Stochastic Linear Learning: Exact Test and Training Error Averages*

Lars Kai Hansen†‡
Technical University of Denmark

Running Title: Stochastic Linear Learning

* *Neural Networks*, in press

†Stimulating discussions with Jan Larsen are gratefully acknowledged. I thank Peter Salamon and the participants of the CONNECT *Journal Club* for useful comments on the manuscript. This research is supported by the Danish Research Counsils for The Natural and Technical Sciences through the Danish Computational Neural Network Center.

‡Request for reprints should be sent to Lars Kai Hansen, CONNECT, Electronics Institute, B.349, Technical University of Denmark, DK-2800 Lyngby, Denmark. Phone: +45 45 93 42 07. Fax: +45 42 88 01 17. Email:lars@eiffel.ei.dth.dk

Stochastic Linear Learning:

Exact Test and Training Error Averages

Abstract

The statistical properties of stochastic linear learning are studied. Exact averages of the test and training errors are derived for a linear model and a finite training set. Our results extend Akaike's asymptotic estimate of the Final Prediction Error in two directions: the randomness arriving from the learning process is included and the results are valid for finite training sets.

Keywords:

Stochastic linear learning, test error estimate, Akaike's Final Prediction Error (FPE) estimate, post training distribution, system identification, inverted Wishart distribution, unbiased noise estimate.

List of Symbols

α :	Number of training set examples pr. model parameter ($\alpha \equiv p/N$)
$A_{jj'}$:	Second moment matrix of the training set input distribution
$\overline{A^{-1}_{jj'}}$:	Average of the inverse of $A_{jj'}$ over the distribution of training sets
β^{-1} :	Effective temperature of the post training distribution
$E^{\alpha}(w)$:	Square of error on example α, when using model with weights w
ϵ_T :	Average over the post training ensemble of training errors for fixed training set
ϵ_G :	Average over the post training ensemble of test errors for fixed training set
$\overline{\epsilon_T}$:	Average of ϵ_T over the distribution of training sets
$\overline{\epsilon_G}$:	Average of ϵ_G over the distribution of training sets
η :	Gradient descent parameter
N :	Number of model parameters
$N(0,\Sigma)$:	The zero mean multivariate normal distribution with covariance matrix Σ
$N(0,\sigma_\nu^2)$:	The zero mean univariate normal distribution with variance σ_ν^2
ν^{α} :	Addititive noise contribution of α'th example
p :	Number of training examples
$P_p(w)$:	Post training distribution of weights
$P_{\sigma_\nu^2}(\nu)$:	Density of the zero mean univariate normal distribution with variance σ_ν^2
$P_{\Sigma}(x)$:	Density of the zero mean multivariate normal distribution with covariance matrix Σ
σ_ν^2 :	Variance of noise distribution
Σ :	Covariance matrix of input vector distribution
w_j :	j'th component of the model parameter vector (weights)
x_j :	j'th component of input vector
y :	Output of the linear model
$Z_p(h)$:	Generating functional for post training ensemble averages

1 Introduction

Test error estimates play an important role in model selection and system identification (Ljung, 1984). While crossvalidation is the ultimate tool for validating a statistical model (Toussaint, 1974), it may be impractical in applications. Often applications do not provide data for thorough testing. Furthermore, crossvalidation is time consuming, eg. during architecture optimization where many models need to be evaluated. Under such circumstances *estimation* of test errors may be invoked. Such estimates are typically based on measurements of the training set error, combined with statistical assumptions about the domain. The standard results on test error estimation are due to Akaike (Akaike, 1969, 1974) and comprise asymptotic expressions valid for the large training sets and essentially linear models (Ljung, 1984).

Much recent attention is devoted to the study of simple linear learning schemes (Krogh, Hertz & Thorbergsson, 1989; Levin, Tishby & Solla, 1990; Le Cun, Kanter & Solla, 1991; Krogh & Hertz, 1992). The simple learning schemes can be analyzed analytically and it is projected that the results may either be applied directly to the more interesting non-linear models, such as feed forward neural networks with hidden representations, or may inspire future analysis of such models. Recent succesful schemes for optimization of neural network architecture are derived basically from local linear approximations (Le Cun, Denker & Solla, 1989). Fogel used Akaike's Information Criterion for comparing neural network architectures (Fogel, 1991). Hoffman and Larsen used Akaike's Final Prediction Error estimate (FPE) for optimal reduction of polynomial models (Hoffmann & Larsen, 1991).

Krogh *et al.* studied the learning process in the limit of large models and large training sets. They found *critical slowing down* of the learning process, reminiscent of the slowing down of relaxation phenomena observed close to *second order phase transitions* in statistical physics, showing that linear learning problems can behave non-trivially (Krogh *et al.*, 1989). We find in this work the exact form of the divergence of the test error close to the critical point.

In this presentation we analyse linear models emphasizing the case of finite training sets. Below we formulate the model, while section 3 defines the post training distribution of parameters following a stochastic learning procedure. In section 4 we compute analytically the estimates for test and training errors, and in section 5 we compare our result with Akaike's asymptotics for the deterministic limit.

2 Linear Modeling

Consider a linear model with N inputs x_j, $j = 1, 2, ..., N$, and a single real-valued output $y(x)$:

$$y(x) = \sum_{j=1}^{N} w_j x_j \tag{1}$$

The N model parameters w_j will be estimated by the standard recursive gradient descent procedure.

The database or *training set* that we will use for estimating the optimal parameters is created by the an unreliable teacher having a set of weights w_j^* (Krogh *et al.*, 1989):

$$y^\alpha = \sum_{j=1}^{N} w_j^* x_j^\alpha + \nu^\alpha \qquad \alpha = 1, 2, ..., p \tag{2}$$

We consider a *finite* training set of p examples. The inputs to the teacher and the noise are independent normally distributed: $x \in N(0, \Sigma)$, $\nu \in N(0, \sigma_\nu^2)$. While we can handle any non-singular correlation matrix $\Sigma = (\Sigma_{jj'})$, it is crucial to our arguments that the noise and the input are independent random variables.

3 The Post Training Distribution

In order to model the generic behavior of a network following a stochastic learning procedure we will compute average properties of an ensemble of networks. The distribution function of the ensemble reflects the learning procedure. Properties of *deterministic* learning procedures can be optained as a limit of the general results.

Levin, Tishby, and Solla have presented general arguments in favor of a *Gibbs distribution* of weights (Levin *et al.*, 1990):

$$P_p(w) = Z_p^{-1} \exp\left(-\beta \sum_{\alpha=1}^{p} E^\alpha(w)\right), \tag{3}$$

where the error on the $\alpha'th$ example is given by:

$$E^{\alpha}(w) = (y(x^{\alpha}) - y^{\alpha})^2 = \left(\sum_{j=1}^{N}(w_j - w_j^*)x_j^{\alpha} - \nu^{\alpha}\right)^2 . \tag{4}$$

and Z_p is the normalization integral.

In many applications learning is achieved by the standard recursive gradient descent method with step-size parameter η:

$$\delta w_j^n \equiv w_j^{n+1} - w_j^n = -\eta \frac{\partial E^{\alpha(n)}}{\partial w_j}, \tag{5}$$

ie. weights are updated after each presented example. The presentations form a random sequence $(\alpha(n))$ drawn from the training set.

It has recently been shown that the distribution of weights, following a stochastic learning procedure as above, solves a Fokker-Planck equation (Radons, 1990; Hansen, Pathria, & Salamon 1991). In the particular case where the covariance matrix of the input vector, is given by $\Sigma_{jj'} = \sigma_x^2 \delta_{jj'}$, it can be shown that the stationary distribution, for small η, can be approximated by a Gibbs distribution as above[1]. In the following we will *assume* the post training distribution to be Gibbsian.

4 Average Test and Training Errors

Given a specific database and using the post training distribution we can compute the *ensemble average* of the training error:

$$\epsilon_T = \int D\tilde{w} P_p(\tilde{w}) \frac{1}{p} \sum_{\alpha=1}^{p} E^{\alpha}(\tilde{w}) \tag{7}$$

with $\tilde{w} \equiv w - w^*$, and using the notation: $\int D\tilde{w} \equiv \prod_{j=1}^{N} \int d\tilde{w}_j$.

The test or *generalization* error is computed as the joint average over the post training ensemble and a random example drawn from the same distribution as the

[1] In fact the effective *temperature* β^{-1} is given by:

$$\beta^{-1} = \eta p^{-1} \sum_{\alpha=1}^{p} (F^{\alpha} - F_0)^2, \tag{6}$$

with $F^{\alpha} = -\partial E^{\alpha}/\partial w$, $F_0 = p^{-1}\sum_{\alpha=1}^{p} F^{\alpha}$. If we assume the *scalar* correlation structure this leads to $\beta^{-1} = 4\eta\sigma_\nu^2\sigma_x^2$ (Hansen *et al.*, 1991).

training set:

$$\epsilon_G = \int D\tilde{w} P_p(\tilde{w}) \int Dx \int d\nu P_\Sigma(x) P_{\sigma_\nu^2}(\nu) \left(\sum_{j=1}^{N} \tilde{w}_j x_j - \nu \right)^2 \tag{8}$$

$P_\Sigma(x)$ and $P_{\sigma_\nu^2}(\nu)$ denote the appropriate normal densities of input and noise. Averaging over the random example this simplifies to:

$$\epsilon_G = \int D\tilde{w} P_p(\tilde{w}) \left(\sum_{j,j'=1}^{N} \Sigma_{jj'} \tilde{w}_j \tilde{w}_{j'} + \sigma_\nu^2 \right) . \tag{9}$$

Both ϵ_T and ϵ_G may be derived from the generating functional:

$$Z_p(h) = \int D\tilde{w} \ \exp \left(-\beta \left(\sum_{\alpha=1}^{p} E^\alpha(\tilde{w}) + \sum_{j=1}^{N} h_j \tilde{w}_j \right) \right) , \tag{10}$$

by computing the derivatives:

$$\epsilon_T = -\frac{1}{p} \frac{\partial}{\partial \beta} [\log Z_p(h)]_{h=0} \tag{11}$$

and

$$\epsilon_G = \frac{1}{\beta^2} \sum_{j,j'=1}^{N} \Sigma_{jj'} \left[\frac{\partial^2}{\partial h_j \partial h_{j'}} \log Z_p(h) + (\frac{\partial}{\partial h_j} \log Z_p(h))(\frac{\partial}{\partial h_{j'}} \log Z_p(h)) \right]_{h=0} + \sigma_\nu^2 \tag{12}$$

Evaluating the gaussian integral in the generating functional we find:

$$\log Z_p(h) = -\frac{1}{2} \log \left(\det(\beta A) \pi^{-N} \right) + \frac{\beta}{4} \sum_{j,j'=1}^{N} a_j a_{j'} (A^{-1})_{jj'} + \beta a_0, \tag{13}$$

where we have defined:

$$A_{jj'} \ = \ \sum_{\alpha=1}^{p} x_j^\alpha x_{j'}^\alpha \tag{14}$$

$$a_j = h_j + 2\sum_{\alpha=1}^{p} x_j^\alpha \nu^\alpha \quad (15)$$

$$a_0 = \sum_{\alpha=1}^{p} (\nu^\alpha)^2 \quad (16)$$

Leading to the following expressions for the training and test errors for the given training set:

$$\epsilon_T = -\frac{1}{p}[-\frac{N}{2\beta} + \sum_{jj'}\sum_{\alpha\alpha'} x_j^\alpha x_j^{\alpha'} \nu^\alpha \nu^{\alpha'} (A^{-1})_{jj'} - \sum_\alpha (\nu^\alpha)^2] \quad (17)$$

and

$$\epsilon_G = \sum_{k'k''} \Sigma_{k'k''}[\frac{1}{2\beta}(A^{-1})_{k'k''} + \sum_{jj'j''}\sum_{\alpha'\alpha''} x_{j'}^{\alpha'} x_{j''}^{\alpha''} \nu^{\alpha'} \nu^{\alpha''} (A^{-1})_{k'j'} (A^{-1})_{k''j''}] + \sigma_\nu^2 \quad (18)$$

These results are valid for the specific training set. We are, as noted above, interested in the generic behavior, hence we average over all possible training sets. We use that $\overline{\nu^\alpha \nu^{\alpha'}} = \sigma_\nu^2 \delta_{\alpha\alpha'}$, to obtain the training error:

$$\overline{\epsilon_T} = \frac{N}{2p\beta} + (1 - \frac{N}{p})\sigma_\nu^2 \quad (19)$$

It should be noted that this exact average was computed without any reference to the distribution of the inputs, except the claim that the A matrix should be non-singular. In this case the gaussian integrals entering the generating functional are well-defined[2]. This means that the expression (19) can be used as an unbiased estimate of the noise level (σ_ν), under rather general circumstances.

The average generalization error becomes:

$$\overline{\epsilon_G} = [\frac{1}{2\beta} + \sigma_\nu^2] \sum_{k'k''} \Sigma_{k'k''}(\overline{(A^{-1})_{k'k''}}) + \sigma_\nu^2 \quad (20)$$

The inverse covariance matrix of the input vectors, $(A^{-1})_{jj'}$, samples an *inverted Wishart* distribution with p degrees of freedom, $W^{-1}(\Sigma^{-1}, p)$, for which the mean is given by (Anderson, 1984):

[2] To be more precise our condition is that the A matrix should be non-singular with probability one.

$$\overline{(A^{-1})_{jj}} = \frac{1}{(p-N-1)}(\Sigma^{-1})_{jj'} \quad (p > N+1) \tag{21}$$

By inserting this we finally obtain:

$$\overline{\epsilon_G} = \frac{N}{2\beta(p-N-1)} + \frac{p-1}{p-N-1}\sigma_\nu^2 \tag{22}$$

It is worth noting that this is an exact finite temperature average of the generalization error, and it is valid for arbitrary correlations among the components of the multi-normal input vector as long as the correlation matrix Σ is non-singular.

5 Discussion and comparison with Akaike's FPE

Expressions (19) and (21) may be combined in a relation between test and training errors:

$$\overline{\epsilon_G} = \frac{p(p-1)}{(p-N-1)(p-N)}\overline{\epsilon_T} - \frac{N-1}{(p-N-1)(p-N)}\frac{N}{2p\beta}, \tag{23}$$

which can be compared to Akaikes Final Prediction Error estimate (Akaike, 1969; Ljung, 1984). Akaike computes the ratio between the test and training errors in the *deterministic limit* $\beta \to \infty$:

$$\overline{\epsilon_G} = \frac{p+N}{p-N}\overline{\epsilon_T}, \tag{24}$$

Using equation (23) and letting $\beta \to \infty$, we find:

$$\overline{\epsilon_G} = \frac{p}{p-N}\frac{p-1}{p-N-1}\overline{\epsilon_T} \tag{25}$$

which coincides with (24) in the limit of large training sets: $p \to \infty$.

Properties of linear models in the limit $p \to \infty$, but with $\alpha \equiv p/N$ constant, have been investigated by several workers (Krogh *et al.*,1989; Le Cun *et al.*, 1991; Krogh & Hertz, 1992). Krogh *et al.* found phase-transition like behaviour for $\alpha \to 1^+$, with divergent relaxation times and infinite test-error. This, of course, is compatible with our expression (25) in which the test-error blows up at $p = N + 1$. In particular we find that our relation reads:

$$\overline{\epsilon_G} = \left(\frac{\alpha}{\alpha - 1}\right)^2 \overline{\epsilon_T} \quad (\alpha > 1) \tag{26}$$

in the limit $N, p \to \infty$. Note that equation (26) only agrees with Akaike's expression (24): $\overline{\epsilon_G} = (\alpha + 1)/(\alpha - 1)\overline{\epsilon_T}$ to first order in $1/\alpha$. In the present context a more precise statement of Akaike's result would read: $\overline{\epsilon_G} = (1 + 2/\alpha)\overline{\epsilon_T} + O(\alpha^{-2})$.

In conclusion we have derived exact averages of the training and test errors of a linear model. Our results are valid for a stochastic learning algorithm if the following conditions are met:

- The post training distribution is Gibbsian.
- The difference between the training set size and the model dimension exceeds *one*: $p - N > 1$.
- The distribution of the inputs is multinormal and the input and the noise terms are independent.

Our results when evaluated in the deterministic limit, extend those of Akaike to finite training set sizes for simple linear learning.

References

Akaike, H. (1969). Fitting Autoregressive Models for Prediction. *Annals of The Institute of Statistical Mathematics*, **21**, 243-347.

Akaike, H. (1974). A new Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, **AC-19**, 716-723.

Anderson, T.W. (1984). **An Introduction to Multivariate Statistical Analysis**. New York: John Wiley & Sons.

Fogel, D.B. (1991). An Information Criterion for Optimal Neural Network Selection. *IEEE Transactions on Neural Networks*, **2**, 490-497.

Hansen, L.K., Pathria, R., and Salamon, P. (1991). Stochastic Dynamics of Supervised Learning. *Electronics Institute Preprint*, The Technical University of Denmark

Hertz, J.A., Krogh, A., and Palmer R.G. (1989). **Introduction to the Theory of Neural Computation**. Santa Fe Institute Lecture Notes **I**. Addison Wesley Publ.Co..

Hoffmann, N. and Larsen, J. (1991). Algorithms for Model Order Reduction in Nonlinear Filters. *Electronics Institute Preprint*, The Technical University of Denmark.

Ljung, L. (1987). **System Identification: Theory for the User**. Englewood Cliffs, New Jersey: Prentice Hall.

Le Cun, Y., Denker, J.S., and Solla, S.A. (1989). Optimal Brain Damage. In D.Touretzky (Ed.) **Neural Information Processing Systems, Denver 1989**, (pp 598-605). Morgan Kaufman.

Le Cun, Y., Canter, I., and Solla, S. A. (1991). Eigenvalues of Covariance Matrices: Application to Neural-Network Learning. *Physical Review Letters*, **66**, 2396-99.

Levin, E., Tishby, N., and Solla, S.A. (1990). A Statistical Approach to Generalization in Layered Neural Networks. *Proceedings of the IEEE*, **78**, 1568-1574.

Krogh, A., Hertz, J.A., and Thorbergsson, G.I. (1989). Phase Transitions in Simple Learning. *Journal of Physics A: Math.Gen.*, **22**, 2133-2150.

Krogh, A. and Hertz, J.A. (1992). Generalization in a Linear Perceptron in the Presence of Noise. *Journal of Physics A: Math.Gen.*, in press.

Radons, G., Schuster, H.G., and Werner, D. (1990). Drift and Diffusion in Backpropagation Learning. In R. Eckmiller (Eds.), **Parallel Processing in Neural Systems and Computers**, (pp 261-266). Elsevier Science Publishers.

Toussaint, G.T. (1974). Bibliography on Estimation of Misclassification. *IEEE Transactions on Information Theory*, **IT-20**, 472-479.