Hinton, G. E., and Sejnowski, T. J., Unsupervised learning: Foundations of neural computation In: G. E. Hinton and T. J. Sejnowski (Eds.) Unsupervised learning: Foundation computation, MIT Press, Cambridge, MA, vii-xv (1999).

# Introduction

Consider the problem of getting a neural network to associate an appropriate response with an image sequence. The obvious approach is to use supervised training. If the network has around $10^{14}$ parameters and only lives for around $10^9$ seconds, the supervision signal had better contain at least $10^5$ bits per second to make use of the capacity of the synapses. It is not immediately obvious where such a rich supervision signal could come from.

A more promising approach depends on the observation that images are not random but are generated by physical processes of limited complexity and that the appropriate response to an image nearly always depends on the physical causes of the image rather than the pixel intensities. This suggests that an unsupervised learning process should be used to solve the difficult problem of extracting the underlying causes, and decisions about responses can be left to a separate learning algorithm that takes the underlying causes rather than the saw sensory data as its inputs. Unsupervised learning can usually be viewed as a method of modeling the probability density of the inputs, so the rich sensory input itself can provide the $10^5$ bits per second of constraint that is required to amke use of the capacity of the synapses.

The papers in this collection provide a sample of research on unsupervised learning. Some areas and important contributions are not represented either because an appropriate paper did not appear in *Neural Computation* or because of the limited space that was available. One entire area of research in unsupervised learning, self-organizing map formation, will appear as a separate volume in this series. Despite these limitations, the wide range of approaches that is included here serves as a guide to the development of the field of unsupervised learning.

## Redundancy Reduction

One of the earliest formulations of unsupervised learning in the context of vision was the concept of redundancy reduction (Attneave 1954; Barlow 1959; Barlow 1989). The goal was to find ways to compress the information contained in images, a goal that was also pursued in the commercial arena to reduce the bandwidth needed to transmit images. In the case of the human visual system, information in the array of photoreceptors in the retina, which number around 100 million, is compressed and represented by spike trains in around 1 million ganglion cells whose axons form the optic nerve. Atick and Redlich (1993) used an entropy reduction measure to show that the center-surround receptive fields found in ganglion cells are

optimal when the mapping is linear and the redundant information in the second-order correlations is removed. This was achieved by having lateral inhibition between neighboring cells.

Linsker (1986) had earlier shown that the same center-surround geometry for the receptive fields could be obtained in the context of a feedforward neural network that used a Hebbian form of synaptic plasticity. This approach, which he called "infomax," used a simple unsupervised Hebbian learning algorithm in the presence of noise on the inputs of the network and converged to the connection strengths needed for the center-surround geometry. It is unlikely that this learning mechanism is actually used in the retina, but it demonstrates that the response properties of neurons can be achieved with local learning rules and transmit visual images in an optimal way, in the sense defined by Atick and Redlich (1993).

Hebbian learning at a synapse depends jointly on the activity of the presynaptic neuron and the postsynaptic neuron. It is a biologically plausible leaning rule because it depends only on signals that are locally available at the synapse. Forms of Hebbian plasticity have been found in the hippocampus (Brown, Kairiss, and Keenan 1990) and the neocortex (Markram, Lubke, Frotscher, and Sakmann 1997).

Hebbian synapses are sensitive to information contained in the second-order correlations of the inputs. Zhang et al. (1993) demonstrate that properties of motion selective cells in the visual cortex, far removed from the retina, can also be understood using Hebbian synaptic plasticity in a feedforward network. The development of the visual cortex can also be modeled by a network with Hebbian plasticity (Miller, Keller and Stryker 1989; Obermayer and Sejnowski 1998).

## Maximizing Mutual Information

There are many possible objective functions for unsupervised learning, each of which can be optimized to produce a representation that is particularly good at achieving some goal. In the case of reducing redundancy this means eliminating correlations and producing a compact code for the input. Linsker (1992) showed that a form of Hebbian learning was able to maximize the mutual information between the inputs and the outputs of a feedforward network in the presence of noise. The learning algorithm has two phases for learning, with the inputs present during the first phase of Hebbian learning and absent during the second, anti-Hebbian phase in which the sign of earning is reversed. The second phase is needed to calibrate the correlations that are induced in the output units by the intrinsic noise on the inputs. Sphased leraning algorithms were introduced by Crick and Mitchison (1983) and Hopfield, Feinstein, and Palmer (1983) for recurrent attractor networks that stored memories as stable states of the network, and for Boltzmann machines, which have hidden units to learn the higher-order structure of the inputs (Hinton and Sejnowski 1986).

Another way to detect higher-order structure of the inputs is to have a multilayer network and to introduce an objective function that measures coherence between different output units. Becker and Hinton (1993) introduced an objective function for maximizing the information that parameters extracted from different parts of an extended sensory input convey about some common underlying cause. The model used columns of feedforward networks and was able to detect disparity as an underlying cause from sterograms presented on the inputs. The learning algorithm, however, was quite complex and required the propagation of global information in the network to optimize the objective function.

In a movie, the temporal sequence of images is correlated and additional information can be extracted by lookking for temporal as well as spatial structure. For example, a movie of a rigidly moving object contains highly redundant information because the image of the object will appear in slightly different spatial locations on successive frames of the movie. Földiák (1991) showed how this translation invariance can be captured in simple feedforward network that used Hebbian synapses and an output layer of units with a short-term memory of previous inputs. The network was trained with moving lines and the response properties of neurons in the network were similar to those found in the visual cortex. This principle was generalized by Stone (1996), who applied it to learning stereo disparity from dynamic stereograms.

## Independent Component Analysis

Neurons in the visual cortex have receptive fields that are compact and elongated, so that the best stimulus is often a thin bar or edge of light (Hubel and Wiesel 1968). The simple cells have separate excitatory and inhibitory subregions. Barlow conjectured that these neurons formed feature detectors that were maximally independent over the ensemble of natural images. Independence is a much stronger property than second-order decorrelation since independence entails that all higher-order correlations between pixels in the ensemble of images must also be zero. Field (1994) recognized that the output values of the feature detectors ought to have a sparse distribution with high kurtosis (many values near zero and a few quite high values). Field and Olshausen (1994) showed that an unsupervised learning algorithm that maximized sparseness produced visual feature detectors that resembled those found in the visual cortex.

Unsupervised learning algorithms have recently been found developed for finding independent components in linear mixtures and blind signal separation (Comon 1994). An efficient algorithm for Independent Component Analysis (ICA) of super-Gaussian signals was derived by Bell and Sejnowski (1995) from Linsker's infomax principle. When applied to natural images, ICA developed localized edge and and bar detectors similar to the simple cells that are found in the visual cortex (Bell and Sejnowski 1997). Amari

(1997) provided an improvement for speeding up this learning algorithm, and Lee, Girolami, and Sejnowski (1999) have extended it to sub-Gaussian sources. Other fast ICA algorithms have also been developed (Hyrarinen and Oja 1997). One of the advantages of these algorithms is that they are able to separate non-Gaussian sources, which are common in auditory and visual signal processing. One of the limitations is that the source model is linear and is not capable of adequately representing visual images with occlusion and other structured properties.

### Clustering and Dimensionality Reduction

The goal of clustering is to group together similar inputs. Many algorithms exist for clustering in low-dimensional spaces, but the problem becomes more difficult as the dimensionality increases. Platt (1991) introduced an on-line clustering method that progressively adjusts the prototypes for each new input and adds a new prototype when none of the existing ones is sufficiently close by. This is an example of a constructive learning algorithm that adds resources sequentially as needed during the learning (Fahlman and Lebiere 1990). This approach has the advantage of starting out with relatively few parameters at the outset and avoids overfitting by increasing the number of parameters only when justified by additional inputs.

The distance measures used in most clustering algorithms are typically simple ones such as the Euclidean distance. However, two objects may be similar despite differences in position, orientation, and scale. Clustering with a graph-matching distance measure that incorporates these invariances, and which is also insensitive to permutation and missing data, was introduced by Gold, Rangarajan, and Mjolsness (1996). Their method is computationally efficient and scales well with the dimensionality of the problem.

A standard engineering technique for reducing the dimensionality of the input is to use Principal Component Analysis (PCA). This is computationally efficient, but it suffers from the limitation that the representation in the principal components is linearly related to the original input. A number of different research groups realized that this limitation of PCA can be overcome by combining it with clustering. The idea is to divide the data into clusters in such a way that the points in each cluster lie close to a plane. This can be done by using a simplified version of the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977), which alternates between an "E-step" that assigns each datapoint to the closest plane and an "M-step" that refits each plane to all the datapoints that have been assigned to it. Kambhatla and Leen (1997) demonstrate that this piecewise linear approach can be quite effective in modeling nonlinear manifolds. If high-dimensional data is projected onto a randomly oriented line it nearly always has an approximately Gaussian distribution. This suggests that directions in the input space that yield non-Gaussian distributions are interesting and

that projections onto these directions would constitute interesting features. Intrator (1992) shows how neurons can discover such directions using a biologically plausible local algorithm called the BCM rule (Bienenstock, Cooper, and Munro, 1982).

## Learning Probability Distributions

During supervised learning, each input vector comes with an associated desired output that is supplied by the teacher. Unsupervised learning is often characterized as supervised lerning with an unknown output. This makes it very hard to decide what counts as success and suggests that the central problem is to find a suitable objective function that can replace the goal of agreeing with the teacher. Many apparently different objectives have been proposed:

Discover clusters in the data.

Discover a mapping from the observed data to a set of pairwise decorrelated or statistically independent features.

Discover temporal or spatial invariances in the data by getting modules to agree with each other.

Discover unlikely coincidences of events whose joint probability far exceeds the product of their individual probabilities.

Discover highly non-Gaussian projections of the data (projection pursuit).

It is less natural, but much more revealing, to view unsupervised learning as supervised learning in which the observed data is the *output* and for which there is no input. This makes it obvious that the model that generates the output must either be stochastic or must have an unknown and varying input in order to avoid producing the same output every time. Given this view, the obvious aim of unsupervised learning is to fit a generative model that gives high likelihood to the observed data.[1] This objective is often accompanied by the hope that the natural causes of the data will come to be represented by the activities of the hidden units.

The introduction of hidden processing units allows a network to represent a larger class of nonlinear functions using latent variables and to extract more complex nonlinear structure from the inputs with unsupervised learning. The Boltzmann machine is a recurrent network of stochastic units

[1] The generative model approach is closely related to the idea of finding an efficient code from which the data can be reconstructed, because any model that assigns high probability density to the data can be used for efficient data-compression in which the number of bits required to communicate a data vector approaches the negative log probability of the vector under the generative model.

with symmetric connections between them (Hinton and Sejnowski 1986) for which there is a Hebbian learning algorithm that reduces the Kullback-Liebler distance between the probability distribution of the inputs and that generated by the free-running network. The Boltzmann machine is a generative model in the sense that after learning is complete, the free-running network generates patterns on the input units with the same probability distribution that occurred during the training phase.

A major advantage of the generative approach is that it cleanly separates inference and learning. The inference process is given an observed data vector and a generative model, and assuming that the data came from the model, it computes the posterior probability distribution across the hidden states of the model, or an approximation to the posterior such as a random sample from it or a local peak. The learning process uses the inferred posterior distribution across hidden states to update the parameters of the generative model so that it is more likely to produce the observed data. In some popular and statistically improper generative models, such as principle components or vector quantization, the inference process is simple and can be performed by projection or by a winner-take-all competition. These degenerate cases have tended to conceal the true nature of unsupervised learning.

It is interesting to see how the five objectives above all make sense from the perspective of generative models. Clustering is just fitting a two-stage generative model in which we first make a discrete choice of a mixture component and then generate from a density determined by this component. The distance measure used for clustering corresponds to the negative log probability of the data under a component of the model. The usual hard assignment of a data vector to the closest cluster corresponds to approximating the posterior distribution over the hidden choices by the single most likely choice.

Discovering independent hidden features can be achieved by fitting a generative movel in which the activities of the hidden units are chosen independently. Some particularly tractable special cases arise when the observed data is modeled as a linear combination of the hidden unit activities plus additive Gaussian noise. If the hidden units have Gaussian priors, then this is factor analysis (Neal and Dayan 1997). If the hidden units have improper uniform priors, then it is principal components analysis. If the priors for the hidden units are independent but non-Gaussian, then it is ICA. The inference process for ICA becomes straightforward as the additive Gaussian observation noise goes to zero. For nonzero noise the posterior distribution over hidden states must be approximated, typically by finding its peak.

Discovering temporal invariances is just fitting a dynamical system (without driving inputs) in which the state-transition matrix for the hidden state space is the identity matrix. This makes it obvious that temporal invariance is naturally subsumed by linear dependence over time. If the observed data is modeled as a linear function of the hidden state space then temporal

invariants can be learning by fitting a linear dynamical system.

Hidden units that represent unlikely coincidences arise naturally when fitting a generative model to data because they allow common conjunctions to have much higher density than they would get if their constituents were generated separately. The same applies to redundancies in general. An advantage of extracting redundancies by fitting a generative model is that it naturally makes differenet units within a layer do different things. No special decorrelating mechanism is required to make different units grab different redundancies. Also, in a multilayer generative model the lower layers can maximize the probability density of the observed data by extracting redundancies among which there are easily extracted redundancies. There is no necessity for the lower hidden layers to extract features that are fully decorrelated or statistically independent. All that is required is that the features should be statistically independent *given* the features in the layer above. So it makes sense to extract natural causes of images like noses and mouths because they are approximately statistically independent *given the face* even though they are very highly correlated.

Discovering very non-Gaussian projections is just what a linear generative model containing a single hidden unit will do if it is given (or allowed to empirically construct) a non-Gaussian prior for the activity of a hidden unit. Assume that the generative model treats the observed data as the output of the model plus Gaussian noise. To maximize the log likelihood of the data the model needs to account for as much of the variance in the data as possible without using improbable states of the hidden unit and without using large generative weights that dilute the probability density of the hidden activities. So the hidden unit needs to have a lot of variance in its activity but as little entropy as possible. This is a natural definition of what it means to be far from Gaussian, since the Gaussian distribution maximizes entropy for a given variance.

Much of the progress in the last few years has come from fitting linear generative models with a single layer of hidden units. In the longer term it seems likely that these simple models will have to be replaced by nonlinear generative models with multiple hidden layers. Consider, for example, the problem of extracting from an intensity image the three position and three orientation parameters of a rigid three-dimensional object. Suppose we want the activities of hidden units to represent these "instantiation" parameters explicitly or to represent posterior distributions in the space of instantiation parameters. The instantiation parameters are nonlinearly related to pixel intensities, so even if we ignore the formidable problems of image segmentation, no linear generative model will suffice. One approach (Gold, Rangarajan, and Mjolsness 1996) is to transform pixel intensities into the image coordinates of identified fragments. The alternative is to have a multilayer, nonlinear model in which units in higher layers somehow represent the instantiation parameters of progressively larger and more complex fragments of objects. To achieve economy and generalization it seems essen-

tial for each fragment to potentially be generated by many different larger fragments in the layer above so the connectivity cannot be restricted to a tree structure. This is precisely the sort of architecture that is found in the cerebral cortex.

A major challenge for unsupervised learning is to get a system of this general type to learn appropriate representations for images. The major difficulty is that it is intractable to compute the full posterior distribution across hidden states in such a complex generative model.[2] One ray of hope is that the standard EM method of fitting models to data can be generalized so that learning can proceed effectively even if the posterior distribution is only approximated. Neal and Hinton (1998) show that a quantity equivalent to free energy is minimized by alternating between a partial M-step that improves the log likelihood of the data given the assumed distribution over hidden states and a partial E-step that improves the approximation to the true posterior distribution. The free energy is equivalent to the description length used by Zemel and Hinton (1995) and it can also be viewed (with a sign reversal) as the log likelihood of the data under the model penalized by a measure of the difficulty of performing inference with the model. The penalty is just the Kullback-Liebler divergence between the approximating distribution and the true posterior distribution. Retrospectively, it is easy to see that a biological organism would much rather have a model in which correct inference can be approximated easily than a model which gives slightly higher likelihood to the data but in which the posterior distribution is hard to approximate. So we arrive at a new objective function for unsupervised learning that recognizes the difficulty of performing inference in sophisticated generative models and builds in a measure of the tractability of inference. Helmholtz machines (Dayan, Hinton, Neal, and Zemel 1996) are an attempt to optimize such an objective function using top-down connections for the generative model and feedforward, bottom-up connections for the approximate inference process.

## References

Amari, S.-I. (1998) Natural gradient works efficiently in learning. *Neural Computation* 10(2):252–276. Reprinted in this volume.

Atick, J. J. and Redlich, A. N. (1993) Convergence algorithm for sensory receptive field development. *Neural Computation* 5(1):45–60. Reprinted in this volume.

Attneave, F. (1954) Informational aspects of visual perception. *Psychological Review* 61:183–193.

[2] For one-dimensional strings of $n$ symbols produced by a stochastic context-free grammar it is possible to consider all $O(n^2)$ possible connected substrings and so the posterior distribution over parses can be computed exactly. This is not possible for two-dimensional images because there are exponentially many connected subregions.

Barlow, H. B. (1959) Sensory mechanisms, the reduction of redundancy, and intelligence. In National Physical Laboratory Symposium No. 10, *The Mechanisation of Thought Processes*, pp. 535–559. London: Her Majesty's Stationary Office.

Barlow, H. B. (1989) Unsupervised learning. *Neural Computation* 1(3):295–311. Reprinted in this volume.

Becker, S. and Hinton, G. E. (1993) Learning mixture models of spatial coherence. *Neural Computation* 5(2):267–277. Reprinted in this volume.

Bell, A. J. and Sejnowski, T. J. (1995) An information-maximization approach to blind separation and blind deconvolution. *Neural Computation* 7(6):1129–1159. Reprinted in this volume.

Bell, A. J. and Sejnowski, T. J. (1997) The "independent components" of natural scenes are edge filters. *Vision Research*, 37:3327–3338.

Bienenstock, E. L., Cooper, L. N., and Munro, P. W. (1982) Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience* 2:32–48.

Brown, T., Kairiss, E., and Keenan, C., (1990). Hebbian synapses: biophysical mechanisms and algorithms, *Ann. Rev. Neurosci.* 13:475–511.

Comon, P. (1994) Independent component analysis: a new concept? *Signal Processing* 36:287–314.

Crick, F. and Mitchison, G. (1983) The function of dream sleep. *Nature* 304:111–114.

Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (1995) The Helmholtz machine. *Neural Computation* 7(5):889–904. Reprinted in this volume.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* 39:1–38.

Fahlman, S., and Lebiere, C. (1990). The cascade-correlation learning architecture. In Touretzky, D. (ed.), *Advances in Neural Information Processing Systems* 2, pp. 524–532, San Mateo, CA: Morgan Kaufmann.

Field, D. J. (1994) What is the goal of sensory coding? *Neural Computation* 6(4):559–601. Reprintedin this volume.

Foldiak, P. (1991) Learning invariance from transformation sequences. *Neural Computation* 3(2):194–200. Reprinted in this volume.

Gold, S., Rangarajan, A. and Mjolness, E. (1996) Learning with preknowledge: clustering with point and graph matching distance measures. *Neural Computation* 8(4):787–804. Reprinted in this volume.

Hinton, G. and Sejnowski, T. (1986) Learning and relearning in Boltzmann machines. In Rumelhart, D. and McClelland, J. (eds.), *Parallel Distributed Processing*, volume 1, chapter 7, pp. 282–317. Cambridge, MA: MIT Press.

Hopfield, J. J., Feinstein, D. I., and Palmer, R. G. (1983) "Unlearning" has a stabilizing effect in collective memories. *Nature* 304(5922):158–159.

Hubel, D. H. and Wiesel, T. N. (1968) Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* 195:215–244.

Hyvarinen, A. and Oja, E. (1977) A fast fixed-point algorithm for independent component analysis. *Neural Computation* 9(7):1483–1492. Reprinted in this volume.

Intrator, N. (1992) Feature extraction using an unsupervised neural network. *Neural Computation* 4(1):98–107. Reprinted in this volume.

Kambhatla, N. and Leen, T. K. (1997) Dimenison reduction by local principal component analysis. *Neural Computation* 9(7):1493–1516. Reprinted in this volume.

Lee, T.-W., Girolam, M., and Sejnowski, T. J. (in press) Independent component analysis using an extended infomax algorithm for mixed sub-Gaussian and super-Gaussian sources. *Neural Computation* 11(2).

Linsker, R. (1986) From basic network principles to neural architecture: emergence of spatial-opponent cells. *Proceedings of the National Academy of Sciences of the United States of America* 83:7508–7512.

Linsker, R. (1992) Local synaptic learning rules suffice to maximize mutual information in a linear network. *Neural Computation* 4(5):691–702. Reprinted in this volume.

Markram, H., Lubke, J., Frotscher, M., and Sakmann, B. (1977) Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* 275:213–215.

Miller, K. D., Keller, J. B., and Stryker, M. P. (1989) Ocular dominance column development: analysis and simulation. *Science* 245:605–615.

Neal, R. M. and Dayan, P. (1997) Factor analysis using delta-rule wake-sleep learning. *Neural Computation* 9(8):1781–1803. Reprinted in this volume.

Neal, R. and Hinton, G. E. (1998) A new view of the EM algorithm that justifies incremental and other variants. In M. I. Jordan (ed.), *Learning in Graphical Models*. Dordrecht: Kluwer Academic Press.

Obermayer, K. and Sejnowski, T. J. (1998) *Self-Organizing Map Formation: Foundations of Neural Computation*. Cambridge, MA: MIT Press.

Olshausen, B. A. and Field, D. J. (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 1381:607–609.

Platt, J. (1991) A resource-allocating network for function interpolation. *Neural Computation* 3(2):213–225. Reprinted in this volume.

Stone, J. V. (1996) Learning perpetually salient visual parameters using spatiotemporal smoothness constraints. *Neural Computation* 8(7):1463–1492. Reprinted in this volume.

Zemel, R. S. and Hinton, G. E. (1995) Learning population codes by minimizing description length. *Neural Computation* 7(3):549–564. Reprinted in this volume.

Zhang, K., Sereno, M. I., and Sereno, M. E. (1993) Emergence of position-independent detectors of sense of rotation and dilation with Hebbian learning: an analysis. *Neural Computation* 5(4):597–612. Reprinted in this volume.