Optimal Ensemble Averaging of Neural Networks

Ury Naftaly* Nathan Intrator† David Horn‡
Raymond and Beverly Sackler Faculty of Exact Sciences
Tel Aviv University, Tel Aviv 69978, Israel

May, 1997

Abstract

Based on an observation about the different effect of ensemble averaging on the bias and variance portion of the prediction error, we discuss training methodologies for ensembles of networks. We demonstrate the effect of variance reduction and present a method of extrapolation to the limit of an infinite ensemble. A significant reduction of variance is obtained by averaging just over initial conditions of the neural networks, without varying architectures or training sets. The minimum of the ensemble prediction error is reached later than that of a single network. In the vicinity of the minimum, the ensemble prediction error appears to be flatter than that of the single network, thus simplifying optimal stopping decision. The results are demonstrated on the sunspots data, where the predictions are among the best obtained, and on the 1993 energy prediction competition data-set **B**.

1 Introduction

In recent years, the use of artificial neural networks (NN) for time series prediction has gained popularity and nowadays, NN can compete with the best time series methods [1]. In this paper we reexamine one of the major techniques for NN performance improvement – ensemble averaging [2, 3]. We argue that it requires a special training methodology, and can be more effective when **not** combined with popular training constraints such as weight decay and early stopping[1].

The theoretical setting of the method is provided by the bias/variance decomposition. Within this framework, we will define a particular bias/variance decomposition for networks differing by their initial conditions only. This is a particularly useful subset of the general set of all sources of variance. We show that while the bias of the ensemble of networks with different initial conditions remains unchanged, the variance error decreases considerably. The theoretical background

*School of Physics and Astronomy, ury@tarazan.tau.ac.il
†School of Mathematical Sciences, nin@math.tau.ac.il
‡School of Physics and Astronomy, horn@vm.tau.ac.il

[1] Under a different setup (training with input noise) weight decay was found useful in conjunction with ensemble averaging [6]

is presented in the next section. We then describe the sunspots data, on which our technique is demonstrated. This includes a simple method of extrapolation to the infinite ensemble. We show that the minimal prediction error of the ensemble is reached *later* in training than that of single networks, and the ensemble error curve appears to be flatter in the vicinity of the minimum error. Results for the sunspots problem are presented in Section 5, where we also evaluate the combination of ensemble averaging with other popular techniques. Our results outperform the best published results [4].

Our method is further evaluated with a data set from the 1993 energy competition (Section 6). The extrapolation method works somewhat differently in this case, signifying the existence of correlations among networks with different initial conditions. The qualitative behavior of the error minima are the same as those of the sunspots analysis.

2 The Bias/Variance Decomposition and Ensemble Averaging

The motivation of our approach follows from a key observation regarding the bias variance decomposition, namely the fact that ensemble averaging does not affect the bias portion of the error, but reduces the variance, when the estimators on which averaging is done are independent.

The classification problem is to estimate a function $f_{\mathcal{D}}(x)$ of observed data characteristics x, predicting class label y, based on a given training set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_L, y_L)\}$, using some measure of the estimation error on $\mathcal{D}$. A good estimator will perform well not only on the training set, but also on new *validation* sets which were not used during estimation.

Evaluation of the performance of the estimator is commonly done via the mean squared error distance (MSE) by taking the expectation with respect to the (unknown) probability distribution P of y:

$$E[(y - f_{\mathcal{D}}(x))^2 | x, \mathcal{D}].$$

This can be decomposed into

$$E[(y - f_{\mathcal{D}}(x))^2 | x, \mathcal{D}] = E[(y - E[y|x])^2 | x, \mathcal{D}] + E[(f_{\mathcal{D}}(x) - E[y|x])^2].$$

The first term does not depend on the training data $\mathcal{D}$ or on the estimator $f_{\mathcal{D}}(x)$, it measures the amount of noise or variability of y given x. Hence f can be evaluated using

$$E[(f_{\mathcal{D}}(x) - E[y|x])^2].$$

The empirical mean squared error of f is given by

$$E_{\mathcal{D}}[(f_{\mathcal{D}}(x) - E[y|x])^2],$$

where $E_{\mathcal{D}}$ represents expectation with respect to all possible training sets $\mathcal{D}$ of fixed size.

To further see the performance under MSE we decompose the error to bias and variance components [5] to get

$$E_{\mathcal{D}}[(f_{\mathcal{D}}(x) - E[y|x])^2] = (E_{\mathcal{D}}[f_{\mathcal{D}}(x)] - E[y|x])^2 + E_{\mathcal{D}}[(f_{\mathcal{D}}(x) - E_{\mathcal{D}}[f_{\mathcal{D}}(x)])^2]. \quad (1)$$

The first RHS term is called the bias of the estimator and the second term is called variance. When training on a fixed training set $\mathcal{D}$, reducing the bias with respect to this set may increase the variance of the estimator and contribute to poor generalization performance. This is known as the tradeoff between variance and bias. Typically variance is reduced by smoothing, however this may introduce bias (since, for example, it may blur sharp peaks). Bias is reduced by prior knowledge. When prior knowledge is used also for smoothing, it is likely to reduce the overall MSE of the estimator.

When training neural networks, the variance arises from two terms. The first term comes from inherent data randomness and the second term comes from the non-identifiability of the model, namely, the fact that for a given training data, there may be several (local) minima of the error surface[2].

Consider the ensemble average $\bar{f}$ of Q predictors, which in our case can be thought of as neural networks with different random initial weights which are trained on a fixed training set:

$$\bar{f}(x) = \frac{1}{Q}\sum_{i=1}^{N} f_i(x).$$

These predictors are identically distributed and thus, the variance contribution (second term on the RHS of Equation 1) becomes: (we omit x and $\mathcal{D}$ for simplicity)

$$\begin{aligned} \mathrm{Var}(\bar{f}) = E[(\bar{f} - E[\bar{f}])^2] &= E[(\frac{1}{Q}\sum f_i - E[\frac{1}{Q}\sum f_i])^2] \\ &= E[(\frac{1}{Q}\sum f_i)^2] - \Big(E[\frac{1}{Q}\sum f_i]\Big)^2. \end{aligned} \tag{2}$$

The first RHS term can we rewritten as

$$E[(\frac{1}{Q}\sum f_i)^2] = \frac{1}{Q^2}\sum E[f_i^2] + \frac{2}{Q^2}\sum_{i<j} E[f_i f_j],$$

and the second term gives,

$$\Big(E[\frac{1}{Q}\sum f_i]\Big)^2 = \frac{1}{Q^2}\sum E[f_i] + \frac{2}{Q^2}\sum_{i<j} E[f_i]E[f_j].$$

Plugging these equalities in Equation 2 gives:

$$E[(\bar{f} - E[\bar{f}])^2] = \frac{1}{Q^2}\sum\{E[f_i^2] - \Big(E[f_i]\Big)^2\} + \frac{2}{Q^2}\sum_{i<j}\{E[f_i f_j] - E[f_i]E[f_j]\}. \tag{3}$$

It follows that

$$\frac{1}{Q}\mathrm{Var}(f_i) \le \mathrm{Var}(\bar{f}) \le \frac{\mathrm{Var}(f_i) + max_{i,j}(E[f_i f_j] - E[f_i]E[f_j])}{Q} \le \max_i \mathrm{Var}(f_i). \tag{4}$$

[2]An example of an identifiable model is (logistic) regression.

More specifically, when replacing $f(x)$ by $\bar{f}(x)$, the reduction in the variance portion of the error is proportional to the degree of independence between the predictors in the ensemble. Due to random initial conditions only, a certain level of independence may be achieved. The independence can be increased by various ways such as different architectures, input noise injection, different training times. Thus the variance portion of the error for ensemble has the form

$$E_{\mathcal{D}}[(\bar{f}_d - E_{\mathcal{D}}[\bar{f}_d])^2] \simeq \frac{E_{\mathcal{D}}[(f_d - E_{\mathcal{D}}[f_d])^2] + \gamma}{Q},$$

where $\gamma \geq 0$ is given by:

$$\gamma = E[f_i f_j] - E[f_i]E[f_j] = E\Big(\{f_i - E[f_i]\}\{f_j - E[f_j]\}\Big),$$

thus, the notion of independence can be understood as independence of the deviations of each predictor from the expected value of the predictor, which can be replaced (due to linearity) by

$$E\Big(\{f_i - E[\bar{f}]\}\{f_j - E[\bar{f}]\}\Big),$$

and is thus interpreted as an independence of the prediction variation around a common mean.

We wish to find the optimal training procedure for reducing the error of our the ensemble average. Traditional training algorithms aim at reducing the error of the individual NN, i.e., the total expression of Equation 1, including both bias and variance. Typically the bias decreases and the variance increases as one employs more and more training epochs, and one aims to stop when their sum reaches a minimum. Since in our algorithm the predictor is defined by an ensemble average we have to search for a different minimum, as we are able to eliminate some portion of the variance of the estimator via ensemble averaging. We should, thus, search for a point with a smaller bias (longer training time) as the optimal tradeoff for ensemble predictor.

3 The Sunspots Problem

Yearly sunspot statistics have been gathered since 1700. The data are plotted in Figure 1. These data have been extensively studied and have served as a benchmark in the statistical literature [7, 8, 9]. Following previous publications [8, 4, 10] we choose the training set to contain the period between 1701 and 1920, and the test-set to contain the years 1921 to 1955. Following [8], we calculate the prediction error according to the average relative variance (ARV)

$$\mathrm{ARV} = \frac{\sum_{k \in S} \left(y_k - f(\vec{x}_k)\right)^2}{\sum_{k \in S} \left(y_k - E\left[y_k\right]\right)^2} \tag{5}$$

which is the MSE divided by the variance of the data set S. The denominator is $\sigma^2 = 1535$ for the training set. The same value is used for the test set.

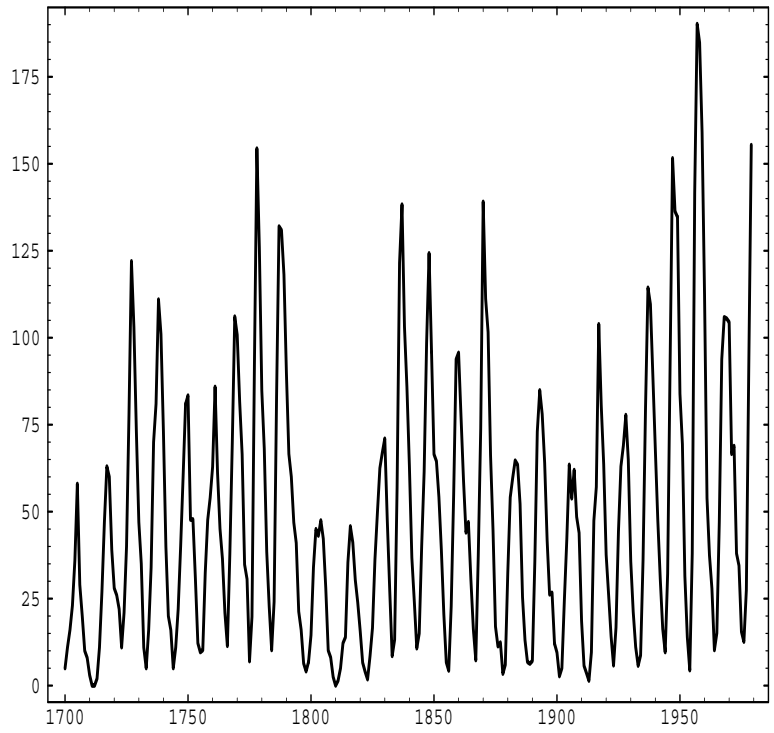

Figure 1: *Sunspots Data: Average sunspot activity from 1700 to 1979*

3.1 Previous Results

In a survey of sunspots prediction models [11], the Threshold Autoregressive model (TAR) of Tong and Lim [12, 13] was favored. The TAR model is a combination of two linear autoregressive models with an activity threshold above which one auto-regression model is used and below which the other is used [14]. The TAR estimator gave a training ARV of 0.097 and the same result for the prediction set.

Weigend et al. [8] used a standard multi-layer perceptron architecture with 12 input units, 8 sigmoidal hidden units and a linear output unit. Possible over-fitting was addressed by the use of weight decay [15]. Their best result was an ARV of 0.082 on the training set and 0.086 on the prediction set.

Nowlan and Hinton [4] imposed a mixture of Gaussians prior on the weights which they called "Soft Weight Sharing" to get an ARV of 0.072 on the test set. Pi and Peterson [10] introduced the δ-test which establishes the dependence of a sequence of numbers on previous element(s) of the sequence. They found that x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4}, x_{t-9} and x_{t-10} are the most important variables in the sunspots series. Thus, the functional form $y = f(x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4}, x_{t-9}, x_{t-10})$

is expected to lead to good prediction results. These lags were used as inputs to a (6,8,1) network. The ARV obtained on the test-set was 0.073.

4 Ensemble Averaging over Initial Conditions

In this section we will demonstrate the method we use for ensemble averaging over initial conditions by applying it to the sunspots problem. We use neural networks with 12 inputs (as in [8]). All nets have one sigmoidal hidden layer consisting of 4 units and a linear output. They are then enlarged to form recurrent networks (SRN) [16] in which the input layer is increased by adding to it the hidden layer of the previous point in the time series. This favors ordered temporal application.

The learning algorithm consists of back propagation applied to an error function which is the MSE of the training set. A learning rate of 0.003 is employed. A validation set containing 35 randomly chosen points was left out during training to serve for performance validation.

The training procedure of a NN starts out with some choice of initial values of the connection weights. We consider then an ensemble of networks that differ from one another just by these initial values. This defines the ensemble that we wish to average over. Since the space of initial conditions is very large we develop a technique which allows us to approximate averaging over the whole space.

Our technique consists of constructing groups of a fixed number of networks, Q. Choosing several different groups of the same size Q, and averaging over them, we define a finite size average. Then we try to estimate the limit $Q \to \infty$. If we regard the specific choice of initial conditions to be equivalent to some random error added to the predictor, we may expect this error to decrease as $1/Q$. This is proved in the Appendix. It turns out that performing a simple regression in $1/Q$ indeed suffices to obtain this limit.

Figure 2 displays our results on the validation set and Figure 3 shows the results on the test set. In both figures we show ARV values as function of the number of training epochs. The highest curves in both figures correspond to $Q = 1$, i.e. the case of single networks. Below it appear the curves of $Q = 2, 4, 10, 20$ followed by the extrapolation to $Q \to \infty$. To demonstrate how the extrapolation is carried out we display in Figure 4 the points obtained for $t = 70$ and $t = 140$ KE for the test set as a function of $\frac{1}{Q}$. It is quite clear that a linear extrapolation is very satisfactory. Moreover, the results for $Q = 20$ are not far from the extrapolated $Q \to \infty$ results.

In this example (Figure 3) ARV is quite flat as a function of t, suggesting that most of the variance was eliminated by the averaging process. The variance that is due to initial conditions can be found by subtracting the $Q \to \infty$ result from $Q = 1$ (Figure 5). Although this variance is due to initial conditions and not to different training sets, it is increasing in time (as would be expected from the variance due to training sets).

Several interesting conclusions follow from this numerical study:

1. The true minimum of the final predictor is obtained at larger t values than that of the $Q = 1$ curve. Note that if we were to use canonical techniques of single network training, we would have stopped around $t = 50$ KE, where the validation set shows minimal ARV for $Q = 1$. This is however very different from where the $Q \to \infty$ curve goes through minimum, both in the cross-validation

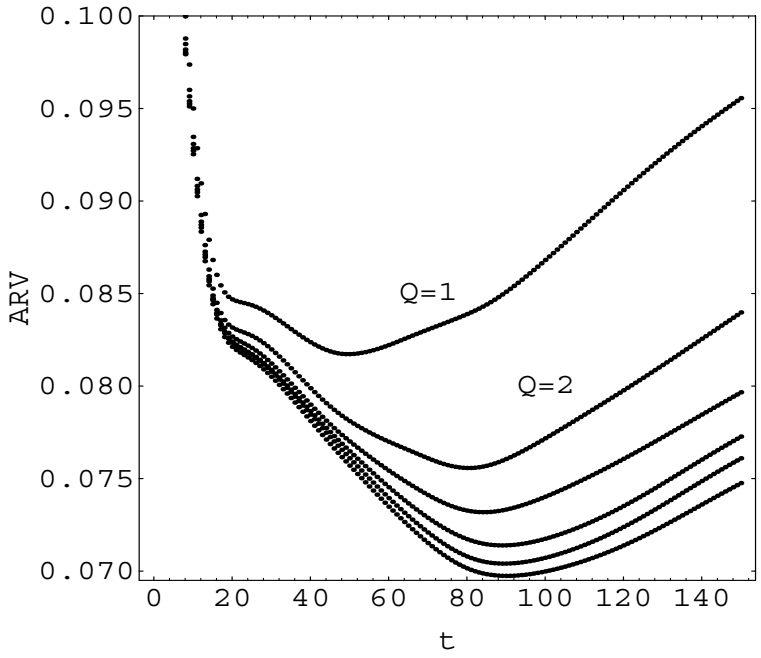

Figure 2: *ARV of cross-validation sets*
ARV is plotted *vs.* training time in kilo epochs (KE). These are results for the cross-validation set of 35 points in the sunspots problem. The curves are shown for different choices of group sizes: $Q = 1, 2, 4, 10, 20$ from top to bottom. The lowest curve is the extrapolation to $Q \to \infty$.

---

set and in the test set. This property of ensemble training follows from the fact that one portion of the variance contribution to the error is removed or reduced by the ensemble averaging, and thus less smoothing (bias) is needed for regularizing the predictor.

2. The error function of the final predictor is shallower than that of the single network, which is obviously due to the reduction in variance.

The final $Q \to \infty$ error function is quite flat, and therefore ARV differences between different stopping points are not large. Nonetheless, for accurate estimates one should beware of the fact that the stopping criterion for the $Q = 1$ problem is very different from the $Q \to \infty$ one. The latter is the one which is relevant to our problem.

The curves shown in the figures of this section were obtained with a learning rate of 0.003. We will see in the next section that better results can be obtained for slower learning rates, as well as for other choices of architecture. When lower errors are obtained, the variation between the single networks and the ensemble is less dramatic. Yet in all cases we find the general characteristics of variance reduction and shift in the temporal structure of the error functions.

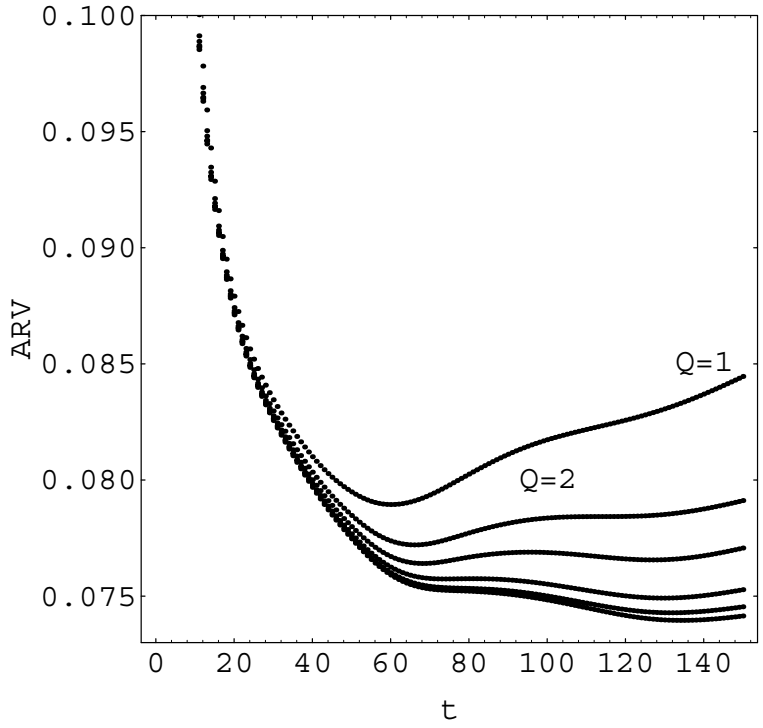

Figure 3: ARV of test set: Results for the test set show a shallow $Q \to \infty$ curve, and the two minima. The setup is the same as in Figure 2.

5 Analysis of the Sunspots Problem

From the curves of Figs. 2 and 3 we may read the performance values of the predictors which can be defined through different stopping criteria. Using $Q = 20$ data and the conventional stopping criterion, i.e. the minimum of the $Q = 1$ curve for the cross-validation set of data at $t = 49$, we are led to ARV=0.0796 on the test set. Stopping at $t = 90$, which is the minimum point of $Q = 20$ on cross-validation data, leads to ARV=0.0752 on the test set. The absolute minimum of the $Q = 20$ curve on the test set, ARV=0.075, is reached at $t = 132$.

We repeat the whole exercise for different learning rates and find that the results are very sensitive to this parameter. This is demonstrated in table 1 which is based on $Q = 20$ groups of networks in each case. Clearly lower learning rates lead to better results. Unfortunately, they require an immense number of training epochs and are very costly.

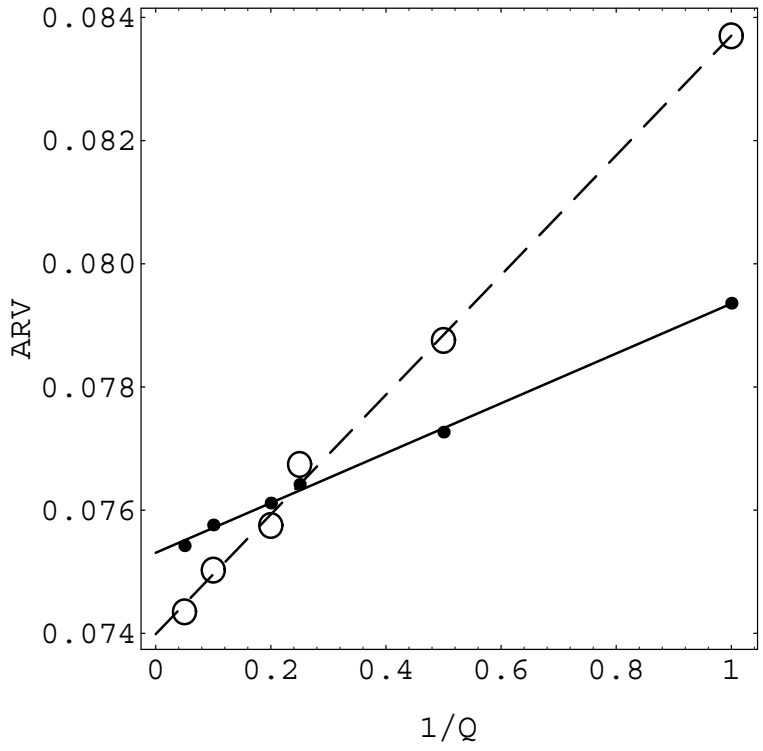

Figure 4: Extrapolation method used for extracting the $Q \to \infty$ prediction. The results for different Q at two different training periods, $t = 70$ (top) and 140KE (bottom), can be extrapolated by a linear regression in $1/Q$.

5.1 Recurrent networks

Recurrent neural networks have been successful as time-series predictors. Simple recurrent networks (SRN) [16] are a practical compromise between fully recurrent dynamics and computational overload, and are being widely used [18].

In such networks, the connections are mainly feed-forward, but include some feedback connections. The recurrency lets the network remember cues from the recent past, but does not appreciably complicate the training. We have already presented in the previous section the results of a simple recurrent network [17]. Here we compare it to a simple feed-forward network (Table 2). We see that recurrency helps lowering the ARV on both ensemble and single nets, but its influence on single nets is about four times stronger than on the ensemble. We note that the improvement of the SRN ensemble is smaller than that of the FF. Since the reduction of errors is due to variance elimination, this means that SRN networks are less independent than FF nets. This sounds

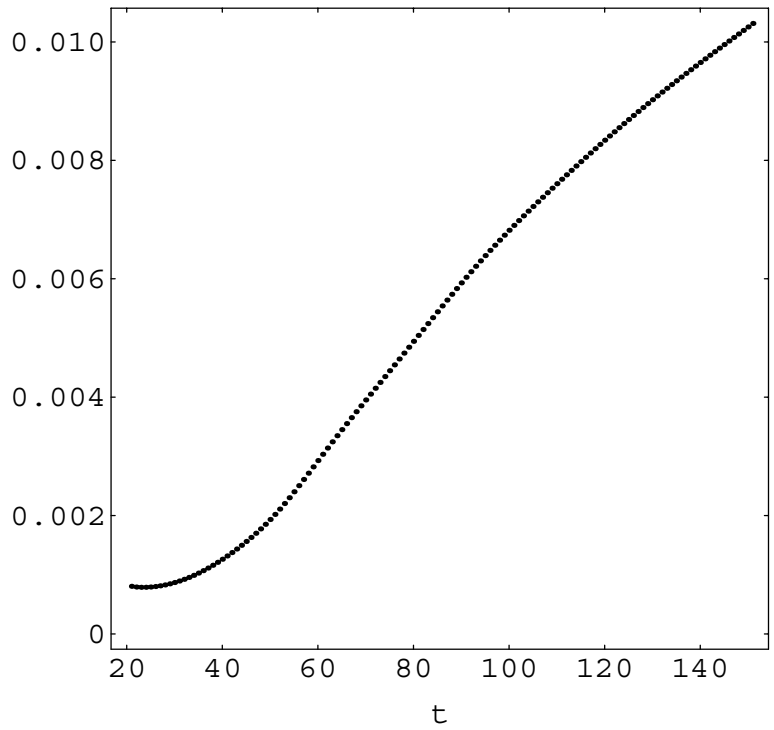

Figure 5: Variance error due to initial conditions is estimated by the difference of the ARV values for $Q = 1$ and for $Q \to \infty$.

---

plausible because of the extra dependence of an SRN on the previous data point.

5.2 Selected Inputs

Pi & Peterson [10] report very good results using a (6,8,1) network where the input vectors consist of the variables x_{t-1}, x_{t-2}, x_{t-4}, x_{t-9} and x_{t-10}. Applying the same inputs to our paradigm, indeed improved the prediction results of the ensemble (Table 3). The result of 0.0674 is better than any previously reported result.

Note that we employed here a very low learning rate which, together with the choice of the Pi & Peterson variables led to this remarkable result. We have already remarked before that in such a case the variance is no longer as strong as in the examples studied above, where a learning rate that is six times faster was employed. For completeness we present in figure 6 the $Q = 1$ and $Q = 20$ curves for this case. Error reduction due to ensemble averaging is small, nonetheless it is significant.

Learning Rate	ARV
0.010	0.0868
0.003	0.0731
0.002	0.0720
0.001	0.0713

Table 1: *The minimal* ARV *on the sunspots test-set as a function of the learning rate for simple networks.*

Learning rate: 0.001		
	SRN	FF
Ensemble	0.0706	0.0713
Single	0.0739	0.0764

Table 2: *The prediction* ARV *of recurrent and non-recurrent networks trained with the same learning and prediction conditions.*

5.3 Weight Decay.

It is well known [18, 8] that a large network cannot generalize well. Therefore, it is advisable to use the smallest network that is able to fit the training data. Many algorithms were used to achieve this goal [18, 8, 4]. Here we use weight-decay [8] which lets the network itself decrease non-useful connections during training. The cost function is

$$\text{MSE} + \frac{1}{2}\lambda \sum_i \frac{w_i^2}{1+w_i^2} \tag{6}$$

where the sum is over the network weights, and λ is the smoothing parameter to be adjusted.

In [8] the authors use this cost function to predict the sunspots series. We repeat their experiment using a $Q = 20$ ensemble of (12,8,1) networks. We get the results presented in Table 4. The main effect of the weight decay algorithm is to reduce the ARV of the single networks, while it has almost no effect on the ensemble ARV.

A simple weight decay which adds a penalty of the form $\lambda \sum_i w_i^2 f(w_0^2 + w_i^2)$ to the energy function, gave no significant improvement on the ensemble prediction, although improving the

Learning rate: 0.0005		
	12-inputs	6-inputs
Ensemble	0.0700	0.0674
Single	0.0730	0.0698

Table 3: *The prediction* ARV *of recurrent networks trained with full input vectors (12-inputs) and of δ chosen lags (6-inputs) using a learning rate of 0.0005.*

	with decay	without decay
Ensemble	0.08138	0.08142
Single	0.08331	0.08578

Table 4: *The prediction* ARV *of networks with and without weight-decay trained with the same learning and prediction conditions.*

results of single networks.

6 Analysis of Energy Competition Data

Here we use another data-set to demonstrate our findings: data-set **B** of the 1993 energy competition, which consists of 3344 measurements of four input variables at hourly intervals during daylight (see figure 7). The physical source of the data were measurements of solar flux from five outdoor devices. Four of the devices were fixed. The fifth, whose output was to be predicted, was driven by motors so that it pointed at the sun. For more information and results see [19].

Our training set consists of 1000 vectors containing the 4 variables, the decimal date and the 5 last values of the target. Test and cross-validation sets that contain 500 vectors each are formed in the same manner. A similar methodology to the one used for the sunspots data was applied to these data: we use 60 simple (non-recurrent) feed-forward networks of 10 inputs, one sigmoidal hidden layer of 8 hidden units and a linear output unit. Setting the learning rate to 0.02 we get the MSE of ensembles consisting of $1, 2, 3, 4, 5, 6, 10, 12, 15, 20, 30, 60$ nets, depicted in Figure 8. Again we see the same effects that were observed on the sunspots data-set: 1) a shift of the minimum and 2) shallower curve of the ensemble MSE.

There is one notable difference between this analysis and the former one on sunspots: we no longer obtain simple $1/Q$ dependence. This is demonstrated in Figure 9 where we plot the behavior of the MSE as a function of $1/Q$. Since a linear regression does not capture the trend of the data, we have employed a power law dependence:

$$\text{MSE} = aQ^{-k} + b. \tag{7}$$

The results of the fit indicate that $k = 0.81$. This serves as the basis for producing the dots in Figure 8 that represent the extrapolation to infinite Q. The fact that $k < 1$ indicates that there exists a correlation between the networks with different initial conditions, as explained in the Appendix.

7 Conclusions

We have shown that ensemble averaging is a powerful procedure which, when used correctly, improves on single network performance. Ensemble averaging is not an alternative to methods for

introducing bias and reducing variance such as smoothing or early stopping, as it does not eliminate variance that is due to training on a limited training set. When the portion of the variance due to initial conditions is large, ensemble averaging is most effective. By using large ensembles we may eliminate it altogether.

As demonstrated on the sunspots data set, we were able to perform a simple extrapolation to infinite ensemble size. The variance turned out to be inversely proportional to the size, demonstrating that, in this case, some independence between predictors can be achieved by varying the initial conditions.

Our experience suggests that when training for optimal ensemble performance, the training method and stopping criteria have to be chosen carefully. This is because stopping criteria based on single network training may not be useful for the ensemble. In fact, ensemble results are improved if single nets are over-trained. Instead of early stopping or even stopping when the validation error reaches a minimum, further training has to be done, so that the bias portion of the error will be further reduced, while paying the price of higher variance for individual networks. Later, the variance portion of the error is reduced by the ensemble average with no effect on the bias. Finally, the reduction of variance, that is inherent in this method, leads to a flattening of the error curve as a function of training time.

We have applied our method both to the sunspots data and to the energy competition data. For both we have obtained considerable decrease of variance by averaging over the space of initial conditions. In the energy competition case we found the Q dependence to be more complicated, signifying nontrivial correlations between the different networks. Some decorrelation methods could turn out to be useful here. In the sunspots case, the Q dependence was as expected from statistical independence. Using the δ-test of Pi and Peterson [10] together with our method, we obtained the best predictions yet for this problem.

References

[1] Weigend, A. S. and Gershenfeld, N. A. (editors), 1994. *Time Series Prediction.* Addison-Wesley.

[2] Lincoln, W.P. and Skrzypek, J. 1990. *Synergy of clustering multiple back propagation networks.* In Touretzky, D. S., editors, *Advances in Neural Information Processing Systems 2,* pages 650–657, SanMateo, CA. Morgan Kaufmann.

[3] Wolpert, D. H. 1992. *Stacked generalization.* Neural Networks 5:241-259.

[4] Nowlan, S. J. & Hinton, G. E. 1992. *Simplifying neural networks by soft weight-sharing.* Neural Comp. 4, 473–493.

[5] Geman, S., Bienenstock, E., and Doursat, R. 1992. *Neural networks and the bias/variance dilemma.* Neural Computation, 4(1):1–58.

[6] Raviv, Y. and Intrator, N. 1996. *Bootstraping with Noise: An Effective Regularization Technique.* Connection Science, Special issue on combining estimators (To appear).

[7] Priestley, M. B. 1981. *Spectral Analysis and Time Series.* Academic Press.

[8] Weigend, A. S., Huberman, B. A., and Rumelhart, D. 1990. *Predicting the future: A connectionist approach.* Int. J. Neural Syst. 1, 193–209.

[9] Morris, J. (1977). *Forecasting the sunspot cycle.* J. of the Roy. Statist. Soc. Ser. A, 140, 437–447.

[10] Pi, H., and Peterson, C., 1994. *Finding the Embedding Dimension and Variable Dependencies in Time Series.* Neural Computation 6, 509–520.

[11] Priestley, M. B. 1988. *Non-linear and Non-stationary Time Series Analysis.* Academic Press.

[12] Tong, H. and Lim, K. S. 1980. *Threshold autoregression, limit cycles and cyclical data.* Journal of the Royal Statistical Society, 42:245.

[13] Tong, H. 1983. *Threshold Models in Non-linear Time Series Analysis,* volume 21 of *Lecture Notes in Statistics.* Springer-Verlag, New York.

[14] Tong, H. 1990. *Non-linear Time Series: A Dynamical Systems Approach.* Oxford University Press.

[15] Hinton, G. E. 1986. *Learning distributed representations of concepts.*In *Proceedings of the 8th Annual Conference of the Cognitive Science Society*, pp 1–12, Hillsdale: Erlbaum

[16] Elman J. L. and Zipser D. 1988. *Learning the Hidden Structure of Speech.* Journal of Acoustical Society of America 83, pp 1615–1626.

[17] Elman, J. L. 1990. *Finding Structure in Time.* Cognitive Science 14, pp. 179–211.

[18] Hertz, J., Krogh, A., and Palmer, R. G. 1991. *Introduction to The Theory of Neural Computation.* Lecture Notes Volume I, Santa Fe Institute. Addison-Wesley.

[19] MacKay D. J. C. 1994. *Bayesian Non-linear Modeling for the Prediction Competition.* In ASHRAE Transactions, V.100, Pt.2, pp. 1053-1062. ASHRAE, Atlanta Georgia.

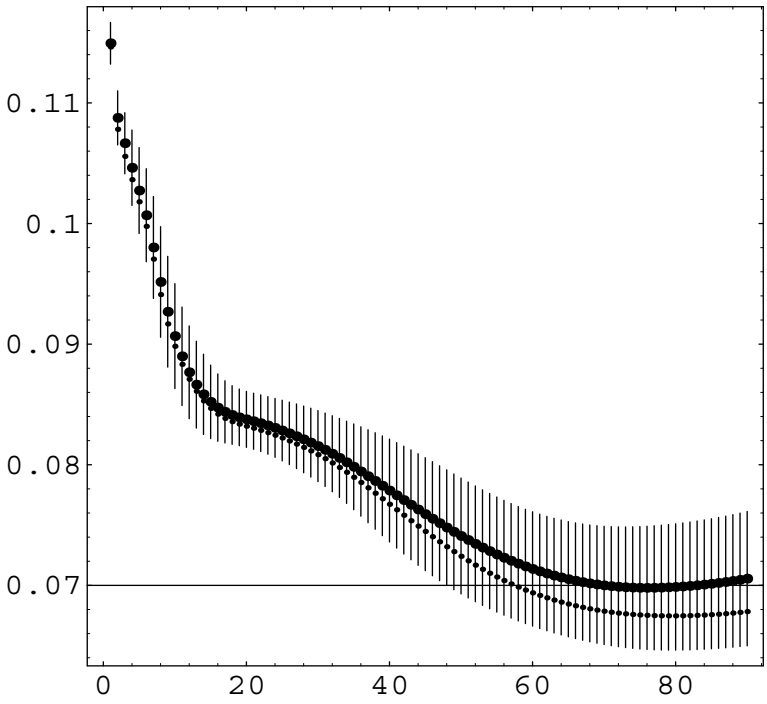

Figure 6: Our best results for the test set of the sunspots problem. Plotted here are $Q = 1$ results for various choices of initial conditions, represented by their averages with error-bars extending over a standard deviation, and $Q = 20$ results (the thinner points), as a function of training time in K-epochs. The network is based on the Pi & Peterson variables, and the learning rate is 0.0005.

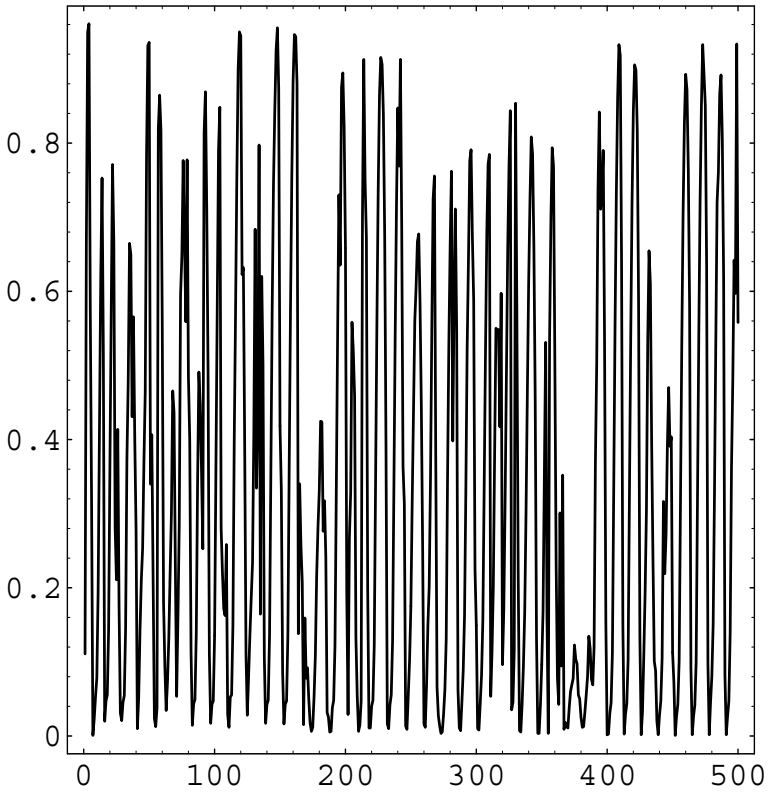

Figure 7: *500 elements of the 1993 energy competition data-set* **B**.

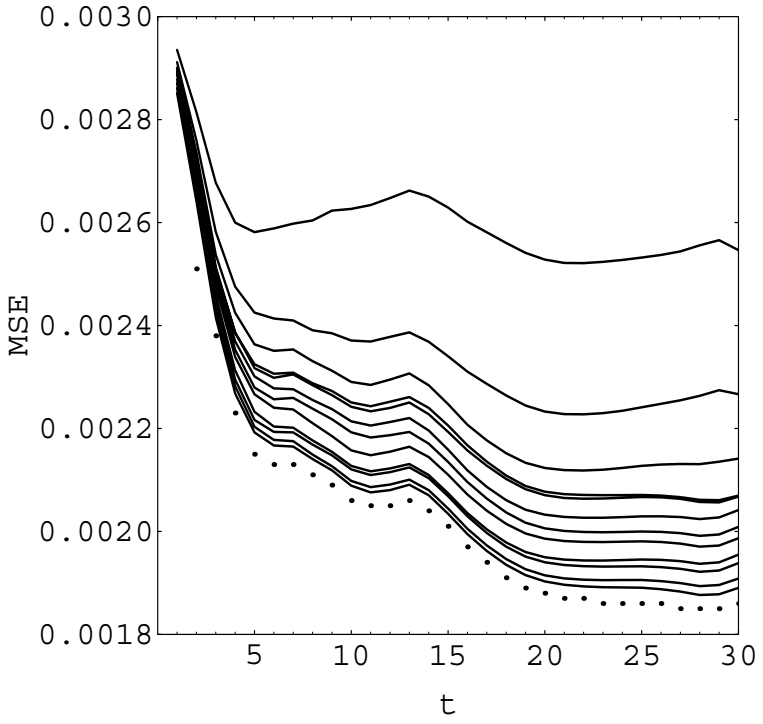

Figure 8: *Prediction* MSE *for the 1993 energy competition data-set* **B**.
Results for a test-set of 500 vectors. The curves are shown for the different choices of ensemble size: $Q = 1, 2, 3, 4, 5, 6, 10, 12, 15, 20, 30, 60$ from top to bottom. The dots are the extrapolation to $Q \to \infty$.

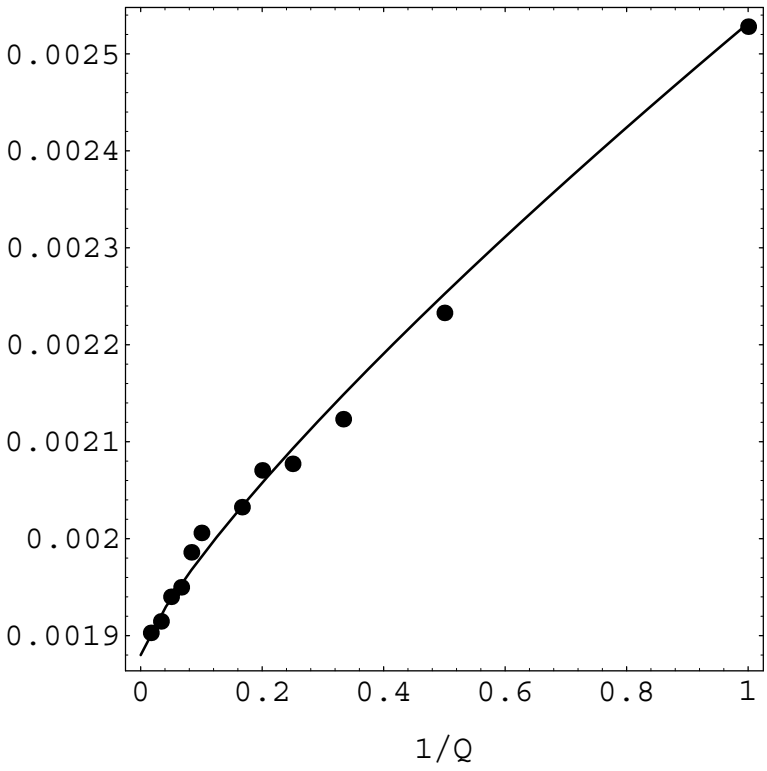

Figure 9: *Extrapolation method used in the energy competition data-set.*
Shown here are results for different Q at $t = 20$ KE. The solid line represents the best fit to a power law, described in the text, which leads to a decreasing component of $Q^{-0.81}$. This indicates that there exist correlations between the networks of different initial conditions.