Generalization: The Hidden Agenda of Learning*

by Jan Larsen and Lars Kai Hansen
Department of Mathematical Modelling
Technical University of Denmark
emails: jl,lkhansen@imm.dtu.dk

Most neural systems are adapted by optimization of a performance index, typically the minimization of a "cost function", based on a finite database (a training set) of N noisy examples derived from the target system. However, there is always the *hidden agenda* that the model should perform well, not only on the training set, but on the much larger set of future inputs to the system.

Reading for your finals you solve previous years tests, but you know very well that if you then test yourself on last years test the result will be biased – too optimistic! Only a test on a fresh data set, a test that was put aside before you started reading, will give the you a reliable prediction of the final performance.

Doing well on unseen data may at first seem unattainable, but the ability to generalize in very complex environments is nevertheless one of the most striking properties of neural systems, and indeed one of the reasons that neural networks have shown useful in practical applications.

As an example: in [10] a neural network system for inspection of handwritten digits was able to classify 99.98% correct after training on a data base of 7291 digits, and classify 95% correct on an additional test set of 2007 digits.

When using a super-flexible model family, like neural networks, which in principle can model arbitrarily complex systems, *overfit* is a major concern, which finds expression in the ubiquitous bias-variance dilemma [4]. The generalization ability of an adaptive system is the quantitative measure of performance on a hypothetical infinite test set. While this quantity cannot be accessed directly, algebraic asymptotic estimates of generalization, valid for large training sets ($N \to \infty$), can be derived [1], [2], [9], [12], [13], [14]. Such asymptotic results were earlier derived for supervised learning; however, it was recently shown that generalization ability for unsupervised learning machines (e.g., principal component analysis and clustering schemes) can be analyzed in a similar framework [7].

If sufficient computational capacity is available, empirical resampling schemes can be invoked. The two basic resampling strategies are cross-validation and bootstrap. Cross-validation [3], [15] is based on a random division of the database into disjunct training and validation sets. The procedure can be repeated, leading to more accurate results at the price of increased computation. The so-called leave-one-out cross-validation is based on using only a single example in the test set, and typically resampling N times. Approximative techniques, by which the computational overhead in leave-one-out is significantly reduced, has been reported [8], [13].

Bootstrap, invented by Efron [6], is based on resampling with replacement. Bootstrap produces pseudo training sets of size N, hence, simulates training set fluctuations at the full sample size, and was applied to control of overfit in a number of investigations [5], [17], [18].

Optimization of the neural network architecture may lead to better generalization ability and preferably lower computational burden. optimizing the network architecture is

*Appears in J.-N. Hwang, S.Y. Kung, M. Niranjan & J.C. Principe (eds.) The Past, Present, and Future of Neural Networks for Signal Processing, *IEEE Signal Processing Magazine*, pp. 43–45, Nov. 1997.

to optimally trade off bias and variance [4], hence, maximizing generalization ability. This can be done *directly* by optimizing the structure of the network by pruning or growing techniques or *indirectly* by using regularization. Regularization – which goes back to Hadamard – consist in adding a penalty term to the cost function. As an example consider predicting the sunspot time series shown in the upper panel of Figure 1. The lower panel [16] shows that generalization error (test error) is reduced by pruning the network.

References

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

[2] P. Craven & G. Wahba: "Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation," *Numerical Mathematics*, vol. 31, 377–403, 1979.

[3] S. Geisser: "The Predictive Sample Reuse Method with Application," *Journal of The American Statistical Association*, vol. 50, pp. 320–328, 1975.

[4] S. Geman, E. Bienenstock & R. Doursat: Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, vol. 4, pp. 1–59, 1992.

[5] B. Efron & R. Tibshirani: "Cross-Validation and the Bootstrap: Estimating the Error Rate of a Prediction Rule," Techn. Report no. 477, Dept. of Statistics, Stanford University, May 1995. To appear in Journ. Amer. Statist.

[6] B. Efron & R. Tibshirani: *An Introduction to the Bootstrap*, New York, NY: Chapman & Hall, 1993.

[7] L.K. Hansen & J. Larsen: "Unsupervised Learning and Generalization," in *Proceedings of IEEE International Conference on Neural Networks*, Washington DC, vol. 1, pp. 25–30, June 1996.

[8] L.K. Hansen & J. Larsen: "Linear Unlearning for Cross-Validation," *Advances in Computational Mathematics*, vol. 5, pp. 269–280, 1996.

[9] 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.

[10] Y. Le Cun *et al.*: "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[11] Y. Le Cun, J.S. Denker & S.A. Solla: Optimal Brain Damage. In D.S. Touretzky (ed.), Advances in Neural Information Processing Systems 2, Proceedings of the 1989 Conference, San Mateo, California: Morgan Kaufmann Publishers pp. 598–605, 1990.

[12] 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, pp. 1–10, 1991.

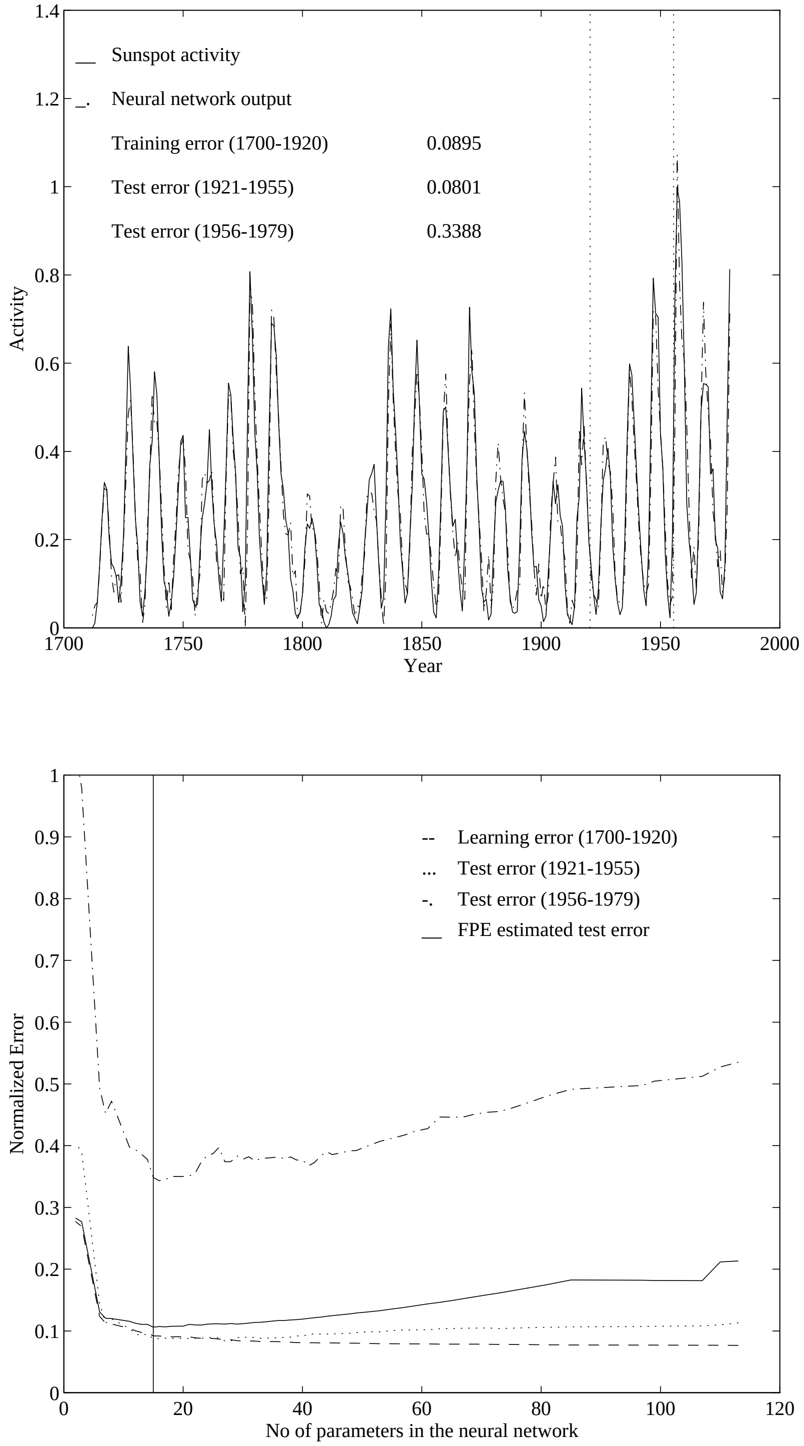

Figure 1: Upper panel: prediction the sunspot time series using an optimally pruned feed-forward neural network. Lower panel: evolution of training and test error during a pruning session using Optimal Brain Damage [11]. FPE is a modified version of the Final Prediction Error estimate [1]. The vertical line indicate the optimal network for which the FPE estimate is minimal.

[13] J. Moody: "Prediction Risk and Architecture Selection for Neural Networks" in V. Cherkassky, J. H. Friedman & H. Wechsler (eds.) *From Statistics to Neural Networks: Theory and Pattern Recognition Applications*, Series F, vol. 136, Berlin, Germany: Springer-Verlag, 1994.

[14] N. Murata, S. Yoshizawaand & S. Amari, "Network Information Criterion — Determining the Number of Hidden Units for an Artificial Neural Network Model," *IEEE Transactions on Neural Networks*, vol. 5, no. 6, pp. 865–872, Nov. 1994.

[15] M. Stone, "Cross-validatory Choice and Assessment of Statistical Predictors," *Journal of the Royal Statistical Society B*, vol. 36, no. 2, pp. 111–147, 1974.

[16] C. Svarer, L.K. Hansen & J. Larsen: "On Design and Evaluation of Tapped-Delay Neural Architectures," in *Proceedings of the IEEE International Conference on Neural Networks*, San Francisco, California, USA, vol. 1, pp. 46–51, 1993.

[17] R. Tibshirani: "A Comparison of Some Error Estimates for Neural Network Models," *Neural Computation*, vol. 8, pp. 152–163, 1996.

[18] A.S. Weigend & B. LeBaron: "Evaluating Neural Network Predictors by Bootstrapping," in *Proceedings of the International Conference on Neural Information Processing (ICONIP'94)*, Seoul, Korea, pp.

1207–1212, 1994.