

Orientation Selective Cells Emerge in a Sparsely Coding Boltzmann Machine

Cornelius Weber and Klaus Obermayer

Informatik, Technische Universität Berlin, cweber@cs.tu-berlin.de

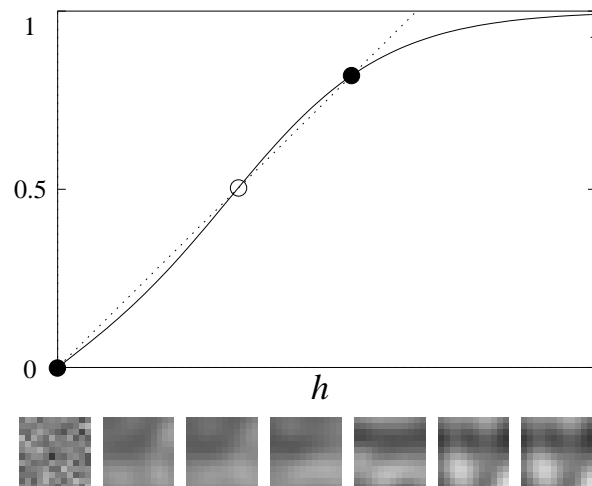
We implemented a recurrent model consisting of weights w_{ij} from every LGN cell j to every V1 cell i and symmetrical feedback weights $w_{ji} = w_{ij}$. There are no lateral connections within a layer. Neuron dynamics is stochastic. Hidden neuron activations u_i take one of three possible values, -1, 0, +1 where 0 is degenerate (occurs often) thus having a high probability to emerge (*sparse coding*). Input neuron activations x_j are evenly distributed. The probability for an activation vector (\vec{u}, \vec{x}) to occur obeys the Boltzmann distribution on an assigned energy value.

The learning rule is derived by maximum-likelihood: the probability for the data distribution to be generated by the model is maximized. We obtain $\Delta w_{ij} \approx \langle r_j s_i \rangle^+ - \langle r_j s_i \rangle^- - \lambda w_{ij} \sum_j w_{ij}^2$. The first term constitutes Hebbian learning in the clamped phase, i.e. data are shown. The second term constitutes anti-Hebbian learning where in absence of data all neurons are freely activated. Both terms are used for the standard Boltzmann machine. Additionally we impose a soft constraint on the weights which is scaled by λ in order to hinder the weights to compensate for sparse coding.

Gibbs sampling on the Boltzmann distribution is approximated by mean-field relaxation of effectively continuous neurons. The resulting mean-field transfer function on the net input h for the hidden neurons (left upper figure) has a stable fixpoint at 0 (ferromagnetic, one-neuron model).

The input data are patches from natural images. Upon training hidden neurons become edge detectors. The figure to the very right shows a selection of the weight vectors sorted from top to bottom according to their absolute length. Each square shows the receptive field of a hidden neuron, black indicating positive, white negative weights to one of the 12×12 input neurons. A larger number of hidden neurons, 12×20 , forms an *overcomplete representation* of the input.

If only the clamped phase is used for training every hidden neuron learns the first principle component of the data (result not shown). Learning by only the free-running phase leads to variety among the receptive fields and also to a continuous rearrangement of weights so that learning never stops.



The lower left figure shows how activations \vec{x} on the 12×12 input neurons relaxate. Black denotes positive, white negative activations. After random initialization (first square) values converge from left to right to final values which are, however larger than those used for learning of the weights (third square). The reason is that a *sparse* activation vector \vec{u} is not stable in the mean-field dynamics.

