BB model is considered in general, and a NN model as a particular example.

Definition 2: The mean BB mapping is defined as

$$\overline{f_{BB}(\boldsymbol{u})} = \lim_{K \to \infty} \left(\frac{1}{K} \sum_{i=1}^{K} \boldsymbol{f}_{BB}(\boldsymbol{u}, \theta_i) \right) \Bigg|_{\forall \boldsymbol{u} \in \psi} . \tag{1}$$

with ψ the domain of interest, given above. □

Definition 3: The variance of a BB model is defined as the variance of an infinite number of BB models that are mapped on the measurement data, when these BB models are compared to each other, or

$$\sigma^2_{f_{BB}} = \lim_{K \to \infty} \left(\frac{1}{K+1} \sum_{i=1}^{K} (\boldsymbol{f}_{BB}(\boldsymbol{u}, \theta_i) - \overline{f_{BB}(\boldsymbol{u})})^2 \right). \tag{2}$$

□

A high variance can result when a BB model with a high number of parameters is mapped on sparse data. Some parameters possibly only operate on a region where no measurements were taken, leading to unprotected behaviour in these regions. Another cause for high variance is the use of an early-stopping algorithm to avoid overfitting on noisy data. If different BB models were mapped starting from random initial parameters, it is highly unlikely that the optimization of the models leads to the same input-output behaviour in all cases, i.e. there is a variance on the model.

Definition 4: The variability of a BB model is defined as the variations on the model output, compared with the true plant, while the mean error is zero, or

$$\nu_{f_{BB}}^2 = \lim_{N\to\infty}\left(\frac{1}{N+1}\sum_{k=1}^{N}(\boldsymbol{y}_k^* - \boldsymbol{f}_{BB}(\boldsymbol{u}_k,\theta))^2\right)\Bigg|_{\lim_{N\to\infty}\left(\frac{1}{N}\sum_{k=1}^{N}\boldsymbol{y}_k^* - \boldsymbol{f}_{BB}(\boldsymbol{u}_k,\theta)\right) = 0} \quad . \tag{3}$$

It is assumed that the different measurement pairs $(\boldsymbol{u}_k, \boldsymbol{y}_k)$ are uniformly spread over the domain of interest. □

Remark that the BB model parameters are optimized using the $(\boldsymbol{u}_k, \boldsymbol{y}_k)$ measurement pairs, while the true outputs $\boldsymbol{y}_k^*$ are expected not known. A typical example of variability is caused by overfitting of measurements with additive zero mean noise on the outputs.

Definition 5: A BB mapping is said to be biased when the mean BB output $\overline{f_{BB}(\boldsymbol{u}^*)}$ does not map the true function $\boldsymbol{g}(\boldsymbol{u}^*)$. The bias measure can then be calculated in the same manner as the calculation of the variability:

$$\beta_{f_{BB}}^2 = \lim_{N\to\infty}\left(\frac{1}{N+1}\sum_{k=1}^{N}(\boldsymbol{y}_k^* - \boldsymbol{f}_{BB}(\boldsymbol{u}_k,\theta_i))^2\right)\Bigg|_{\lim_{N\to\infty}\left(\frac{1}{N}\sum_{k=1}^{N}\boldsymbol{y}_k^* - \boldsymbol{f}_{BB}(\boldsymbol{u}_k,\theta_i)\right) \neq 0} \tag{4}$$

in which the only difference is that the mean error doesn't need to be zero. □

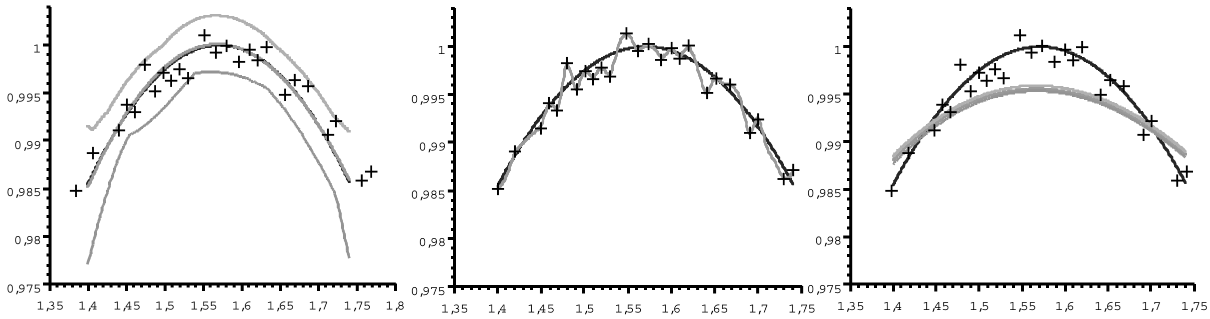

Fig. 1 Examples of NN mappings with [LEFT] high variance and zero bias, [MIDDLE] high variability but low variance, and [RIGHT] low variance, low variability but high bias. The solid line is the true function, the gray lines the mean, minimum and maximum values of 100 NN mappings.

Remark that, given a set of measurements, a BB model can have a very high variability with zero variance, thus mapping all measurement data points exactly during each optimization. Meanwhile a good model for the true plant is never reached and the model could be of little use, e.g. when used for simulation purposes.

If a BB model is optimized with a high variance, it is very likely that it also has a high variability, since it probably doesn't map the true functional. The mean result of a very large number of BB mappings, however, must give the true functional, i.e. $\overline{f_{BB}(\boldsymbol{u}^*)} \equiv g(\boldsymbol{u}^*)$. A model can, therefore, have a high variance, while its mean outcome has zero variability.

Making the difference between variability and variance is rather important when discussing generalization. The goal of a good generalization should be to achieve a low variability, low variance, together with low bias. As will be seen further, many researchers content themselves whenever a low variability with a low bias is reached.

II. IMPROVING GENERALIZATION WITH JITTER, THE BIAS/VARIABILITY TRADE-OFF [1]

Within the NN domain a seemingly simple to use technique exists that is generally accepted to improve the generalization behaviour of a NN model. In this technique *jitter* (call it noise) is added to the existing measurements, thus increasing the existing data set with added noisy data samples. The NN model parameters are then optimized using this enlarged

1. In fact, all references mentioned, speak about the bias/variance trade-off. What is really meant in these references, however, should be described as the bias/variability trade-off, given the definitions 2 through 5. It is not possible to swap the definitions of variance and variability here, since the definition 3 is consistent with the generally accepted definition of the variance.

data set.
Bishop [2] states that enlarging training sets by adding noisy data, allows for more complex models which then reduce bias. For his statement, Bishop uses the knowledge that one of the causes for biasing is a too low number of model parameters used. Adding noise, however, introduces bias at the level of the cost function, as is shown further in this paper. This bias cannot be overcome by picking a more complex BB model, or by increasing the number of parameters!

No source can be found where it is proven that adding output noise indeed helps generalization. On the other hand, different sources have studied the effects of adding input noise to NN models (An [1], Bishop [2], Geman et al. [4], Haykin [5], Mackay [8], Reed et al. [9] and Twomey et al. [12]). It is stated that input noise can be used, as long as it is balanced with the model variability. Sometimes the use of noise with low, but nonzero variances is recommended (e.g. [3] and [8] for NN models), while other sources stress the importance of using large enough variances such that generalization is guaranteed. Within the NN domain, this is called the *bias/variance* trade-off. Because of the unpredictable results of adding noise, it is generally accepted that the level of noise must be designed, e.g. by putting constraints in the BB architecture [5] which makes the simple method of adding noise suddenly more complex. Yet, this complexity cannot be avoided and the variance of the noise must be chosen carefully [2].
Geman [4] reports an example in which adding noise results in a nonlinear mapping that comes very close to the true functional in at least one case. First, this implies that the true functional is known, which is usually not the case. Second, the model variance is completely left out of the picture in this discussion, as only one realization was taken as an example. It is, therefore, always necessary to analyse the mean BB mapping $\overline{f_{BB}}$.
To clarify the idea that is brought in this paper, we will only deal with the simple SISO case and show why adding noise leads to errors in the model performance. The analysis is based on a higher order decomposition of the cost function. It will be shown that adding noise increases the probability that higher order derivatives of a BB model are suppressed.

III. THE EFFECTS OF ADDING NOISE ON MEASUREMENT DATA

Consider a SISO system that is measured N times with measured input samples $u_k \in \mathbb{R}$ and output samples $y_k \in \mathbb{R}$ where $k = 1, 2, ..., N$. The input-output pairs (u_k, y_k) are used to optimize the parameters of a BB or NN model

$$f_{BB}: \mathbb{R} \to \mathbb{R} : y_k = f_{BB}(u_k, \theta) \tag{5}$$

Assumption 1 The parameters θ of the nonlinear model are optimized by minimizing the LS cost function

$$C_{LS} = \frac{1}{N}\sum_{k=1}^{N}(y_k - f_{BB}(u_k, \theta))^2 \quad . \tag{6}$$

This assumption means that no specific care is taken to minimize the effects of noisy measurements, as can be done using a WLS or EIV cost function or added regularization terms. □

Assumption 2 In the case that enough noiseless measurements are available and that the measurements cover the whole domain of interest, it is possible to find the true BB parameters $\theta^* = \text{argmin}_\theta(C_{LS})$. The idea behind this assumption is that the BB model is expected to be a universal approximator. □

Now consider the case where the experimenter decides to add noise as a means of generalization. In the most general case the addition of noise is done by taking the measurements R times and add noise on the resulting measurement set of size $R \times N$, or

$$u_k^{[r]} = u_k + n_{k,u}^{[r]} \quad \text{and} \quad y_k^{[r]} = y_k + n_{k,y}^{[r]} \text{ , such that } . C_{LS} = \frac{1}{N}\sum_{k=1}^{N}\left\{\frac{1}{R}\sum_{r=1}^{R}\left[\left(f_{BB}(u_k^{[r]}, \theta) - y_k^{[r]}\right)^2\right]\right\} \tag{7}$$

with $n_{k,u}^{[r]} \in \mathbb{R}$ and $n_{k,y}^{[r]} \in \mathbb{R}$ random values, and $r = 1, 2, ..., R$. The former LS cost function (6) now also sums the added noisy samples.

Assumption 3 $n_{k,u}^{[r]}$ and $n_{k,y}^{[r]}$ are independent, mutually uncorrelated, zero mean Gaussian distributed random variables with known standard deviations $\sigma_{u,k}, \sigma_{y,k}$ and such that $\sigma_{uy,k} = 0; \forall k$. □

Assumption 4 It is assumed that a very large number of noisy samples is taken, or $R \to \infty$. This assumption is used for the infinite case of the following theorems. Further is shown that the theorems also apply in the finite case. □

Lemma 1 Under assumptions 3 and 4 the LS cost function (7) converges strongly to its expected value, or:

$$C_{LS} = \underset{R \to \infty}{\text{a.s.Lim}}(E\{C_{LS}\}) \tag{8}$$

Proof: see Lukacs [7]. □

Hence to study the influence of the added noise samples for $R \to \infty$ it is sufficient to study $E\{C_{LS}\}$ which is more tractable than (7). Since the added noisy samples on inputs and outputs are mutually uncorrelated, the two cases are treated differently in the next paragraphs. Further is shown that this approach is justified.

Theorem 1 Under assumptions 1, 3 and 4 the addition of noisy output samples for training BB models, has no effect on finding the BB parameters θ^* that minimize the LS cost function, i.e. adding output noise is useless.

Proof: see section VII.: Appendix. □

Theorem 2 Under assumptions 1, 3 and 4 the addition of noisy input samples leads to a suppression of higher order derivatives of the BB function. This suppression is proportional to the variance of the added noise, due to a regulizer term in the cost function, such that

$$E\{C_{LS}\} = \frac{1}{N}\sum_{k=1}^{N}\left[\left(f_{BB}(u_k, \theta) - y_k\right)^2 + \varepsilon_k^2(\sigma_u)\right] \tag{9}$$

with $\varepsilon_k^2(\sigma_u)$ a regulizer term that is only dependent on the higher order derivatives of $\boldsymbol{f}_{BB}$ and the noise variance.

Proof: see section VII.: Appendix. □

Based on this theorem we can conclude that the noise puts a constraint on the higher order derivatives of the BB model. The regulizer term ε_k^2 increases when σ_u increases. In effect, adding noise makes the resulting model smoother and influences directly the variability of the model, but it must be well understood that no conclusion can be made concerning the variance. The variance of the model can still be very large.

The noise, however, also introduces bias. The ε_k^2 regulizer is not θ independent and it is highly unlikely that both the regulizer and the original function have the same minimum for θ. As a result the regulizer pulls the BB parameters away from the ideal solution θ^*. This effect also increases with the amplitude of the added noise.

Theorem 3 Under assumptions 1, 3 and 4 the addition of noisy input samples and noisy output samples, has the same effect as adding noisy input samples only.

Proof: The proof is not given in detail in this paper, but using the same lines as the proof of Theorem 1, it can be shown that the case of Theorem 3 folds back to the case of Theorem 2. □

IV. THE FINITE SAMPLE CASE

The above theorems are given for the case that an infinite number of noisy measurements is added. This is not very practical and the remaining question is how many noisy samples are needed in order to use the above theorems.

Assume therefore that the number of repeated and noisy measurements R is a finite value. The cost function (7) is still used, but as R is finite, only a sample mean version of the cost can be calculated.

Definition 6: The sample mean Least Squares cost function $\hat{C}_{LS}$ is defined as

$$\hat{C}_{LS} = \frac{1}{N}\sum_{k=1}^{N}\left(f_{BB}(\hat{u}_k, \theta) - \hat{y}_k\right)^2 \tag{10}$$

in which the sample means $\hat{u}_k$ and $\hat{y}_k$ are calculated as $\hat{u}_k = \frac{1}{R}\sum_{r=1}^{R} u_k^{[r]}$ and $\hat{y}_k = \frac{1}{R}\sum_{r=1}^{R} y_k^{[r]}$. □

Lemma 2 Under Assumption 3 and for $R \geq 6$ it can be proven that

$$E\{C_{LS}\} = \frac{R-3}{R-1}E\{\hat{C}_{LS}\} \tag{11}$$

Proof: See Schoukens et al. [10]. □

The $(R-3)/(R-1)$ term is independent of the NN parameters θ and both sides of equation (11) have the same minimizer for θ. Conclusions that were drawn for the infinite case, therefore also apply for the finite case if more than five noisy samples were added. This is usually the case: most references assume the addition of noisy samples that are 100 to 1,000 times larger in size than the original amount of data samples.

V. CONCLUSION

In this paper was shown that adding noisy input samples while modelling a Black Box model in the general, or a Neural Network in the particular case, is a bad idea. To clarify the idea that is brought, this paper defined the concept of variability of a model. It was shown that adding input noise indeed decreases the variability of a model. The resulting model has a smoother transfer function and the feeling exists that the model is more accurate. This is not the case, as the variance of the model still exists, while bias errors are introduced. It is our observation that the decrease in variability does not outweigh the increase in bias, while the variance remains the same. We suggest the use of interpolation techniques instead of adding input noise (see also references [12] and [14]).

VI. ACKNOWLEDGMENTS

This paper presents research results of the Belgian Programme on Interuniversity Poles of Attraction (IUAP 4/2), initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture; and the Flemish Government (GOA, IMMI2). The scientific responsibility rests with its authors.

VII. APPENDIX: PROOFS

A. Proof of Theorem 1: effects of output noise on BB modelling

Assume only noise on the output samples. Using Lemma 1 the expected value of the LS cost function (7) becomes

$$E\{C_{LS}\} = E\left\{\frac{1}{N}\sum_{k=1}^{N}\frac{1}{R}\sum_{r=1}^{R}\left(f_{BB}(u_k, \theta) - y_k^{[r]}\right)^2\right\}. \tag{12}$$

Put (7) in (12). Since

$$E\left\{\frac{1}{R}\sum_{r=1}^{R}(n_{k,y}^{[r]})^2\right\} = \sigma_{k,y}^2 \quad \text{and} \quad E\left\{\frac{-2}{R}\sum_{r=1}^{R}n_{k,y}^{[r]}\left(f_{BB}(u_k, \theta) - y_k\right)\right\} = 0\ , \tag{13}$$

the expected value (12) becomes

$$E\{C_{LS}\} = E\left\{\frac{1}{N}\sum_{k=1}^{N}\left[\left(f_{BB}(u_k, \theta) - y_k\right)^2 + \sigma_{k,y}^2\right]\right\}. \tag{14}$$

The $\sigma_{k,y}^2$ term in this equation is independent of the BB parameters θ. The minimizer of (14) for θ is therefore exactly the same as the minimizer of (6) such that adding noise to the output samples doesn't change the BB parameters, i.e. has no regularization effect. □

B. Proof of Theorem 2: effects of input noise on BB modelling

Assume only noise on the input samples. Using Lemma 1, the expected value of the LS cost function (7) becomes

$$E\{C_{LS}\} = E\left\{\frac{1}{N}\sum_{k=1}^{N}\frac{1}{R}\sum_{r=1}^{R}\left(f_{BB}(u_k^{[r]}, \theta) - y_k\right)^2\right\}. \tag{15}$$

Put (7) in (15). The BB model is assumed continuous with finite higher order derivatives such that we can replace the BB model by its Taylor expansion

$$f_{BB}(u_k^{[r]}, \theta) = f_{BB}(u_k + n_{k,u}^{[r]}, \theta) = f_{BB}(u_k, \theta) + \sum_{l=1}^{\infty}\frac{(n_{k,u}^{[r]})^l}{(l)!}\frac{\partial^l f_{BB}(u_k, \theta)}{(\partial u_k)^l} \tag{16}$$

and put this expansion into the cost function. The calculation of the expectation value becomes

$$E\{C_{LS}\} = E\left\{\frac{1}{N}\sum_{k=1}^{N}\frac{1}{R}\sum_{r=1}^{R}\left[f_{BB}(u_k, \theta) + \sum_{l=1}^{\infty}\frac{(n_{k,u}^{[r]})^l}{(l)!}\frac{\partial^l f_{BB}(u_k, \theta)}{(\partial u_k)^l} - y_k\right]\right\}. \tag{17}$$

After computing the summation terms and knowing that (Stuart et al. [11])

$$E\{(n_{k,u}^{[r]})^{2l+1}\} = 0 \quad \text{and} \quad \frac{1}{\sigma_{k,u}\sqrt{2\pi}}\int_{-\infty}^{+\infty}(n_{k,u}^{[r]})^{2l}\exp\left(-\frac{(n_{k,u}^{[r]})^2}{2\sigma_{k,u}}\right)d(n_{k,u}^{[r]}) = \frac{\sigma_{k,u}^{2l}}{2^l}\frac{(2l)!}{l!}\ , \tag{18}$$

it is possible to write (17) as

$$E\{C_{LS}\} = \frac{1}{N}\sum_{k=1}^{N}\left(f_{BB}(u_k, \theta) - y_k\right)^2 + \frac{1}{N}\sum_{k=1}^{N}\left(2\mu_k\left(f_{BB}(u_k, \theta) - y_k\right) + \varepsilon_k^2\right) \tag{19}$$

with

$$\mu_k = \sum_{l=1}^{\infty} \frac{\sigma_{k,u}^{2l}}{2^l (l)!} \frac{\partial^{2l} f_{BB}(u_k, \theta)}{(\partial u_k)^{2l}} \qquad \text{and} \qquad \varepsilon_k^2 = E\left\{\sum_{l=1}^{\infty} \frac{(n_{k,u}^{[r]})^l}{(l)!} \frac{\partial^l f_{BB}(u_k, \theta)}{(\partial u_k)^l}\right\} \tag{20}$$

where ε_k^2 is the expectation of the square term which can be decomposed as

$$\varepsilon_k^2 = \sum_{l=1}^{\infty}\left[\sigma_{k,u}^{2l}\frac{(2l)!}{2^l(l!)^3}\left(\frac{\partial^l f_{BB}(u_k, \theta)}{(\partial u_k)^l}\right)^2\right] + 2\sum_{l=2}^{\infty}\left[\sum_{t=1}^{l-1}\sigma_{k,u}^{2l}\frac{(2l)!}{(l!)2^l(2t-l)!}\frac{\partial^t f_{BB}(u_k, \theta)}{(\partial u_k)^t}\frac{\partial^{2l-t} f_{BB}(u_k, \theta)}{(\partial u_k)^{2l-t}}\right]. \tag{21}$$

The term $(f_{BB}(u_k, \theta) - y_k)$ in equation (19) contains values that are spread around zero and is of the order $\vartheta(\sqrt{K})$, while the ε_k^2 term is of order $\vartheta(K)$. For large K it is possible to skip the terms in μ_k. This was also concluded for SISO Neural Networks by An [1] and Bishop [2] for the case $\rightarrow \infty$. In the analysis given in this paper, K is assumed large enough, so that the analysis is restricted to

$$E\{C_{LS}\} \cong \frac{1}{N}\sum_{k=1}^{N}\left[\left(f_{BB}(u_k, \theta) - y_k\right)^2 + \varepsilon_k^2\right] \quad . \tag{22}$$

A closer view on (21) shows that ε_k^2 is a summation of variance-derivative products, such that it acts as a regulizer term that adds higher order derivatives of the BB model into the cost function. ε_k^2 has become a regulizer term that is dependent on the noise variance in a very complex way. □

VIII. REFERENCES

[1] An G., "The Effects of Adding Noise During Backpropagation Training on a Generalization Performance", Neural Computation, Vol. 8, 1996, pp. 643 - 674.

[2] Bishop C. M., "Neural Networks for Pattern Recognition", Clarendon Press, Oxford, 1995.

[3] Fausett L., "Fundamentals of Neural Networks, Architectures, Algorithms and Applications", Prentice Hall, 1994.

[4] Geman S., Bienenstock E., Doursat R., "Neural Networks and the Bias/Variance Dilemmma", Neural Computation, Vol. 4, 1992, pp. 1 - 58.

[5] Haykin S, "Neural Networks, a Comprehensive Foundation", Prentice Hall, 1999.

[6] Juditsky A., Hjalmarsson H., Benveniste A., Delyon B., Ljung L., Sjöberg J., Zhang Q., "Nonlinear Black-box Models in System Identification: Mathematical Foundations", Automatica, Vol. 31, No. 12, 1995, pp. 1725 - 1750.

[7] Lukacs E., "Stochastic Convergence", Academic Press, New York, 1975.

[8] MacKay D. J. C., "Bayesian Methods for Adaptive Models", PhD thesis, California Institute of Technology, Pasadena, California, 1992.

[9] Reed R., Marks R. J. II, Oh S., "Similarities of Error Regularization, Sigmoid Gain Scaling, Target Smoothing, and Training with Jitter", IEEE Transactions on Neural Networks, Vol. 6, No. 3, May 1995, pp. 529 - 538.

[10] Schoukens J., Pintelon R., Rolain Y., "Maximum Likelihood Estimation of Errors-In-Variables Models Using a Sample Covariance Matrix Obtained from Small Data Sets", Recent Advances in Total Least Squares Techniques and Errors-In-Variables Modeling, 1997, pp. 59 -68.

[11] Stuart A., Ord J. K., "Kendall's Advanced Theory of Statistics, Distribution Theory", Vol. 1, Charles Griffin & Co., 1987.

[12] Twomey J. M., Smith A. E., "Bias and Variance of Validation Methods for Function Approximation Neural Networks Under Conditions of Sparse Data", IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews, Vol. 28, No. 3, August 1998, pp. 417 - 430.

[13] Van Gorp J., Rolain Y., "An Interpolation Technique for Learning With Sparse Data", accepted for SYSID 2000, Santa Barbara, California, June 2000, see also http://elecftp.vub.ac.be/Papers/Jurgen.Van.Gorp.html.

[14] Van Gorp J., Schoukens J., Pintelon R., "Adding Input Noise to Increase the Generalization of Neural Networks is a Bad Idea", ANNIE 1998, Intelligent Engineering Systems Through Artificial Neural Networks, Volume 8, pp. 127 - 132.