A robust biologically plausible implementation of ICA-like learning

Felipe Gerhard, Cristina Savin and Jochen Triesch

Frankfurt Institute for Advanced Studies
Ruth-Moufang-Strasse 1, Frankfurt am Main - Germany

Abstract. We present a model that can perform ICA-like learning by simple, local, biologically plausible rules. By combining synaptic learning with homeostatic regulation of neuron properties and adaptive lateral inhibition, the neural network can robustly learn Gabor-like receptive fields from natural images. With spatially localized inhibitory connections, a topographic map can be achieved. Additionally, the network can solve the Földiák bars problem, a classical nonlinear ICA task.

1 Introduction

Many neural network models have been proposed for independent component analysis. Although grounded on information theoretic principles, they fail to provide a full story of how ICA-like computation could be performed by biological neurons with simple, local, biologically plausible rules. Additionally, previous models typically require tight regulation of some parameters (transfer function, lateral inhibition) based on the statistics of the input [1, 2]. Clearly, additional plastic changes must account for this parameter tuning in biological networks.

We hypothesize that intrinsic plasticity (IP), a homeostatic mechanism known to bring the activity of the neuron back to baseline in response to experimentally induced changes in firing rates [3], may contribute to ICA-like processing in neural networks. Our model extends previous work on how a single neuron could recover a single independent component by combining IP and Hebbian learning [4]. Similar to classic single-unit implementations of ICA [1], the output of different neurons is decorrelated by adaptive lateral inhibition. Our model offers a simple, biological plausible account of how ICA-like computation could be performed in a neural system. Moreover, unlike previous models, our system maintains its function for a wide range of parameters, a robustness reminiscent of biological systems.

The paper is organized as follows. After shortly presenting the IP model from [4], we introduce our network model, which is shown to learn Gabor filters from natural images. A topographic map is then derived, by limiting the range of the lateral inhibition to the spatial neighborhood of each neuron. Additionally, our model can solve the Földiák bars problem, a classical nonlinear ICA task. We conclude by comparing our work with related models for V1 receptive field development.

2 Model

2.1 Intrinsic plasticity

It has been hypothesized that IP may maximize the information transmission between the neuron's total input and output, under the constraint of a fixed mean firing [5, 4]. This is equivalent to forcing the output distribution of the neuron to an exponential. As in [4], we consider a parameterized sigmoid for the transfer function $y = S_{ab}(h) = \frac{1}{1+\exp(-(ah+b))}$ with $a \in \Re^{>0}$ and $b \in \Re$. The distance from the desired distribution, measured by the Kullback-Leibler divergence, is minimized by stochastic gradient descent, yielding the parameter update rules: $\Delta a = \eta_{\text{IP}} \left(\frac{1}{a} + h - (2 + \frac{1}{\mu})hy + \frac{1}{\mu}hy^2 \right)$, $\Delta b = \eta_{\text{IP}} \left(1 - (2 + \frac{1}{\mu})y + \frac{1}{\mu}y^2 \right)$, where μ is the desired mean activity level and η_{IP} is a small learning rate.

2.2 Neural network

The network consists of N neurons, which receive a 2×100 input $\vec{x}$ from a ON- and OFF- population in the lateral geniculate nucleus (LGN), and are laterally connected by inhibitory synapses. The structure employed here is similar to that in [6]. Initially, feed-forward weights W have random values taken uniformly in the range $[0, 1]$, while the inhibitory weights U are all set to zero.

The excitatory drive to the neurons is calculated as: $\vec{h}^{\text{exc}} = W\vec{x}$, while the all-to-all lateral inhibition is given by: $\vec{h}^{\text{inh}} = A_{\text{inh}}U\vec{y}$. Here, $\vec{y}$ denotes the output of each neuron and A_{inh} is a scaling factor. The neuron output is determined by the equation $y_{\text{i}} = S_{a_{\text{i}}b_{\text{i}}}(h_{\text{i}})$, with $\vec{h} = \vec{h}^{\text{exc}} - \vec{h}^{\text{inh}}$. This equation is implicit with respect to y_{i} and must be solved numerically at each iteration.

The feed-forward weights are updated according to a Hebbian-like covariance rule: $\Delta W_{\text{ij}} = \eta_{\text{Hebb}}(y_{\text{i}} - \mu) \cdot (x_{\text{j}} - W_{\text{ij}})$, and inhibitory weights are updated by a classical anti-Hebbian rule: $\Delta U_{\text{ij}} = \eta_{\text{anti}}(y_{\text{i}}y_{\text{j}} - \mu^2)$, where η_{Hebb} and η_{anti} are learning rates. The diagonal elements are set to zero and all values are rectified.

The learning rule for the feed-forward connections —through the factor $(y_i - \mu)$— makes sure that the receptive fields stay stable once the desired output activity is reached, while the normalization term prevents the weights from unbounded growth. The anti-Hebbian learning on the lateral connections increases the inhibition between two neurons which repeatedly get excited by the same stimulus, again with a normalization term implementing synaptic scaling. A stationary state ($E[\Delta U_{ij}] = 0$) is reached once the average outputs of the neurons have the desired mean μ and their activities are uncorrelated, such that $E[y_i y_j] = E[y_i]E[y_j] = \mu^2$. Similar learning rules are used in [6] (see Section 4).

3 Results

3.1 Learning Gabor receptive fields

Images from the van Hateren dataset [7] are preprocessed by a Difference of Gaussians (DoG) filter, with the center width of 1 pixel and 1.2 pixels for the surround, similar to [8]. Image patches of size 10 × 10 pixels are extracted randomly across all possible positions. The mean of each patch is subtracted and the resulting vector is normalized to unit (Euclidean) length. The rectified pixel values are simulated as a population of ON- and OFF-response cells.

Figure 1 shows the results of a run with 3×10^5 iterations, when the system has converged to a stationary state. Simulation parameters are: $\eta_{\mathrm{Hebb}} = 0.01$, $\eta_{\mathrm{anti}} = 0.002$, $\eta_{\mathrm{IP}} = 0.02$, $\mu = 0.005$, $A_{\mathrm{inh}} = 1$. Initially, $a_i(0) = 800$ and $b_i(0) = -500$ for all i. The neurons develop different Gabor-like receptive fields, qualitatively comparable to those of other computational models [9, 6, 8]. Since uncorrelated outputs are only a necessary condition for independence, we estimated the pairwise mutual information to assess independence between the activities of different neurons. The normalized pairwise mutual information (estimated during the last 10^5 iterations) decreases to a small average value of 0.024 (see [8] for details on normalization of the mutual information estimate). This corresponds to 2.4% of the maximal possible value.

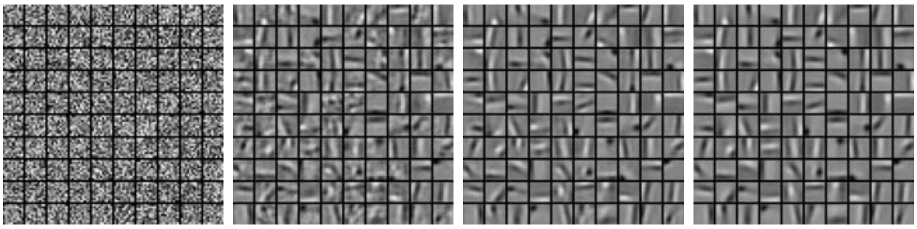

Fig. 1: Natural images. The connection strengths W_{ij} of the $N = 100$ neurons. The color map is chosen such that gray corresponds to zero filter strength. From left to right: Initial feed-forward connections; receptive fields after 5×10^4, 10^5, resp. 3×10^5 iterations.

3.2 Learning topographic representations

We extend the model above, by adding a spatial structure to the neural network. We consider each neuron to be located on a two-dimensional rectangular grid with periodic boundary conditions. A kernel function $k(d)$ is used to modulate the strength of the lateral inhibition between neurons of (Euclidean) distance d on the grid defined above. In particular, we use a rectified DoG kernel with two parameters σ_1 and σ_2.

Figure 2 shows the resulting orientation maps for a network with $N = 400$ neurons (i. e. a 4 times overcomplete basis) using two different choices for the kernel, but otherwise the same parameters as in section 3.1. Even without explicit excitatory lateral connections, the imposed connection pattern leads to the formation of patches of similar orientations whose size is determined by the width of the used kernel. On a more coarse-grained level, one can observe sharp discontinuities as well as rather smooth transitions in the orientation directions across the map.

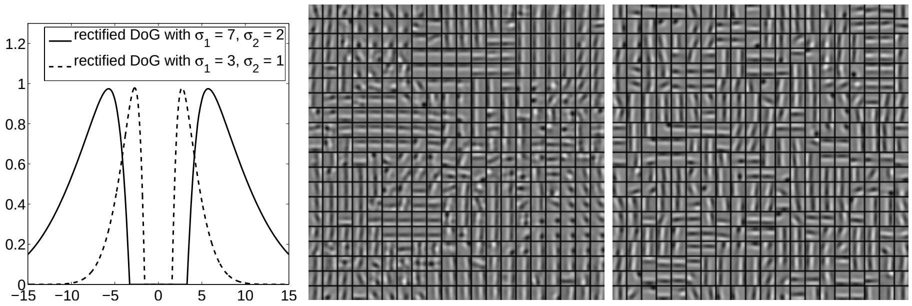

Fig. 2: Emergence of orientation maps. $N = 400$ neurons are used and receptive fields are shown after convergence. Left: DoG kernels used (normalized to unit height). Middle: Receptive fields using DoG kernel $k(d)$ with $\sigma_1 = 7$ and $\sigma_2 = 2$. Right: Receptive fields using DoG kernel $k(d)$ with $\sigma_1 = 3$ and $\sigma_2 = 1$.

3.3 The Földiák bars

The Földiák "bars test" [10] defines input images as a collection of horizontal and vertical bars (with a width of one pixel), superimposed non-linearly (intersection points are as bright as the rest of the bar). Each bar occurs independently with a fixed probability p.

For consistency, the mean of each images is subtracted to obtain both positive and negative intensity values which can be fed to the ON- and OFF-cells. The model structure is the same as described in section 2.2 with $N = 20$ neurons (a complete basis). Figure 3 shows the receptive fields for various intermediate stages and the final configuration after 3×10^5 iterations. The network converges to a stationary state where each neuron learns a different bar. The parameters for the simulation are: $\eta_{\text{Hebb}} = 0.01$, $\eta_{\text{anti}} = 0.01$, $\eta_{\text{IP}} = 0.01$, $\mu = 0.05$, $A_{\text{inh}} = 1$. Initially, a and b are set to the same value for all neurons, namely $a_i(0) = 50$ and $b_i(0) = -50$.

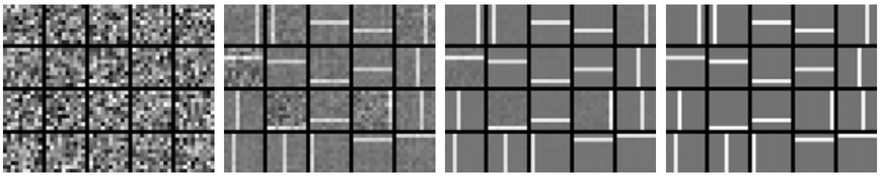

Fig. 3: The Földiák bars. Receptive fields evolution for the $N = 20$ neurons. The color map is chosen such that gray corresponds to zero filter strength, and white to the maximum positive value. From left to right: Initial feed-forward connections; receptive fields after 4×10^4 , 6×10^4, resp. 3×10^5 iterations.

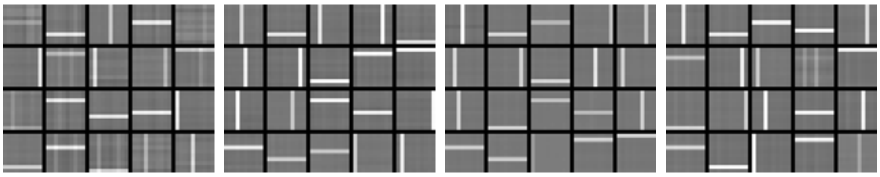

Fig. 4: Receptive fields for bars stimuli for different parameter settings. From left to right: $A_{\text{inh}} = 0.5, 0.8, 4, 6$.

3.4 Robustness

We investigated how sensitive the development of receptive fields is with respect to the network parameters, for both the natural image patches and the bars problem. Due to computational constraints, only variations along one dimension at the time in parameters space were considered. Table 1 summarizes the results. For both the Gabor-like receptive fields and the bars the results are stable in a wide range of parameters. Due to the fact that the sigmoid output is bounded, a reasonable match with an exponential can only be achieved for small means. Exemplarily, Figure 4 shows the receptive fields for the extrema of valid values for A_{inh} and —for comparison— two non-convergent cases.

parameters	η_{IP}	η_{Hebb}	η_{anti}	μ	a_0	b_0	A_{inh}
bars: lower bound	0.01	0.002	0.005	0.02	1	-1000	0.8
bars: upper bound	0.1	0.1	0.1	0.12	1000	0	4
images: lower bound	0.01	0.005	0.001	0.002	1	-1000	0.01
images: upper bound	0.2	0.02	0.01	0.008	1000	0	2

Table 1: Robustness to model parameters. For the bars, we consider parameters for which the network learns the complete basis after at most 2×10^5 iterations. For natural images, parameter ranges are shown for which the system still develops localized Gabor-like receptive fields after at most 3×10^5 iterations.

4 Conclusions and future work

Information-theoretic approaches have been successfully used to explain certain firing properties of biological neurons [9]. We were able to learn receptive fields resembling those of simple cells in visual cortex in a new neural network model that combines IP, Hebbian learning and adaptive lateral inhibition. Furthermore, our method solves the bars problem on which traditional ICA or PCA algorithms are prone to fail [11]. While other approaches require significant parameter tuning, our system shows great robustness to wide parameter changes, as expected for biological neural networks. This is owed to the IP, which serves a homeostatic purpose stabilizing the network dynamics, especially during the decorrelation process.

The present model has some similarities to earlier work. In particular, it shares the use of an intrinsic plasticity rule with the model in [8], which could produce a topographic map of V1-like receptive fields by spatially selective lateral inhibition. However, our model replaces its neighborhood function gating the learning with the more biologically plausible adaptive lateral inhibition. Also, Gabor-like receptive fields were developed in [6], using all-to-all lateral inhibition. Instead of the IP, this work considered a first-moment dependent sensitivity rule regulating parameter b, but kept a fixed, adjusting it "by hand" as to yield exponential-like output distributions. This adaptation is done automatically by the here-used intrinsic plasticity rule.

As output activities remain somewhat correlated in the stable state, it is sensible to assume that a second network layer, receiving this activity as input, would discover some non-trivial structure. There have been only few successful attempts to construct such iterative or hierarchical models for ICA [12]. The difficulty lies in the fact that linear ICA models require some additional non-linear processing for converting the output of one layer into a suitable input for the subsequent one. The nonlinear model proposed here naturally works with non-negative, heavy-tailed distributions for both input and output, making our network a feasible building block for more complex hierarchical structures.

References

[1] A. Hyvärinen. One-unit contrast functions for independent component analysis: a statistical analysis. *Neural Networks for Signal Processing*, pages 388–397, 1997.

[2] A. Bell and T. Sejnowski. An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[3] W. Zhang and D.J. Linden. The other side of the engram: experience-dependent changes in neuronal intrinsic excitability. *Nature Reviews Neuroscience*, 4:885–900, 2003.

[4] J. Triesch. Synergies between intrinsic and synaptic plasticity mechanisms. *Neural Computation*, 19:885–909, 2007.

[5] M. Stemmler and C. Koch. How voltage-dependent conductances can adapt to maximize the information encoded by neuronal firing rate. *Nature Neuroscience*, 2(6):521–527, 1999.

[6] M.S. Falconbridge, R.L. Stamps, and D.R. Badcock. A Simple Hebbian/Anti-Hebbian Network Learns the Sparse, Independent Components of Natural Images. *Neural Computation*, 18(2):415–429, 2005.

[7] J.H. van Hateren. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society B: Biological Sciences*, 265(1394):359–366, 1998.

[8] N.J. Butko and J. Triesch. Learning sensory representations with intrinsic plasticity. *Neurocomputing*, 70(7-9):1130–1138, 2007.

[9] B. Olshausen and E. Simoncelli. Natural image statistics and neural representation. *Annu. Rev. Neuroscience*, 24:1193–1216, 2001.

[10] P. Földiák. Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics*, 64:165–170, 1990.

[11] S. Hochreiter and J. Schmidhuber. Feature extraction through lococode. *Neural Computation*, 11(3):679–714, 1999.

[12] H. Shan, L. Zhang, and G.W. Cottrell. Recursive ICA. *Advances in Neural Information Processing Systems*, 19, 2007.