Optimal, Unsupervised Learning in Invariant Object Recognition

Guy Wallis[1]

Max-Planck Institute für biologische Kybernetik, Spemannstraße 38, 72076 Tübingen, Germany. Email: guy@mpik-tueb.mpg.de

Roland Baddeley

Dept. Experimental Psychology, University of Oxford, South Parks Road, Oxford OX1 3UD, United Kingdom.

A means for establishing transformation invariant representations of objects at the single cell level is proposed and analysed. The association of views of objects is achieved by using both the temporal order of the presentation of these views, as well as their spatial similarity. Assuming knowledge of the distribution of presentation times, an optimal linear learning rule is derived. If we assume that objects are viewed with presentation times that are approximately Jeffrey's distributed ($P(\tau) \approx \tau^{-1}$), then the optimal learning rule is very well approximated using a simple exponential temporal trace. Simulations of a competitive network trained on a character recognition task are then used to highlight the success of this learning rule in relation to simple Hebbian learning, and to show that the theory can give quantitative predictions for the optimal parameters for such networks.

1 Introduction

How might we learn to recognise an object irrespective of the precise relationship between viewer and object, characterised by viewing angle, distance and translation. If we believe the view center approach to object recognition (Tarr & Pinker, 1989; Bülthoff & Edelman, 1992) then the problem becomes one of associating together a series of different views of the object which may share few, if any, of the features supporting the recognition. A broadly tuned feature based system would be sufficient to perform recognition over small transformations (Poggio & Edelman, 1990), and the form of the necessary receptive fields might be learnt via a simple competitive network with Hebbian learning. However, large shape transformations would either require separate pre-normalisation for size and translation, or separate feature detectors feeding into a final arbitration layer. This then raises the question of how such a final layer might be trained - dismissing the use of any form of supervised training signal.

[1] To whom correspondence should be sent.

The use of pre-normalisation and a 'final arbiter' is in fact in contrast to the evidence we have from the responses of real neurons implicated in object recognition, which seem to build up invariance from cells with initially restricted receptive fields, over a series of visual stages (Perrett 1993), culminating in the types of cell responses found in inferior temporal (IT) cortex. Cells in this region exhibit invariance to combinations of the types of transformations discussed here (Rolls, 1992; Desimone, 1991; Tanaka *et al.*, 1991).

Work by (Miyashita, 1988) has illustrated that stimuli can be arbitrarily associated together by a single IT neuron distinct from many other spatially similar images. The key to the training he used was the association of these images, not in space, but in time. The hypothesis expressed here is that temporal relations in the appearance of objects underpin learning. In the real world we see objects for protracted if variable periods whilst they undergo any manner of natural transformations - such as when we approach an object or rotate it in our hand. The consistent stream of images serves as a cue that all of these images belong to the same object and that it would be expedient to associate them together.

This idea has been described elsewhere (Edelman & Weinshall, 1991; Földiák, 1991; Wallis & Rolls, 1995), but in this work we derive the form an unsupervised learning rule should take, given a general probability function governing the length of each exposure to images of the same object. This learning rule bases training upon nothing more than the local activity of the neuron. It is the aim of this paper to build a theoretical framework for analyzing the response of a neuron over time and to derive the exact form of an optimal training rule, itself determined solely by the temporal structure of the world to which the neurons are exposed.

2 Mapping the output of a neuron to the optimal training signal

2.1 Introduction

The ideal signal with which to train a neuron selective for images of a particular stimulus or stimulus set, is simply the one which permits training if and only if an image of its preferred stimulus is presented. In practice, this external signal is not available to a neuron. Instead, perhaps the most obvious signal present in a neuron for performing learning is the activity of the neuron itself and indeed instantaneous activity forms the training signal in straightforward Hebbian learning. Unfortunately, this Hebbian training signal will not typically be optimal - due to erroneous classifications made by the neuron to spatially similar images from different objects, and spatially dissimilar images derived from the same object.

There is however, one extra piece of information available to a neuron, namely that objects tend to be viewed over variable but extended periods of time. In this work the consequent temporal structure will be used to provide information for improving the use of local neural activity as a training signal. This is achieved here by using a weighted sum of previous neuronal activity rather than by attending to the current neuronal output alone.

Assuming the neuronal output to be a signal described by the function $y(t)$, and the ideal training signal to be $s(t)$, discrepancies between the two signals can be regarded as being due to errors in stimulus classification, themselves due to some noise signal $n(t)$. Couched in these terms, the optimal recovery of $s(t)$ from $y(t)$ can be regarded as a classical filtering problem.

The form of the optimal linear filter for retrieving the original signal can be derived by a process usually referred to as 'Optimal Linear' or 'Wiener' Filtering (Wiener, 1949; Press, 1992)[2]. The Wiener Filter $\phi(t)$, gives an optimal estimate $\tilde{s}(t)$, of the true signal $s(t)$, when applied to the noise contaminated output $y(t)$. Adopting the convention that capital letters denote functions transformed into the frequency domain, the theorem states:

$$\tilde{S}(f) = Y(f)\Phi(f)$$

where

$$\Phi(f) = \frac{|S(f)|^2}{|S(f)|^2 + |N(f)|^2}$$

Hence, the optimal linear filter is a simple combination of the estimated system noise and a characterisation of $s(t)$, the perfect signal sought. The following two sections consider the predicted optimal filter for three different forms of $s(t)$, each describing a different regime for stimulus presentation.

2.2 Fixed Length Presentation Times

As a simple first case a set of stimuli are assumed to appear in identical, fixed length sequences. Possible forms of the ideal training signal and neural output associated with this training regime appear in figure 1.

The form of $s(t)$ is easily characterized but there remains the question of how to represent the noise signal $n(t)$. In these calculations it is assumed to be white, i.e. that it is uncorrelated across time. There may, in fact, be cases where this assumption will fail, including the probable increase in false negatives over false positives - i.e. the rate of classification errors during the period when the neuron should fire but does not over when it should not fire but does. However,

[2]Note that what is sought here is the optimal *linear* filter. There are presumably innumerable, more optimal, *nonlinear* filters, which are beyond the scope of the Wiener Filtering theory given here.

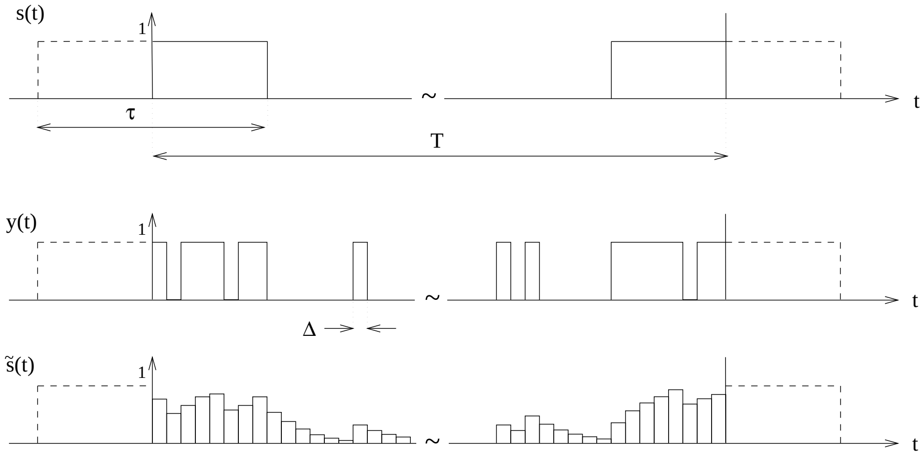

Figure 1: The idealized training signal $s(t)$, a possible neural output function $y(t)$ and an example of an approximation $\tilde{s}(t)$ to $s(t)$, gained by taking a weighted average of $y(t)$ over the current and a small range of previous time steps. T represents the period of an entire training epoch, in which all stimuli are seen once by the network, τ corresponds to the period over which all versions of a single stimulus are presented and Δ the time for a single stimulus presentation. Note that the notches and narrow peaks in $y(t)$ represent the erroneous responses of the neuron which appear smoothed out in the signal $\tilde{s}(t)$.

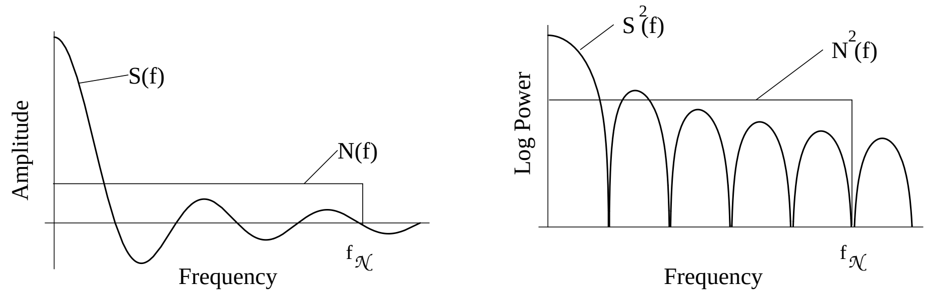

Figure 2: Amplitude spectrum and log power spectrum for the ideal training signal $s(t)$ and the generalized noise signal $n(t)$, for which $f_{\mathcal{N}} = 1/2\Delta$ is the Nyquist frequency.

the validity of this assumption should improve when taken as the average affect of noise across all neurons, and can indeed also be examined, at least in the restricted cases presented in this paper. The results of such a test form part of the experimental section which succeeds the following theory.

The noise is hence characterized by a flat response across all frequencies, making $N(f)$ a constant for all f - denoted ρ. An expression for $S(f)$ can be derived by standard Fourier analysis, allowing us to view the frequency and power domain representations of $s(t)$ and $n(t)$. These are shown in figure 2 and demonstrate the large low frequency bias of the training signal $s(t)$, relative to the noise signal $n(t)$, and the rapid flattening of the power spectrum for higher frequency terms. This all strengthens the notion that a low pass filter would be effective in removing noise in the training signal with minimal distortion of the true training signal sought, making $\tilde{s}(t)$ a good approximation of $s(t)$. Knowing the form of $S(f)$ and $N(f)$ then leads to the following definition of the optimal filter $\Phi(f)$:

$$\Phi(f,\tau,T) = \begin{cases} \frac{\tau^2}{\tau^2+\rho^2T^2} & : \; f = 0; \\ \frac{4\sin^2\left(\frac{\pi\tau f}{T}\right)}{4\sin^2\left(\frac{\pi\tau f}{T}\right)+\pi^2 f^2\rho^2} & : \; f \in \mathcal{Z}, \; f > 0; \\ 0 & : \; \text{otherwise.} \end{cases}$$

Since the training signal is effectively sampled every Δ time units, all signal frequencies above the Nyquist frequency $f_{\mathcal{N}}= 1/2\Delta$, are subject to aliasing. Noise power is hence constrained to lie between D.C. and $f_{\mathcal{N}}$. Signal amplitude above $f_{\mathcal{N}}$ is assumed negligible, since in general $a_1 \gg a_{f_{\mathcal{N}}}$[3].

With this explicit expression for $\Phi(f)$ it is now possible to generate images (impulse responses) of the optimal filter $\phi(t)$ for differing noise levels ρ - for the example case $\tau = 10\Delta, T = 100\Delta$. The first two diagrams in figure 3 depict the associated filters over two ranges of noise, in which the weightings at each time step are proportional to the weighting of the cells activation at this time in the past, denoted $\bar{\phi}(t)$. The level of noise clearly affects the form of the Wiener Filter[4]. In practice, as the neuron learns, its associated error rate will decrease, and with it, the form of the optimal linear filter. The change in error rate, from the neuron's initial state to fully trained, describes a series of steepening filters, moving from a more linear form at high noise values to a distinctly exponential form at lower noise values.

The large negative trough apparent in all of these filter profiles at time T/τ, comes as a direct consequence of fixing the ratio τ/T. Clearly, for 2τ steps in T, the response of the neuron τ time steps ago is opposite to what the current response should be, and hence the response at that time carries a great deal of information about the correct current response of the neuron. It is to be expected that this trough will disappear with variable length presentation times - as is confirmed in the next section.

2.3 Variable Length Presentation Times

In the previous section the optimal filter for the neural response was derived on the basis of fixed presentation lengths. In the 'real world' one might expect to see objects over variable periods of time. This leads to a natural alternative to the previous presentation scheme in which the presentation time τ is not fixed, but rather described by a decaying probability distribution. Since presentation time is a scale parameter, it is reasonable to propose that the presentation times are Jeffrey's distributed ($P(\tau) \propto 1/\tau$ or equivalently that log presentation times are uniformly distributed, see (Jaynes, 1983)), which implies that objects tend to be seen for short periods but are occasionally seen for much longer periods.

Given a certain probability density function $P(\tau)$ for the presentation length of each stimulus, the product of the filter shape, $\phi(t)$, and its associated density function, $P(\tau)$, can now be summed over all possible values of presentation length[5]. This yields the average Wiener Filter, $\langle\phi(t)\rangle$ for all τ.

In this section new optimal filters are derived under this new presentation length paradigm with two different values of maximum presentation length $\hat{\tau}$. In the first of these experiments $\hat{\tau}$ is set to be 100Δ, and the minimum presentation length to be Δ - i.e. the presentation time for one stimulus. In the second, the

[3] For example, if $T = 100\Delta$ and $\tau = 10\Delta$, $a_1 \approx 50 a_{f_{\mathcal{N}}}$.

[4] $\rho = 0.45$ is equivalent to an error rate of 10%, $\rho = 0.14$ to an error rate of 1%

[5] Summed rather than integrated since presentation lengths can only take integer values between the minimum and maximum presentation lengths.

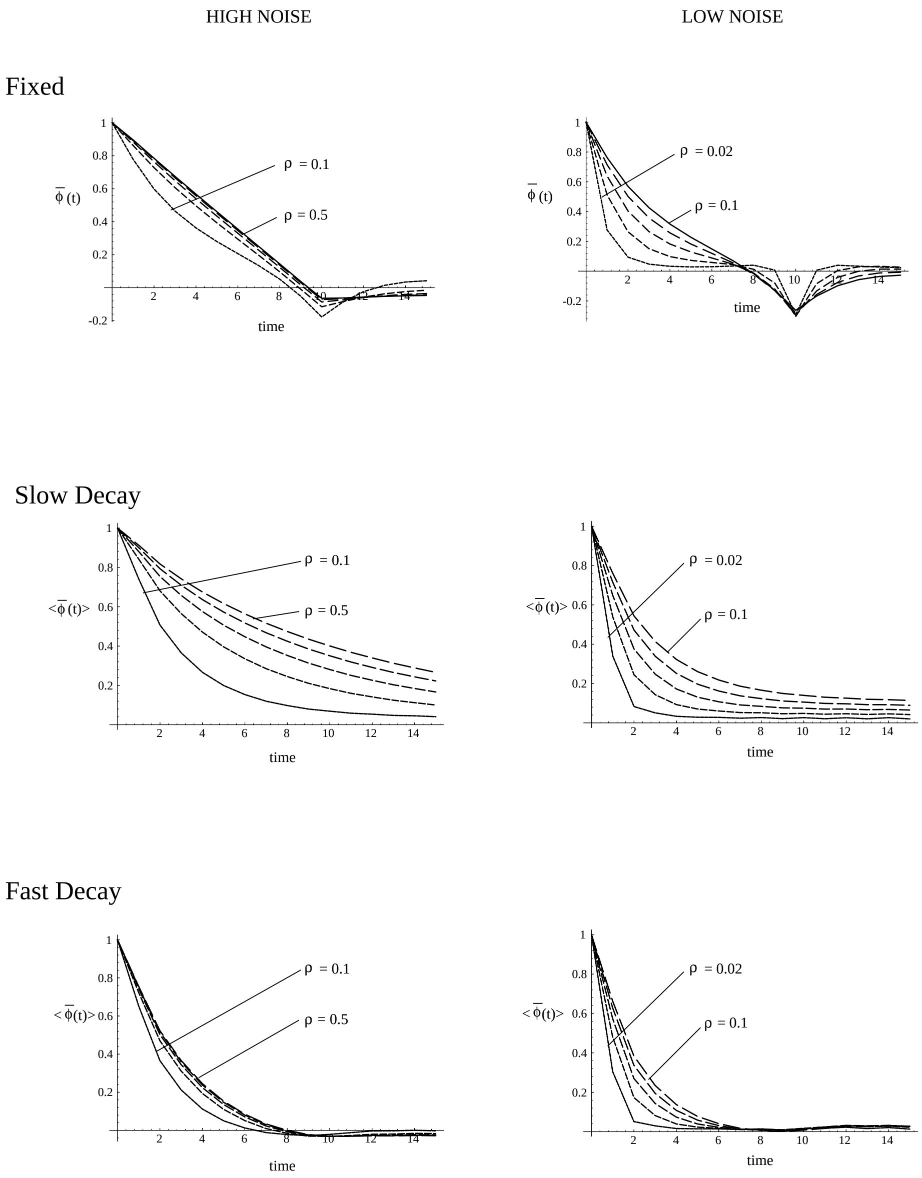

Figure 3: Normalized Wiener filters for the three presentation paradigms. The left hand graphs are all for higher noise values and the right hand graphs for a range of lower noise values. The vertical axis represents the weighting accorded to each sample of neural activity and the horizontal axis, the sample number in multiples of Δ time steps previous to the current sample at time zero.

presentation length probabilities will decay much more quickly with $\hat{\tau}$ being reduced to 10Δ.

Note that the period T, over which all stimuli are presented in any one training epoch is now distributed about a mean value rather than fixed, since the sub-period presentation lengths are of variable, random length. In these calculations the value of T is assumed reasonably constant and hence taken to be fixed at the expected, or mean, total presentation time for k distinct stimuli. Since it depends upon the maximum presentation time $\hat{\tau}$ it is now given the symbol $T_{\hat{\tau}}$.

Presentation length probabilities are defined by a simple Jeffrey's distribution of the form $1/\tau$ extending within the range $\Delta < \tau < \hat{\tau}$, such that $P(\tau)|_{\tau=\hat{\tau}} = 0$, as follows[6]:

$$P(\tau) = Z_{\hat{\tau}}^{-1}(\tau^{-1} - \hat{\tau}^{-1}) : \Delta \leq \tau \leq \hat{\tau}$$

$$Z_{\hat{\tau}} = \sum_{s=\Delta}^{\hat{\tau}} (s^{-1} - \hat{\tau}^{-1}) \ : \quad \begin{array}{l} Z_{100\Delta} = 4.18738\Delta \\ Z_{10\Delta} = 1.92897\Delta \end{array}$$

and

$$T_{\hat{\tau}} = k \sum_{\tau=\Delta}^{\hat{\tau}} \tau P(\tau) \ : \quad \begin{array}{l} T_{100\Delta} = 11.8212\Delta k \\ T_{10\Delta} = 2.33285\Delta k \end{array}$$

Hence, from the fore-running definition of $\phi(\tau, t)$:

$$\begin{aligned} \langle \phi(t) \rangle &= \sum_{\tau=\Delta}^{\hat{\tau}} P(\tau)\ \phi(t) \\ &= \sum_{\tau=\Delta}^{\hat{\tau}} \left(P(\tau) \sum_{f=0}^{\infty} \Phi(f, \tau, T_{\hat{\tau}}) \cos\left(\frac{2\pi f t}{T_{\hat{\tau}}}\right)\right) \end{aligned}$$

The lower four graphs of figure 3 show the various representations of the ideal filter through different noise values for the two different presentation probability distributions. As predicted in the previous section, switching to variable sequence presentation lengths sees the disappearance of the trough present in the fixed length presentation case. In addition, as one would expect, the optimal filter in the case of the shorter presentation times yields temporal filters with commensurately shorter time constants. Note that in all the cases the general form is of a decaying exponential, which will be important in formulating a local learning rule for implementing the filtering.

[6] We cannot use the pure τ^{-1} distribution since this is improper - it does not sum to one. We are therefore obliged to slightly modify this distribution to ensure that it can be used as a standard probability distribution.

2.4 The trace rule

The previous two sections have described the shape of the appropriate filters for transforming neural output into an improved training signal. This section describes a locally implementable learning rule which can realise the general form of all of the filters sought with the adaptation of a single parameter. The learning rule produces a running average of neural activity based upon a decaying exponential weighted average of the neuron's previous activity - of the type seen in the Wiener filter graphs in figure 3. The simple recursive form of the learning rule may also be important in that it lends itself to implementation locally within a cortical neuron (Wallis *et al.*, 1993).

The learning rule has a long history but was used most recently in invariant object recognition for orthogonal images (Földiák, 1991) and non-orthogonal images (Wallis *et al.*, 1993), and can be summarized as follows:

$$\Delta w_{ij}{}^{(t)} = \alpha \overline{y}_i{}^{(t)}.x_j \quad : \quad \sum_j w_{ij}{}^2 = 1 \; for \; each \; i^{th} \; neuron$$

and

$$\overline{y}_i{}^{(t)} = (1-\eta) y_i{}^{(t)} + \eta \overline{y}_i{}^{(t-1)}$$

where x_j is the j^{th} input to the neuron, y_i is the output of the i^{th} neuron, w_{ij} is the j^{th} weight on the i^{th} neuron, η governs the relative influence of the trace and the new input, and $\overline{y}_i{}^{(t)}$ represents the value of the i^{th} cell's trace at time t.

Using the above theory it is now possible to predict which values of η in the trace rule model will best fit the predicted optimal linear filter. The left most diagram of figure 4 represents the crucial predictive result for this section. The optimal value of the trace rule parameter η, is plotted as the joined line, and varies as a function of the noise ρ. The vertical bars show the relative quality of fit of the filter implemented by the trace rule to the optimal filter as a least mean square fit for the first 20 steps in the filter. Values for the actual least mean squared fitting error at each value of ρ are shown on the top of each bar. Clearly, the optimal value of η gradually drops as the noise level drops, but remains around 0.8 over a very wide range of noise. This firm value of η is crucially important because it suggests that a constant trace rule could be used throughout training, even as the classification error rate drops - a prediction which will be tested in section 3.

The remaining two graphs of figure 4 show the change in the best fitting value of η as a function of ρ for the variable presentation time cases. The general form of the graphs for the slow decay are similar to the fixed presentation time case, except that the optimal value of η is a little higher - between 0.8 and 0.9 - and the average fitting error (grey bars) is considerably smaller. This indicates that the training rule fits the optimal filter more closely under the probabilistic presentation regime. Note that the optimal value of η in the fast

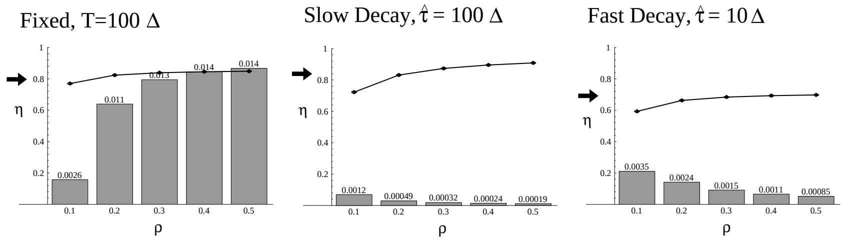

Figure 4: Graphs showing the change in the predicted optimal value of η with changing error rate, plus the least mean square fit error at each noise level. The thick arrows indicate the predicted optimal value of η over the range of noise values shown.

decay case is lower than for the other two cases lying as it does, slightly below $\eta = 0.7$. In practice it is likely that the with such short training sequences the final network performance will be worse than in the previous cases and hence the noise value will remain higher, thus $\eta = 0.7$ will probably produce better results than $\eta = 0.6$. The accuracy of these and previous predicted optimal values of η appear in the next section.

3 Simulating the predictions of the Wiener Filtering

3.1 Introduction

A series of predicted optimal linear filters, parameterized for stimulus presentation length and the classification error rate of the neurons ρ, have now been produced. In order to gauge the accuracy of the predictions a series of simulations were run using a range of filter shapes in accordance with the calculations of the previous section.

3.2 Methods

A two layer network was constructed - see figure 5. The first layer acts as a local feature extraction layer and consists of a 16x16 grid of neurons arranged in 4x4 inhibitory pools. Each pool fully samples a corresponding 4x4 patch of a 16x16 input image. Competition is only effective within these pools and is implemented by reducing all neurons within a pool with less than maximum firing rate, to one third of their initial rate. All learning in this layer is standard Hebbian. Above the input layer is a second layer consisting of a single inhibitory

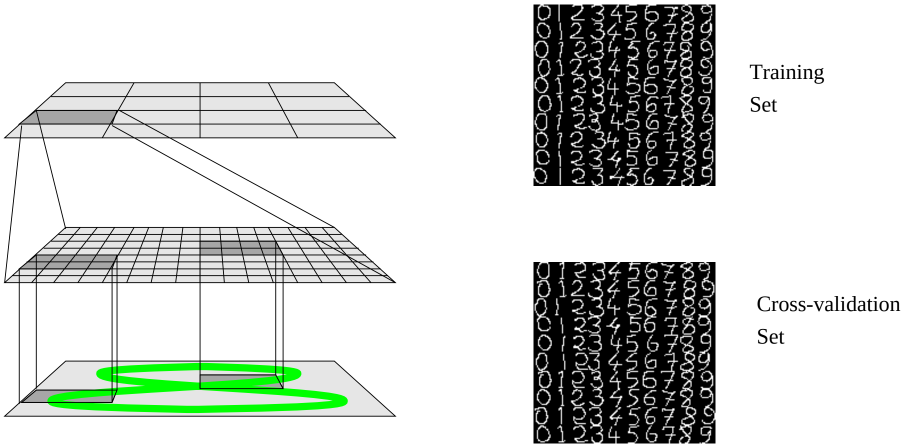

Figure 5: Architecture of the network used in the simulations and the two sets of digits used during training and testing.

pool of ten neurons which fully samples the first layer. The neurons in this layer are trained with the trace rule - as described in the following experiments. All neurons in both layers also have a separate, non-linear activation function which transforms the cell's calculated weighted input into an output firing rate. This was achieved by raising the activation value to a fixed power, namely 7, before scaling the outputs within each inhibitory pool to 1.

Digits from a character set used by LeCun *et. al.* (1989) were presented to the network. During learning, 100 digits - 10 of each type[7] - were presented in permuted random sequence according to the three stimulus presentation distributions described in the previous section. Cross validation testing was conducted by presenting a further 100 digits from the same database. Performance of the network was then gauged by the ability of neurons in the final layer to categorize digits from each class together, and further, to yield good categorization performance on the cross-validation data set. In addition, the noise measured on a neuron trained from start to steady response was calculated for all three presentation paradigms. From this the spectral form of the noise was calculated to confirm whether it is indeed ostensibly flat, as the forerunning calculations have assumed.

[7]Hence, k in the previous equations was 10

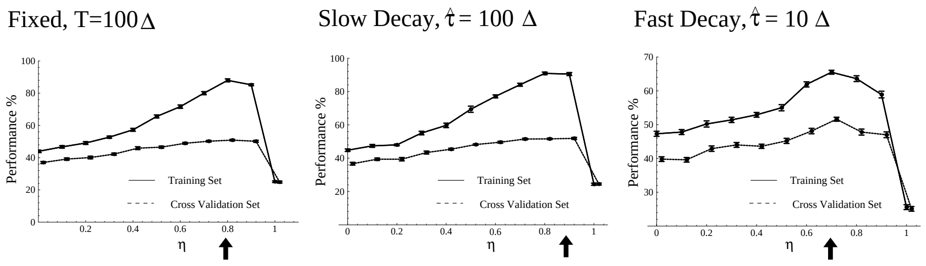

Figure 6: Classification performance for the network at differing values of η for the three presentation length paradigms. The thick arrows indicate the optimal value of η predicted by the earlier theory, which are seen to be in accord with the optimal values to come out of the simulations. The error bars correspond to the estimated standard error, which are in most cases vanishingly small. Note that in the fast decay case the graph only extends from 0% to 70% to improve clarity.

3.3 Results

The results shown in figure 6 depict the effect upon overall classification performance of varying the length of the trace, governed by the parameter η. The results are averaged over ten separate runs of the network. Under each of the three presentation paradigms the value of η predicted in the previous section is seen to correspond very closely to the peak in the performance graph displayed alongside - see the dark arrows in each of the six graphs. Therefore, despite two possible theoretical problems (the assumptions of independent errors and of linearity), the predictions from the theory are very close to the results of the simulations.

Note that for $\eta = 0$ the results correspond to simple Hebbian learning in the final layer. These results are clearly much worse than that those achieved with use of an optimal version of the trace rule. Further comparisons, with other architectures and other learning rules are described elsewhere (Wallis, 1995).

The exact calculations depend on the errors being approximateley random. Figure 7 shows the average power spectrum recorded for a neuron monitored from a random starting point to the point at which a steady state response had been achieved, for each of the three presentation paradigms. The spectrum is averaged across all 10 output cells and also across time, with the signals being binned into 100 consecutive stimulus presentations - the minimum time for the net to see all stimuli at least once. All three graphs show a peak in the low frequency end of the noise spectra - presumably due to the difference in false positive and false negative error rates discussed earlier. However, the power of this signal is considerably less than the signal power, and the spectrum is ostensibly flat for all higher frequencies up to and beyond the point at which

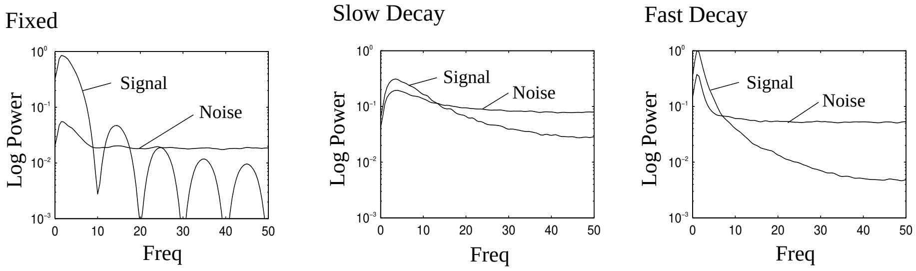

Figure 7: Log power spectrum for noise $n(t)$ and ideal training signal $s(t)$ in the response of a neuron under each of the three presentation length training paradigms. As can be seen, the general form reasonably approximates the graphs shown in Figure 2, with the noise being reasonably flat, and the signal having most of its power at low frequencies.

noise power exceeds signal power, in accordance with the form of the graph in figure 2.

4 Conclusions

The original proposal of this work was that temporal as well as spatial correlational structure existing in our environment could be used to construct representations of objects which are tolerant to large changes in view - such as are seen in the neurons of IT cortex. It was shown that if objects were not presented in random order, but presented as sequences of views of the same object, then this regularity could be used to help learn invariant representations.

If the distribution of presentation times was distributed with a Jeffreys distribution (the distribution of log presentation times was uniform), then the optimal linear filter was well approximated by an exponentially weighted average of the previous responses. This is easily implemented using the so called trace rule.

There are a number of reasons such an analysis may not be appropriate. The errors made by a neuron may be far from random, and the mathematics assumes that the system is linear whereas classification requires non-linearity. Despite this, when the theory was applied to a non trivial classification task, the system worked and the optimal parameters were very well predicted from the theory.

References

Bülthoff, H., & Edelman, S. 1992. Psychophysical support for a two–dimensional view interpolation theory of object recognition. *Pages 60–64 of: Proceedings* of the National Academy of Science, USA*, vol. 92.

Desimone, R. 1991. Face–selective cells in the temporal cortex of monkeys. *Journal of Cognitive Neuroscience*, **3**, 1–8.

Edelman, S., & Weinshall, D. 1991. A self–organising multiple–view representation of 3D objects. *Biological Cybernetics*, **64**, 209–219.

Földiák, P. 1991. Learning invariance from transformation sequences. *Neural Computation*, **3**, 194–200.

Jaynes, E.T. 1983. *Papers on Probability, Statistics and Statistical Physics.* Synthese library. Reidel, Dordrecht.

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., & Jackel, L.D. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, **1**, 541–551.

Miyashita, Y. 1988. Neuronal correlate of visual associative long–term memory in the primate temporal cortex. *Nature*, **335**, 817–820.

Poggio, T., & Edelman, S. 1990. A network that learns to recognize three-dimensional objects. *Nature*, **343**, 263–266.

Press, W.H. 1992. *Numerical Recipes in C : The Art of Scientific Computing.* Cambridge: Cambridge University Press.

Rolls, E.T. 1992. Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical areas. *Philosophical Transactions of the Royal Society, London [B]*, **335**, 11–21.

Tanaka, K., Saito, H., Fukada, Y., & Moriya, M. 1991. Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology*, **66**, 170–189.

Tarr, M.J., & Pinker, S. 1989. Mental rotation and orientation–dependence in shape recognition. *Cognitive Psycholgy*, **21**, 233–282.

Wallis, G. 1995. Using spatio-temporal corrolations to learn invariant object recognition. *To appear in Neural Networks.* www: ftp://ftp.mpik-tueb.mpg.de/pub/guy/nn.ps.Z.

Wallis, G., & Rolls, E.T. 1995. A Model of Invariant Object Recognition in The Visual System. *Submitted to Journal of Computational Neuroscience for Review.* www: ftp://ftp.mpik-tueb.mpg.de/pub/guy/jcns7.ps.Z.

Wallis, G., Rolls, E.T., & Földiák, P. 1993. Learning invariant responses to the natural transformations of objects. *Pages 1087–1090 of: International Joint Conference on Neural Networks*, vol. 2.

Wiener, N. 1949. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series : with Engineering Applications*. New York: John Wiley & Sons.