Reduced Representation by Neural Networks with Restricted Receptive Fields

Marco Idiart
Barry Berk
L. F. Abbott
Center for Complex Systems, Brandeis University, Waltham, MA 02254 USA

Model neural networks can perform dimensional reductions of input data sets using correlation-based learning rules to adjust their weights. Simple Hebbian learning rules lead to an optimal reduction at the single unit level but result in highly redundant network representations. More complex rules designed to reduce or remove this redundancy can develop optimal principal component representations, but they are not very compelling from a biological perspective. Neurons in biological networks have restricted receptive fields limiting their access to the input data space. We find that, within this restricted receptive field architecture, simple correlation-based learning rules can produce surprisingly efficient reduced representations. When noise is present, the size of the receptive fields can be optimally tuned to maximize the accuracy of reconstructions of input data from a reduced representation.

1 Introduction

Hebbian learning rules commonly used in model neural networks are closely related to principal component techniques for data reduction (Oja 1982; Linsker 1988; Hertz *et al.* 1991). Principal component analysis is a standard method for reducing the dimension of data sets by projecting onto coordinate axes that are eigenvectors of the correlation matrix. In a linear network, weights that develop according to an appropriately constrained correlation-based Hebbian learning rule will project input data onto the principal component axis with the largest eigenvalue (Oja 1982; Linsker 1988; Hertz *et al.* 1991; Miller and MacKay 1994). This suggests that networks of linear units with Hebbian learning rules might develop efficient reduced representations of high-dimensional data sets automatically, without supervision. However, when more than one unit is involved in developing such a representation, a problem arises. Acting independently, correlation-based learning on each unit will find the same maximal principal component axis and therefore each unit will provide the same information. Although the representation for each unit by

itself is optimal, collectively the network produces a representation that is highly redundant and very far from optimal.

Various solutions have been proposed for this dilemma. Using more complex learning schemes (Oja 1989; Sanger 1989; Foldiak 1989; Fyfe 1993; Linsker 1993) it is possible to construct multidimensional principal component representations. However, none of these schemes is related convincingly to known mechanisms in biological networks. Instead, it appears that biological systems may prevent redundancy by providing different neurons with different views of the input space. This is the result of restricted receptive fields, that is, not all inputs are connected to all network units or, equivalently, some weights are permanently set to zero. If different units have different nonzero weights, each will have access to a different subspace of the full input space. A simple correlation-based learning rule applied to these restricted weights will find the principal component of the input data set in a subspace that is different for each unit. Thus, the network units will not all develop the same representation of the input data and redundancy is reduced. However, the reduction of redundancy has a price. The directions along which the network units project the input vector will no longer be the optimal ones because the individual units do not have access to the full input correlation matrix. Nevertheless, as we will see, finite receptive fields provide an effective solution to the problem of building optimal nonredundant reduced representations using simple correlation-based learning rules.

Minimizing the redundancy between units is an effective strategy for building efficient representations (Barlow *et al.* 1989; Atick and Redlich 1990). However, redundancy can be useful in the presence of noise because it allows averaging of the noisy signal. We will show that receptive field sizes can be adjusted to optimize the balance between averaging and redundancy for a particular noise level.

2 Network Architecture and Data Reconstruction Method _____

We consider a simple, single-layer feedforward network with D inputs fed into N units. We do not include horizontal interactions between units. The D inputs form the coordinates of points in the data set being represented by the network output. These D inputs are coupled to the N linear network elements through a matrix of weights. The weight coupling input coordinate x_i to network unit a is W_{ai}. If we use a vector notation for the i index, the couplings of unit a to all the inputs can be represented by $\mathbf{W}_a$. The response of unit a, denoted by y_a, is given by

$$y_a = \sum_{i=1}^{D} W_{ai}x_i + \eta_a = \mathbf{W}_a \cdot \mathbf{x} + \eta_a \tag{2.1}$$

where η_a represents an uncorrelated noise term that has zero mean and standard deviation σ.

To admit a reduced representation, a data set must have certain characteristics that are best expressed in terms of its correlation matrix

$$C_{ij} = \langle x_i x_j \rangle \tag{2.2}$$

We label the eigenvectors of the correlation matrix by X_a and the corresponding eigenvalues by λ_a. We number the eigenvectors and eigenvalues according to the size of the eigenvalue so that X_1 is the eigenvector with the largest eigenvalue, λ_1, X_2 has the next largest eigenvalue, and so forth. The eigenvectors are normalized to have unit length. For a data set to be reducible, a subset of the correlation matrix eigenvalues must be significantly larger than all the others. This allows the data points to be represented by and reconstructed from a reduced representation. Note that we are considering a rather specialized data set, one that lends itself particularly well to reduced representation because there is a gap between large and small eigenvalues. While restrictive, this case is ideal for illustrating the role of finite receptive fields.

The output of the network consists of the values y_a with $a = 1, 2, \ldots, N$ of the N network elements. These outputs form the reduced representation of the input data. To make use of this reduced representation, and to evaluate its accuracy, we must have a means of reconstructing the full D-dimensional data points from these N outputs. This reconstruction is not something done by the network itself but rather performed either by downstream networks or, in this case, by us to evaluate the quality of the representation. We do the reconstruction by computing an optimal linear estimate of the input vector x,

$$x_{\text{est}} = \sum_{a=1}^{N} y_a D_a \tag{2.3}$$

with appropriately chosen D_a. The accuracy of the reconstruction will be measured by defining the average normalized reconstruction error as

$$E = \frac{\langle |x - x_{\text{est}}|^2 \rangle}{\langle |x|^2 \rangle} \tag{2.4}$$

where the notation $\langle \ \rangle$ indicates an average over the input data set. This error is minimized by choosing (Salinas and Abbott 1994; Sanger 1994)

$$D_a = \sum_{b=1}^{N} \left[Q^{-1} \right]_{ab} L_b \tag{2.5}$$

where $Q_{ab} = \langle y_a y_b \rangle$ and $L_a = \langle x y_a \rangle$. With this choice, the average error is

$$E = 1 - \frac{\sum_{a=1}^{N} D_a \cdot L_a}{\sum_{a=1}^{D} \lambda_a} \tag{2.6}$$

To illustrate the use of equation 2.6 we will examine the accuracy of two different networks. The most redundant representation results when all the network units are coupled to all the inputs and all use the same independent correlation-based learning rule to construct their input weights. In this case, each unit develops the same set of weights corresponding to the principal component axis with the largest eigenvalue so

$$\mathbf{W}_a = \mathbf{X}_1 \tag{2.7}$$

for all a. In this case, we find from 2.6, that the error is

$$E = 1 - \frac{N\lambda_1^2}{(N\lambda_1 + \sigma^2)\sum\limits_{a=1}^{D}\lambda_a} \tag{2.8}$$

This approaches a finite limit as $N \to \infty$ due to the redundancy of the representation. For high levels of noise (large σ), the factor of N in the numerator of the second term indicates increasing accuracy for large networks due to signal averaging.

As a second example we consider the N-dimensional optimal reduced representation of a network that uses the N principal component eigenvectors with the largest eigenvalues as weight vectors. In our notation this means that

$$\mathbf{W}_a = \mathbf{X}_a \tag{2.9}$$

for $a = 1, 2, \ldots, N$. The corresponding average error is

$$E = 1 - \frac{\sum\limits_{a=1}^{N}\lambda_a^2/(\lambda_a + \sigma^2)}{\sum\limits_{a=1}^{D}\lambda_a} \tag{2.10}$$

The error depends on the percentage of the trace of the correlation matrix represented by the N largest eigenvalues. This is the optimal reduced representation but note that for high noise levels it produces a larger error than the highly redundant network of equation 2.8.

3 Restricted Receptive Fields

Ideally, we would like to combine the best features of the two different reduced representations that we have discussed. In other words, we would like to construct a reduced representation that requires only a simple correlation-based learning rule but that provides a nonredundant reduced representation. Restricted receptive fields provide one way of doing this. If the receptive fields of the network units are restricted as in Figure 1, the weights constructed by a simple correlation-based learning
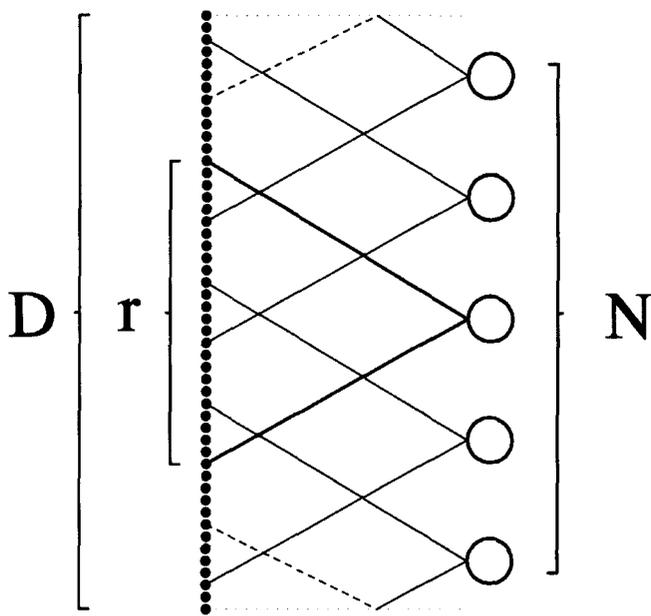
Figure 1: The architecture of the networks under study. In this example, $D = 50$ inputs represented by small filled circles drive $N = 5$ network units denoted by large unfilled circles. Receptive fields are restricted to those inputs lying between the two straight lines diverging from each unit. In this case each unit is coupled to $r = 26$ inputs. At the edges of the input array we impose periodic boundary conditions as indicated by the dashed receptive field lines.

rule will not be equal to the maximal eigenvector of the correlation matrix. Instead, they will be the eigenvectors with maximum eigenvalue of submatrices of the full correlation matrix.

We construct the restricted receptive fields as shown in Figure 1. The D inputs are divided into N subgroups consisting of r elements. At the edges of the input array we impose periodic boundary conditions. Because they will be important in the discussion of our results, we will review here the definition of several symbols:

D = the number of inputs to the network.

N = the number of network units.

d = the number of "large" eigenvalues of the input correlation matrix.

r = the number of inputs connected to each network unit.

A restricted receptive field means that some of the weights for each unit are forced to take the value zero. It is convenient to define a factor Z_{ai} that is zero if W_{ai} is forced to be zero and one if it is not. A simple correlation-based learning rule applied to these restricted weights will construct the maximal eigenvectors of a submatrix of the full correlation matrix that is different for each network unit. For unit a, this submatrix is

$$\tilde{C}_{ij}^a = Z_{ai}C_{ij}Z_{aj} \tag{3.1}$$

The average representation error for a network with finite receptive fields can be computed from equation 2.6, although we must resort to numerical techniques to perform the calculation. We begin by constructing an appropriate correlation matrix corresponding to a reducible data set. We randomly choose orthonormal eigenvectors $\mathbf{X}_a$ with eigenvalues λ_a and write

$$C_{ij} = \sum_a \lambda_a X_{ai} X_{aj} \tag{3.2}$$

We then partition the D inputs onto the N network units, r at a time to define the elements Z_{ai}. Submatrices are computed from equation 3.1 and the weight vector $\mathbf{W}_a$ for each unit is set equal to the eigenvector of the corresponding submatrix with the largest eigenvalue. This computation is done numerically. From the resulting weights $\mathbf{W}_a$ we determine the vectors $\mathbf{L}_a$ and $\mathbf{D}_a$ and insert these into formula 2.6 to obtain the average error. The entire procedure is repeated several times to get a good statistical sample.

4 Results

The accuracy of a reconstruction from a reduced representation depends most strongly on the percentage that the large set of eigenvalues contributes to the trace of the correlation matrix. We found that our results were not very sensitive to the values of D or d provided that this percentage was held fixed. Therefore, we show a representative case, $D = 50$ and $d = 5$, in the figures. Figure 2A shows our results for the average normalized error (2.6) as a function of the size of the receptive field. For small receptive fields, little information is extracted from the data by each unit so the reconstruction is inefficient and the reconstruction error is large. When we increase the size of the receptive field, the error decreases rapidly to a plateau. The beginning of the plateau represents a critical receptive field size beyond which, at least on average, network performance does not improve appreciably. The critical value is near the point where different network units begin to have common inputs and the receptive fields start to overlap.

As the size of the receptive field grows, each individual unit can construct a better projection axis. However, this also causes the units to have
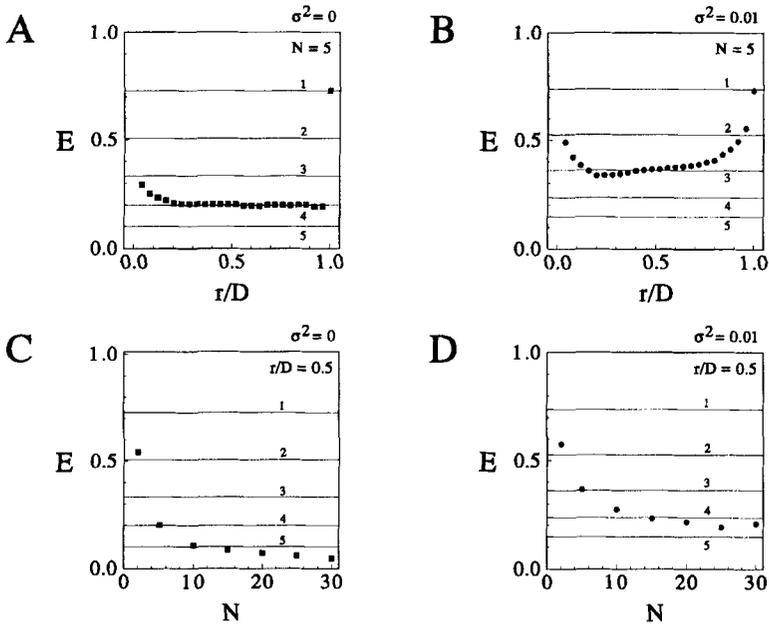
Figure 2: Results for the average reconstruction error for networks with restricted receptive fields. $D = 50$ and the five largest eigenvalues of the correlation matrix are $\lambda_1 = 0.27$, $\lambda_2 = 0.22$, $\lambda_3 = 0.18$, $\lambda_4 = 0.13$, and $\lambda_5 = 0.1$. The trace of the correlation matrix is normalized to one and the sum of the five principal eigenvalues totals 90% of the trace. The data points show the results for networks with restricted receptive fields while the solid lines indicate the performance of fully connected networks using multiple principal components with $1, 2, \ldots, 5$ units in decreasing order of error. (A) The reconstruction error as a function of receptive fields size in the absence of noise, $\sigma^2 = 0$ with $N = 5$. (B) The reconstruction error as a function of receptive fields size with $\sigma^2 = 0.01$ and $N = 5$. (C) The reconstruction error as a function of network size with $r/D = 0.5$ and $\sigma^2 = 0$. (D) The reconstruction error as a function of network size with $r/D = 0.5$ and $\sigma^2 = 0.01$.

increasingly similar inputs and the increased redundancy cancels this improvement. As a result the performance of the full system remains fairly constant. Ultimately, when the receptive field size goes to $r = D$, the system is equivalent to a fully coupled, maximally redundant network that uses a single eigenvector. This transition is, however, discontinuous as seen in Figure 2A. As long as there is any difference between the representations constructed by the different units, the optimal reconstruction

technique can provide a fairly accurate reconstruction. Figure 2A also shows the reconstruction errors for multiple principal component representations of various sizes. Although not as accurate as a principal component reduction of the same size, the network with finite receptive fields performs fairly well.

The ability of the optimal reconstruction technique to exploit small differences in network unit outputs is limited if there is noise in the system. This is seen in Figure 2B. The impact of the noise is more pronounced for large receptive field sizes and the plateau is no longer as flat as it was without noise in Figure 2A. In the example shown, the optimal performance occurs near the point where the receptive fields just begin to overlap.

How efficient is the restricted receptive field architecture compared to the optimal multiple principal component method? Figure 2C shows that, without noise, it takes about 10 units in a network with restricted receptive fields ($r/D = 0.5$) to equal the performance of an optimal network with 5 units. When noise is present, the approach to the optimal performance is only asymptotic because adding more noisy units increases the total network noise level. This is shown in Figure 2D.

In Figure 2B, the error rises for large receptive field sizes when the network configurations are highly redundant. This is because the small differences between network unit responses are swamped by the noise. However, when the noise level is high, redundancy can be advantageous. With high noise levels, the best strategy may be to project onto a small number of directions and to cover these directions with multiple units to average out the noise. Thus, for higher noise levels the receptive field size that produces the minimum error increases as seen in Figure 3. For low noise the optimal receptive field size is near the value where the fields just begin to overlap. However, as the noise increases, the optimal receptive field size grows and approaches the limit of full redundancy with all units receiving the same inputs. This indicates that the receptive field size can be tuned to a value that is optimal for a given level of noise.

Figure 4 shows the reconstruction error for a network with finite receptive fields as a function of two characteristics of the distribution of input data. In Figures 4A and B, the error is shown as a function of the eccentricity of the data set defined as the fraction of the trace of the correlation matrix carried by the d largest eigenvalues,

$$\epsilon = \frac{\sum\limits_{a=1}^{d} \lambda_a}{\sum\limits_{a=1}^{D} \lambda_a} \tag{4.1}$$

The error for both a network using conventional principal components and one using finite receptive fields drops as the eccentricity increases to the maximum value of one.
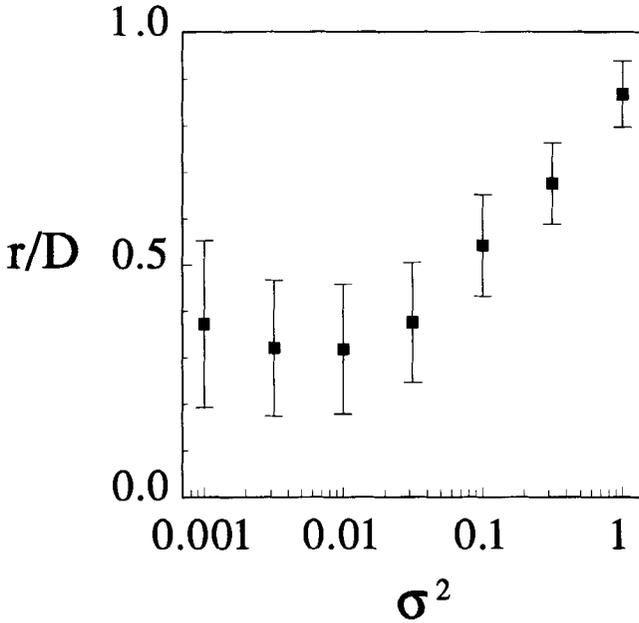
Figure 3: The optimal receptive field size as a function of the noise level. For each value of σ^2 the receptive field size r giving the minimum reconstruction error was determined. The results shown are for $D = 50$ and $d = N = 5$. The error bars indicate the standard deviations over different trials. For low noise levels, the optimal receptive field size is small, producing a modest amount of overlap between neighboring fields. For large noise levels, the optimal receptive field size grows.

Figure 4C and D shows the same network reconstruction error as a function of a parameter that controls how close the d largest eigenvalues are to each other. Specifically, we have taken

$$\lambda_k \propto [100 - (k - 1)p]^2 \tag{4.2}$$

with p as a parameter and the constant of proportionality determined by the value of ϵ. The particular expression used here is arbitrary and was chosen to illustrate the effect. Interestingly, the principal component networks and the finite receptive field network show different dependencies. For the latter, the error is dominated by the ability of this architecture to find the largest eigenvalues and is not very sensitive to their distribution.
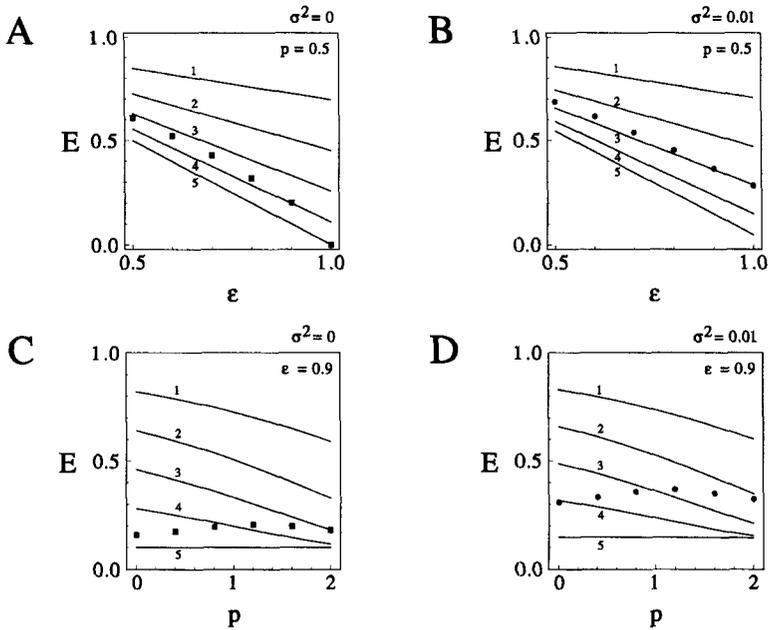
Figure 4: Network performance for different data structures. The correlation matrix is parameterized by two variables, ϵ and p, defined in the text, that characterize the eccentricity and variability of large eigenvalues for the input data set. (A,C) The reconstruction error in the absence of noise, $\sigma^2 = 0$; (B,D) $\sigma^2 = 0.01$. In (A) and (B), $p = 0.5$ and in (C) and (D) $\epsilon = 0.9$. The data points show the error for a network with $D = 50$, $N = 5$, and $r/D = 0.5$ and the solid lines indicate the performance for networks using a multiple principal component algorithm with $1, 2, \ldots, 5$ output units.

5 Discussion

Our results indicate that restricted receptive fields provide an effective way of building nonredundant reduced representations. The network shown in Figure 2A performs with $N = 5$ elements as well as an optimal $N = 4$ network. In Figure 2C, a restricted field network of 10 elements performs as well as an optimal network with 5 elements. When noise is present, the receptive field size should be tuned to the level of noise to minimize reconstruction errors. It would be interesting to see if receptive field sizes in biological systems are adjusted in this way to produce optimal results.

Acknowledgments

References

Atick, J., and Redlich, N. 1990. Towards a theory of early visual processing. *Neural Comp.* **2**, 308–320.

Barlow, H. B., Kaushal, T. P., and Mitchison, G. J. 1989. Finding minimum entropy codes. *Neural Comp.* **1**, 412–423.

Foldiak, P. 1989. Adaptive network for optimal linear feature extraction. In *Proceedings of the IEEE/INNS International Joint Conference on Neural Networks*, pp. 401–405. IEEE Press, New York.

Fyfe, C. 1993. PCA properties of interneurons. In *Proceedings of the International Conference on Artificial Neural Networks*, S. Gielen and B. Kappen, eds., pp. 183–188. Springer-Verlag, London.

Hertz, J., Krogh, A., and Palmer, R. G. 1991. *Introduction to the Theory of Neural Computation*. Addison-Wesley, New York.

Linsker, R. 1988. Self-organization in a perceptual network. *Computer* **21**, 105–117.

Linsker, R. 1993. Local synaptic learning rules suffice to maximize mutual information in a linear network. *Neural Comp.* **4**, 691–702.

Miller, K. D., and MacKay, D. J. C. 1994. The role of constraints in Hebbian learning. *Neural Comp.* **6**, 100–126.

Oja, E. 1982. A simplified neuron model as a principal component analyzer. *J. Math. Biol.* **15**, 267–273.

Oja, E. 1989. Neural networks, principal components and subspaces. *Intl. J. Neural Syst.* **21**, 61–68.

Salinas, E., and Abbott, L. F. 1994. Reconstruction of vectors from firing rates. *J. Comp. Neurosci.* **1**, 89–107.

Sanger, T. D. 1989. Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks* **2**, 459–473.

Sanger, T. D. 1994. Theoretical considerations for the analysis of population coding in motor cortex. *Neural Comp.* **6**, 29–37.