Available online at www.sciencedirect.com
ScienceDirect
Neurocomputing 70 (2007) 1130–1138
NEUROCOMPUTING
www.elsevier.com/locate/neucom

Learning sensory representations with intrinsic plasticity

Nicholas J. Butko[a,*], Jochen Triesch[a,b]

[a]Department of Cognitive Science, University of California San Diego, 9500 Gilman Dr., MC 0515, La Jolla, CA 92093-0515, USA
[b]Frankfurt Institute for Advanced Studies, Johann Wolfgang Goethe University, Max-von-Laue-Str. 1, 60438 Frankfurt am Main, Germany

Available online 10 December 2006

Abstract

Intrinsic plasticity (IP) refers to a neuron's ability to regulate its firing activity by adapting its intrinsic excitability. Previously, we showed that model neurons combining a model of IP based on information theory with Hebbian synaptic plasticity can adapt their weight vector to discover heavy-tailed directions in the input space. In this paper we show how a network of such units can solve a standard non-linear independent component analysis (ICA) problem. We also present a model for the formation of maps of oriented receptive fields in primary visual cortex and compare our results with those from ICA. Together, our results indicate that intrinsic plasticity that tries to locally maximize information transmission at the level of individual neurons may play an important role for the learning of efficient sensory representations in the cortex.

Keywords: Intrinsic plasticity; Information theory; Unsupervised learning; Independent component analysis; Primary visual cortex

1. Introduction

1.1. Mechanistic vs. functional models

Computational models of unsupervised learning of sensory representations in the brain abound. Frequently, they fall into one of two categories: *mechanistic models* or *functional models*. Mechanistic models start from neuroscientific data about the structure of cortical networks and cortical plasticity mechanisms (cell types, connection patterns, plasticity rules, etc.) which are distilled into simplified models. These models are trained on actual sensory data or noise patterns and the learned representations can be compared to neurophysiological observations. If the resulting representations are similar to those found in the brain then this provides evidence that the processes in the brain have been accurately captured, but it does not clarify why the brain operates this way or in what sense the brain's solution may be optimal. An example of a model of this kind is by Linsker [15], where V1-style orientation columns are learned from random prenatal visual noise through Hebbian learning. Later Miller extended this work to learn many of the various map-structures in V1, and used model neurons that were somewhat more plausible [19].

Functional models focus on the abstract computational goal of the problem. For the case of learning sensory representations they start by asking: what is the *optimal* way to represent sensory stimuli such as natural images, where optimality is usually defined with respect to certain statistical criteria (e.g. sparseness, independence, temporal coherence, etc.) and additional constraints. Algorithms are derived to learn the optimal solution to the problem, which can again be compared to neuroscientific data. If the found solution resembles the biological solution, then this provides evidence that the brain may in fact be trying to optimize a similar objective function. Through what mechanisms the brain may achieve this goal is typically not answered, however. Some examples of such an approach will be given below.

Both mechanistic and functional models have their merits, but for a comprehensive understanding of sensory coding in the cortex we arguably have to develop models that bridge functional and mechanistic levels of description. Such models should explain how the physiological mechanisms contribute to optimizing the system's information processing properties in a meaningful way. In the

*Corresponding author.
E-mail address: nbutko@cogsci.ucsd.edu (N.J. Butko).

0925-2312/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2006.11.006

following, we develop a model that can be viewed as a step in this direction.

1.2. Information maximization

A central idea in many functional models of the development of sensory representations is information maximization [1,3,16,20,22]. According to some formulations of this idea, individual neurons should maximize the entropy of their firing rate distribution. If the firing rate is constrained to lie in a fixed interval between zero and the neuron's maximum firing rate, then entropy maximization means that the neuron should use all its firing rate levels equally often. In order to achieve this, it should spread out its responses in dense regions of the input space and compress responses in sparse regions such that it maps the distribution of its inputs to a uniform distribution of its outputs, maximizing entropy. Biological evidence for this idea comes from Laughlin, who showed that blowfly large monopolar cells have been adapted so that their input/output transfer functions nearly optimally represent the contrast statistics of the blowfly's visual environment [13].

Information maximization may not be the only important objective, however, and energy considerations may also play an important role for sensory coding in the brain, e.g. [14]. In particular, Baddeley et al. found that neurons in different visual cortical areas of cats and monkeys show exponential distributions of their firing rate. They have argued that this maximizes a neuron's information transfer given a fixed energy budget [2]. This is because the exponential distribution has the maximum entropy among all distributions of a positive random variable (the firing rate) with a fixed mean. This and other reasons suggest that *sparse* representations, where individual units are highly active only rarely, may be an important principle of sensory coding [10].

On the modeling side, Olshausen and Field showed that localized, oriented, and bandpass receptive fields similar to those observed in primary visual cortex (V1) arise when optimizing image reconstruction error subject to lifetime sparseness constraints [21]. They imposed a sparse prior on the contribution of each basis function in a generative model with the intuition that among the space of possible sources of an image, each one is present only rarely. In a closely related approach, Bell and Sejnowski showed that the information maximization principle can be applied to the independent component analysis (ICA) problem. They applied their technique to natural images and also found localized, oriented, and bandpass sources [4].

1.3. What is the role of intrinsic plasticity for learning sensory representations?

Most work on the learning of sensory representations has focused on synaptic plasticity as the only mechanism for learning efficient codes. But it is becoming increasingly clear that biological neurons also regulate their pattern of firing by adapting their intrinsic excitability through the modification of voltage-gated channels in their membrane. Such *intrinsic plasticity* (IP) seems to be a ubiquitous phenomenon in the brain [30]. For example, Desai et al. showed that neurons that had been prevented from spiking for two days increased their response to current injection [6]. Consistent with this finding, it is frequently assumed that IP contributes to the homeostasis of a neuron's mean firing activity. A few computational models do in fact incorporate a mechanism for regulating the mean activity level of a unit by controlling a "threshold" parameter [7–9]. But it is also plausible that IP may help to optimize the encoding and transmission of information in a more sophisticated fashion. Concretely, it has been speculated that IP may be instrumental in achieving approximately exponential firing rate distributions in cortical neurons [23]. More recently, we have shown that an IP mechanism that drives a neuron to exhibit an exponential firing rate distribution can synergistically interact with Hebbian learning at the synapses. The two processes lead to the discovery of heavy-tailed directions in the input space [24,26].

In this paper we extend these results to networks of neurons with IP and Hebbian learning. Our specific goal is to explore the potential role of IP for learning efficient map-like representations for sensory stimuli. The model we present in the following attempts to bridge the gap between mechanistic and functional models. On the one hand, it has a clear connection to the idea of information maximization and energy efficient coding [28]. On the other hand, it has a mechanistic formulation that is biologically viable because the learning mechanisms make use of information that is local in time and space. While similar bridges have been attempted before, e.g. [5,8], our model is distinguished by utilizing an IP model derived from information theory as a mechanism for the learning of efficient sensory representations.

2. Network model with intrinsic plasticity

We consider a network of units learning to represent a sensory input vector $\mathbf{x}$. The activity of unit i in the network in response to input $\mathbf{x}$ is given by

$$y_i(h_i) = [1 + \exp(-a_i h_i - b_i)]^{-1} \quad \text{with } h_i = \mathbf{x} \cdot \mathbf{w}_i, \tag{1}$$

where $\mathbf{w}_i$ is the neuron's weight vector, "$\cdot$" denotes the inner or dot product, and a_i and b_i are adjustable parameters of the neuron's transfer function that are controlled by IP (compare Fig. 1a). In particular, a_i and b_i are adapted in such a way that the unit's output y_i assumes an approximately exponential distribution. To this end we have previously derived a learning rule for a_i and b_i that performs stochastic gradient descent on the Kullback–Leibler divergence between the unit's output distribution and the desired exponential distribution. This leads to the

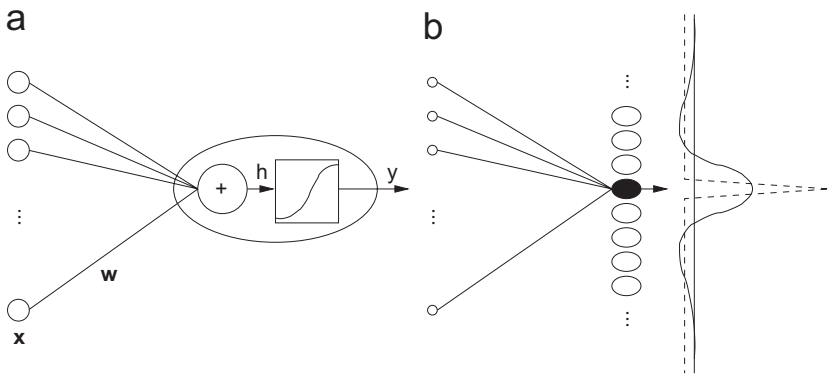

Fig. 1. (a) Illustration of an individual unit of the network. The weights **w** are adapted through Hebbian learning, the sigmoidal non-linearity is adapted through intrinsic plasticity. (b) Network architecture. The most activated unit (shaded) determines the sign and amount of synaptic learning in neighboring units via a neighborhood function. Two examples of neighborhood functions are shown (not drawn to scale).

following learning rule [25,26]:

$$a_i \leftarrow a_i + \eta_{\mathrm{IP}}[a_i^{-1} + h_i - (2+\mu^{-1})h_i y_i + \mu^{-1} h_i y_i^2],$$

$$b_i \leftarrow b_i + \eta_{\mathrm{IP}}[1 - (2+\mu^{-1})y_i + \mu^{-1} y_i^2], \qquad (2)$$

where "$\leftarrow$" denotes assignment, η_{IP} is a small learning rate and μ is the desired mean activity of all units. Since this learning rule has the effect of making the distribution of y_i a sparse, approximately exponential distribution, it maximizes the unit's entropy under the constraint of a fixed average activity: the unit transmits information efficiently. Note that this rule is local in space and time, making it physiologically viable.

Plasticity of the weight vectors $\mathbf{w}_i$ is modeled with a Hebbian learning rule. In [24], we considered a single unit learning rule of the form $\Delta\mathbf{w} \propto \mathbf{x}y$. We showed that the coupling of IP with this form of Hebbian learning allowed the unit to discover heavy-tailed directions in the input. We have generalized this result to other Hebbian learning rules in [26]. To extend this approach to a network of model neurons, we introduce a *neighborhood function* $\mathcal{N}$ as illustrated in Fig. 1b. The value of the neighborhood function for neuron i is determined by its activity y_i and the activities of all other neurons, *i.e.* $\mathcal{N}(y_i;\mathbf{y})$. In particular, we are considering neighborhood functions that depend on a unit's distance to the most activated unit in the layer—as frequently used in self-organizing maps. Specific forms of $\mathcal{N}$ are introduced below. The general idea is that the neighborhood functions can take on positive and negative values, such that learning is Hebbian for some units and anti-Hebbian for others. This is used to correlate and decorrelate weight updates in specific sets of units, allowing different units to develop different stimulus preferences and facilitating the formation of maps of smoothly varying stimulus preferences. The decorrelation serves the goal of reducing redundancy in the representation, the map formation contributes to wiring length minimization, because units with similar properties will be grouped together. After each stimulus presentation, the weights are updated according to:

$$\Delta\mathbf{w}_i = \mathbf{x}y\mathcal{N}(y_i;\mathbf{y}), \quad \mathbf{w}_i \leftarrow \frac{\mathbf{w}_i + \eta_{\mathrm{Hebb}}\Delta\mathbf{w}_i}{\|\mathbf{w}_i + \eta_{\mathrm{Hebb}}\Delta\mathbf{w}_i\|}, \qquad (3)$$

where η_{Hebb} is a learning rate and the normalization of the weight vector to unit length mimics competition between synapses on a neuron's dendritic tree [19].

3. The "bars" problem

As a first test bed for studying the learning of sensory representations with networks of units with intrinsic plasticity we consider the "bars" problem. This is a standard non-linear ICA problem introduced by Földiák [9]. Horizontal and vertical bars are presented on a retina of R-by-R pixels. The presence or absence of a bar is independent of that of any other bars. The unsupervised learning problem is to learn filters that correspond to the individual independent components, i.e. the bars. The problem is non-linear because the pixel at the intersection of two bars is just as bright as any other pixel of the bars, not twice as bright. In our previous work [24,26], we showed that a single model neuron with IP and Hebbian learning robustly discovers one of the bars when exposed to stimuli from the bars problem. Here we use a population of units to learn the complete problem. We use a retina of size 10-by-10 pixels and the probability of any of the 20 bars occurring in a given stimulus is 10%. The bar stimuli are unnormalized such that every "on" pixel has value 1.0 and every "off" pixel has value 0. Since we want filters that respond highly when bars are present and not otherwise, the desired mean firing rate is set to $\mu = 0.1$ which corresponds to 10% of a unit's maximum activation. $\mathcal{N}$ is chosen to enforce a winner-take-all competition between the units, so that the maximally activated neuron updates its weight vector in a standard Hebbian fashion, and all other units update their weight in an anti-Hebbian manner

regulated by a decorrelation parameter β:

$$\mathcal{N}_{\text{bars}}(y_i;\mathbf{y}) = \begin{cases} 1, & y_i = \max(\mathbf{y}), \\ -\beta & \text{else.} \end{cases} \quad (4)$$

All units update their intrinsic parameters independently, as described in (2).

We examined the learning of bars within the described framework, systematically probing the value of the neighborhood-interaction parameter β, which ranged from 0 to 0.5 in steps of 0.05. Other parameters were: $\eta_{\text{Hebb}} = 0.01$, $\eta_{\text{IP}} = 0.005$, and $\mu = 0.1$. The networks always consisted of 20 units (the number of individual bars). For each value of β we ran 30 independent experiments with 300,000 randomly generated bars stimuli each. Typical examples of bars stimuli and learned representations for different values of β are shown in Fig. 2. We found that the learning result fell in one of three regimes depending on whether there was too little neighborhood interaction, a good amount of interaction, or too much. Perfect learning results were obtained for β values from 0.1 to 0.2 as illustrated in Fig. 3. This means that every unit in the network learned to represent one distinct bar. Learning substantially worsened when β was less than 0.1 or greater than 0.25. When β is too low, all bars are learned, but some are duplicated in the population (some filters learn more than one bar). When β is too high, all bars are learned exactly once, but some filters learn two bars, leaving other filters to learn no bars (see examples in Fig. 2).

Varying the learning rates η_{Hebb} and η_{IP} affected learning little, provided both remained above 0. The complete set of filters would *not* be learned without intrinsic plasticity, however. We also studied the influence of μ on the learning result. When μ was 0.05, redundant filters were learned, i.e. multiple units learned to represent the same individual bar while some bars were not represented at all. When it was 0.2, multiple bars were represented within single filters. This suggests that when the true mean of the components is unknown, it may be a better strategy to choose μ too high rather than too low. This way, all true sources will likely be captured because individual filters each learn to represent several sources.

Since its introduction by Földiák, a number of different network architectures for solving the bars problem have been proposed and a number of variations on the problem have also been considered in the literature. The performance of some of the more complex approaches has been tested quite thoroughly, e.g. [17]. While a comprehensive review of this literature is beyond the scope of this paper, it is worth pointing out that our approach shares certain similarities with Földiák's original method [9] and some subsequent approaches. First, our IP mechanism has a similar function as the adaptive threshold regulation in his network. Second, we also utilize a combination of Hebbian and anti-Hebbian weight updates, because the neighborhood function changes the sign of the weight update (positive for most activated unit, negative otherwise). In contrast to Földiák's original method, however, our network does not require adaptable lateral weights between the y-units to function. Thus, our solution is conceptually particularly simple.

4. Modeling the emergence of orientation maps

Receptive fields of simple cells in primary visual cortex (V1) are oriented, localized, and bandpass. In addition, neighboring neurons in V1 will have a similar orientation preference, giving rise to smooth orientation maps. For modelling the emergence of orientation maps, we consider the neurons in our network to be located on a two-dimensional sheet, with neuron i at grid position $(j,k)_i \in \mathbb{N} \times \mathbb{N}$ after the fashion of a self-organizing map (SOM).

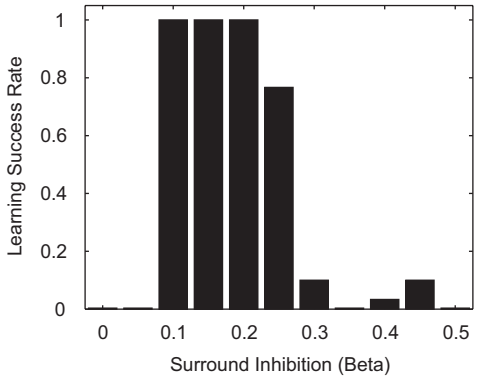

Fig. 3. Fraction of simulations (out of 30) in which a correct representation was learned for various values of β. When β was 0 or 0.05, a correct representation was never learned. When β was 0.1, 0.15, or 0.2, a correct representation was always learned. When β was 0.3 or greater, correct representations were learned only rarely. For typical examples of representations learned in each regime, refer to Fig. 2.

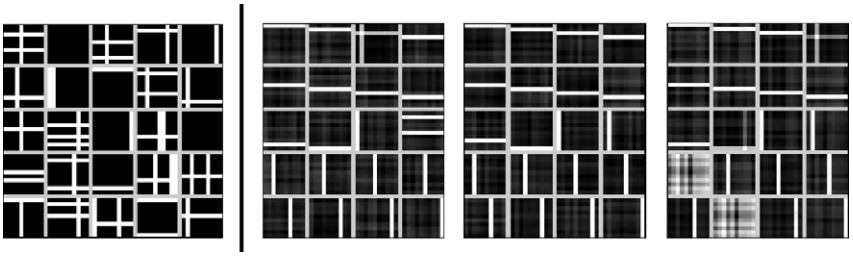

Fig. 2. Left: Example bars stimuli. Stimuli are created by adding bars independently with 0.1 probability. Right: Examples of bars learned when β is too low ($\beta = 0$), just right ($\beta = 0.2$), and too high ($\beta = 0.5$), respectively.

The most active unit exhibits a center-surround influence on learning in its neighbors according to a difference of Gaussians (DoG) neighborhood function centered around it. Let $d_i^2 \equiv (j_i - j_*)^2 + (k_i - k_*)^2$ be the squared distance of neuron i to the most activated unit in the layer at (j_*, k_*). We define

$$\mathcal{N}_{\text{map}}(y_i; \mathbf{y}) = \frac{1}{2\pi\sigma_c^2} \exp\left(\frac{-d_i^2}{2\sigma_c^2}\right) - \frac{1}{2\pi\sigma_s^2} \exp\left(\frac{-d_i^2}{2\sigma_s^2}\right), \qquad (5)$$

where σ_c and σ_s determine the range of the center and surround interaction. In our case, this neighborhood function serves a slightly different role than the Gaussian weighting function usually used in traditional SOMs. The role of $\mathcal{N}$ in our case is short-range cooperation among units combined with a decorrelation of weight updates for units that are less close. Units that are very far away from the winning unit are prevented from learning altogether. This simple mechanism avoids the development of a large amount of redundancy in the learned representation and it facilitates the formation of maps with smoothly varying orientation preference.

4.1. Experiment 1: learning over-complete representations for natural image patches

We trained networks on natural images collected by Van Hateren [27]. We used log-intensity images because these have greater contrast and this transform is performed in the early visual pathway [27]. We convolved the images with a DoG filter to model the center-surround opponency of neurons in the lateral geniculate nucleus (LGN) [19]. For the DoG filter, we used a center width of 1 pixel and a surround width of 1.2 pixels. From each of 375 images, 500 image patches of size 10-by-10 pixels were drawn at random, and were presented once to each neuron in our population (one epoch). The input had positive and negative values simulating populations of ON and OFF cells in the LGN [15]. We used networks of various sizes ranging from 10-by-10 units to 25-by-25 units. Each unit had a 10-by-10 receptive field size, making the populations 1 to 6.25 times over-complete. Parameters were: $\eta_{\text{Hebb}} = 0.05$, $\eta_{\text{IP}} = 0.01$, $\mu = 0.15$, $\sigma_c = 1$, $\sigma_s = 1.5$. Training lasted for 50 epochs each consisting of 3000 image patch presentations for a total of 150,000 natural stimulus presentations.

Typical results of learning are shown in Fig. 4 for networks of three different sizes. Learning was robust to changes in the parameters over a wide range of values. The learned filters are Gabor-like and exhibit a variety of orientations, frequencies, and locations. Moreover, they exhibit smooth interpolation in local regions of the map. This is reminiscent of the orientation-map structure in V1.

We studied the amount of redundancy in the learned representation by measuring the mutual information between all pairs of units in a given network. Here we used a normalized mutual information measure:

$$\text{MI}^*(X, Y) = \frac{2\text{MI}(X, Y)}{\text{H}(X) + \text{H}(Y)}, \qquad (6)$$

where $\text{MI}(X, Y)$ denotes the mutual information between random variables X and Y and $\text{H}(.)$ denotes the entropy. This measure varies between 0 and 1, with 0 indicating independence and 1 indicating maximal dependence of the filter responses. We calculated the average pairwise normalized mutual information by analyzing the empirical firing histograms with six equally spaced bins for networks of different sizes. Fig. 5 plots the average normalized mutual information as a function of the amount of over-completeness of the network. The generally small values of below 0.04, i.e. less than 4% of the maximum possible mutual information, indicate that on average a unit's responses are highly correlated to only a small number of other units. The networks successfully avoid learning many redundant filters, which implies that the network's representation of its input can be considered efficient. Interestingly, as over-completeness increases, the values of the average mutual information actually slightly decrease. This decrease in the per-unit redundancy in over-complete maps

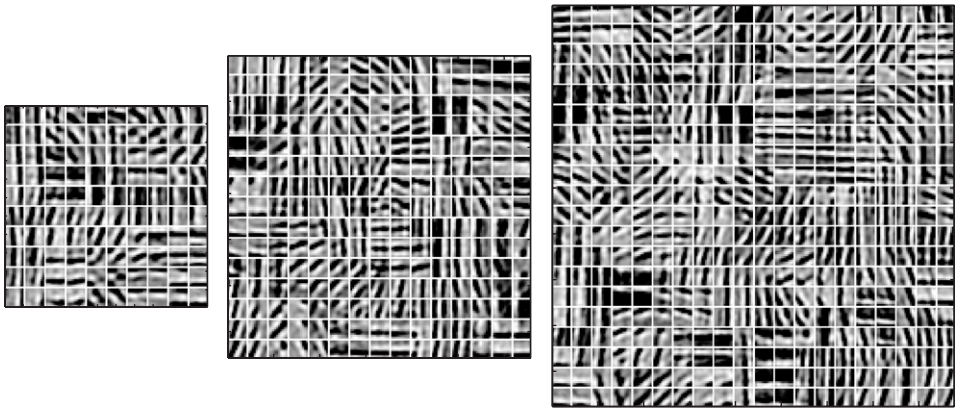

Fig. 4. Receptive fields learned on various map sizes from natural image patches. We plot the set of resulting weight vectors for networks of three sizes. Left: 10-by-10 (100 units, complete); middle: 15-by-15 (225 units, 2.25 times over-complete); right: 20-by-20 (400 units, 4 times over-complete). Parameters were $\eta_{\text{Hebb}} = 0.05$, $\eta_{\text{IP}} = 0.01$, $\mu = 0.15$, $\sigma_c = 1$, $\sigma_s = 1.5$.

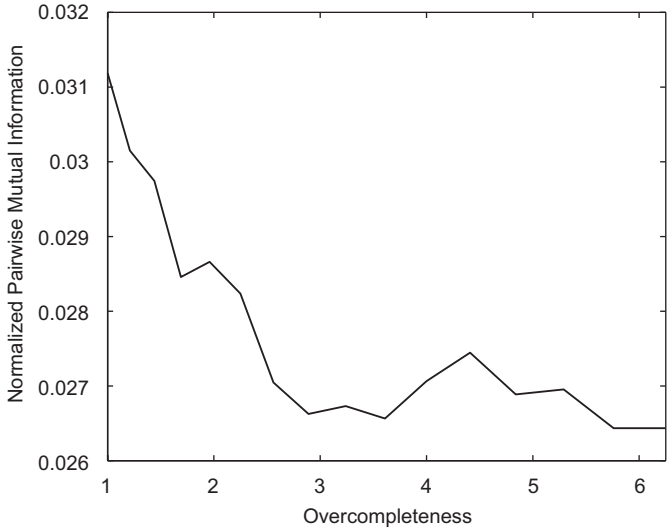

Fig. 5. Average normalized pairwise mutual information between units in networks with different degrees of over-completeness. The generally low values demonstrate that the network successfully avoids learning many redundant filters. Small degrees of over-completeness actually reduce the average pairwise mutual information measure.

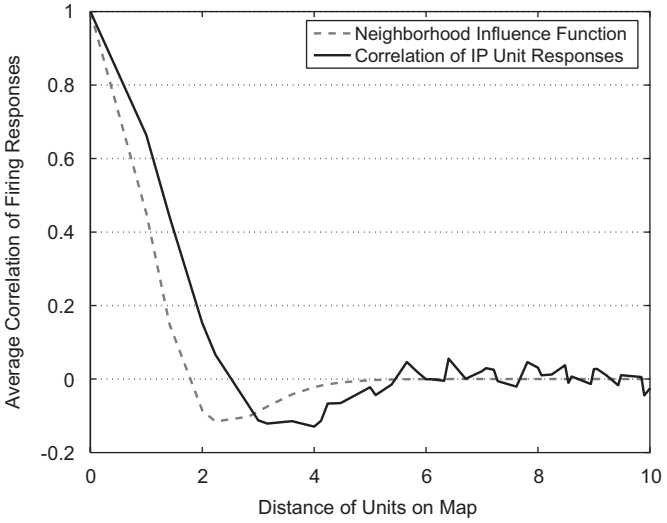

Fig. 6. Average correlation of units' activities as a function of their spatial separation for a network with 15-by-15 units (2.25 times over-complete representation).

implies that as the number of units increases, representation space is covered more evenly and efficiently.

The map-formation mechanism based on the neighborhood function $\mathcal{N}_{\text{map}}$ encourages close neighbors to develop similar weight vectors, making their responses positively correlated, while somewhat more distant units are driven to develop anti-correlated responses. In Fig. 6 we plot the average correlation in the responses of pairs of neurons as a function of their separation for a network with 15-by-15 units. As predicted, close neighbors have positively correlated responses while more distant neurons have anti-correlated responses. Very distant neurons are uncorrelated. This pattern mirrors the shape of the neighborhood function $\mathcal{N}_{\text{map}}$. Thus, the pattern of correlations can be influenced by specific choices of $\mathcal{N}_{\text{map}}$. This result also reflects the low levels of redundancy in the learned representation discussed above.

4.2. Experiment 2: role of IP in the learning process

In order to better understand the role of IP in the learning process, we systematically varied the strength of IP and observed its impact on the learned filters. Since the networks develop units whose receptive fields are similar to Gabor filters, we assessed network performance by measuring how well the learned filters matched Gabor filters—the standard model of V1 simple cell responses—for different learning rates η_{IP}. To this end, we compared each learned filter to a large number of Gabor filters by computing the dot product between the learned filters and standard Gabor filters. All vectors were normalized to unit length, so a dot product of 1 indicates identical vectors and a dot product of 0 indicates orthogonal vectors. The Gabor filters used for comparison covered odd and even symmetry, 100 center locations, six sizes of the Gaussian envelope (ranging from 0.75 to 4.5), 15 values for the spatial frequency (covering the range from 0.03 cycles per pixel up to 0.45 cycles per pixel) and eight different orientations (22.5 degree steps). These values were chosen to fully cover the range of filters learned on 10-by-10 image patches by both our IP model and ICA (see below) [11].

The results are shown in Fig. 7. We compared four conditions: *High IP* and *Low IP* used the method described above with $\eta_{\text{IP}} = 10^{-2}$ and $\eta_{\text{IP}} = 10^{-5}$, respectively. Condition *No IP* used a fixed, non-adaptive sigmoid nonlinearity that was chosen to be $a = 5$ and $b = -2.5$, corresponding to a sigmoid that is roughly linear on the input range 0 to 1. Finally, condition *Linear* used fixed linear units. As shown in Fig. 7 (left panel), condition *High IP* was fastest to obtain Gabor-like receptive fields. Interestingly, however, we found that IP is not strictly necessary to learn Gabor-like receptive fields. Even in conditions *No IP* and *Linear*, Gabor-like receptive fields will develop in the network, but at a dramatically slower rate. This suggests that IP's role in our networks may be primarily to ensure efficient information transmission in individual units and to speed the learning process of the weights, but it does not dramatically alter the resulting weight vectors. The interesting result that somewhat Gabor-like receptive fields even emerge in linear units is caused by the neighborhood function $\mathcal{N}_{\text{map}}$, which forces units to perform anti-Hebbian weight updates whenever the most activated unit is close but not very close to them.

We also measured if and how fast the four different conditions would lead to exponential activity distributions in the units of the network. To this end we measured the marginal activity distributions of individual neurons using a discrete binning with 50 equally spaced bins and compared them to the desired exponential distribution using the L-2 norm. The *High IP* and *Low IP* conditions

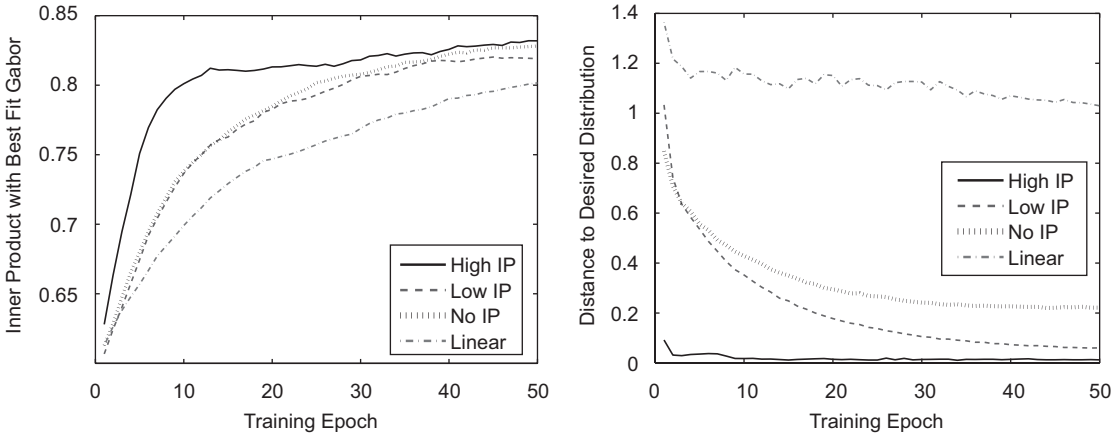

Fig. 7. Dynamics of learning with and without intrinsic plasticity (IP). The left panel plots the average similarity of learned filters to Gabor filters as a function of the number of learning epochs. Similarity to Gabor filters is calculated as the dot product of a filter with its best-fitting Gabor filter. While IP is not necessary to learn Gabor-like receptive fields, it speeds learning substantially. The right panel shows the average similarity of the marginal distribution of filter responses to that of an exponential distribution with the desired mean. With IP, units quickly assume exponential activity distributions. This effect is not observed in linear units and is less pronounced in units with a fixed sigmoidal non-linearity. Each epoch contains 3000 image patch presentations.

produce activity distributions that are very close to exponential—the *High IP* condition achieves this much faster, however. In the *No IP* condition (fixed sigmoidal nonlinearity) the units' activity distributions move closer to an exponential shape as their weight vectors are changing, but the units stop short of exhibiting close-to-exponential activity distributions in their firing patterns. In the *Linear* condition, activity distributions of individual neurons remain very far from exponential distributions.

4.3. Experiment 3: comparison with ICA

In order to better understand the relation of our model to conventional approaches, we compared the population of learned filters with those resulting from ICA. All simulations were done using Hyvärinen and Hoyer's imageica package (http://www.cis.hut.fi/projectsica/imageica/) [11]. We used the ICA algorithm with 100 filters of 10-by-10 pixels. The training set contained 15,000 image patches and we learned for 300 iterations. No extra pre-processing was performed beyond the whitening procedure that is part of this ICA algorithm. Fig. 8 displays the learned receptive fields from the ICA algorithm. As expected, we also observe filters that are localized, bandpass, and oriented, and resemble Gabor filters.

Fig. 8. Set of filters learned by ICA. Each filter has been individually normalized.

Our first analysis aimed to quantify how well-receptive fields learned with our network or with ICA-matched standard Gabor filters. We found the best fitting Gabor filter for each learned receptive field by an exhaustive search over a set of different Gabor filters covering the complete range of learned filters as described in the previous section. These discrete filters were chosen to fully cover the support of the empirical learned-filter distributions that resulted from both the ICA and IP models. We found that changing the number or range of discrete filters did not significantly alter the shape of the resulting histograms, suggesting continuous underlying filter distributions. Learned filters were compared to their best matching Gabor filters by computing the inner product of the two. On average, filters from our network are more similar to Gabor filters than the filters resulting from ICA. The average dot product to the best matching Gabor filter is 0.8921 for the filters in a 15-by-15 network with IP and only 0.7675 for ICA. A possible reason for this poor fit of ICA filters is that they tend to be quite elongated, while we

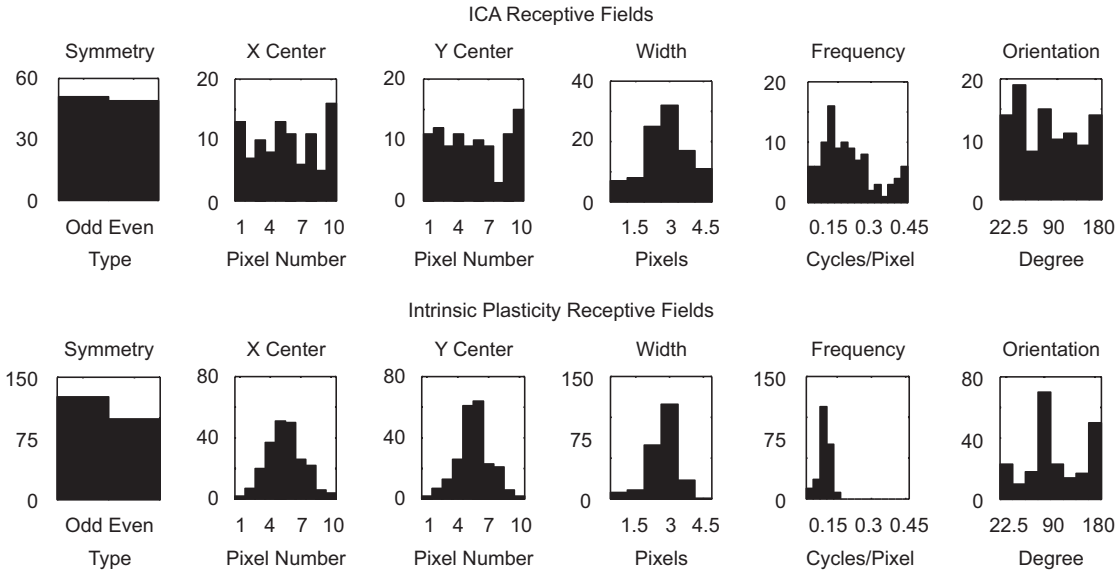

Fig. 9. Comparison of filters learned by our network with those resulting from ICA.

only consider Gabor filters with rotationally symmetric Gaussian envelopes.

Our second analysis considered the variety of different filters learned by the network with IP or by ICA. Fig. 9 shows the distribution of various filter properties in both cases. Generally, the ICA filters tend to exhibit a greater variety along many different dimensions. A point in case is the wider range of spatial frequencies that are covered by the ICA filters. A part of the explanation for this behavior is that while ICA tries to achieve independence between all filters, our simple network merely works to decorrelate the responses of filters that are sufficiently far apart in the layer while close-by units are actually encouraged to develop positively correlated responses.

5. Discussion

Different forms of plasticity are involved in shaping sensory representations in the brain and it is important to understand how these different mechanisms interact. In [24,25] we developed model neurons that maintain sparse lifetime distributions of their individual activities through IP. We showed that when IP is combined with various forms of Hebbian learning at the synapses, a single unit will discover heavy-tailed directions in its input [24,26]. Here we constructed networks of such neurons whose learning was coupled using different neighborhood interaction mechanisms: a direct decorrelation method and an approach facilitating the formation of smooth maps of stimulus preferences. In the former case, we solved the ''bars'' problem, a standard non-linear ICA task, and in the latter we found maps of Gabor-like receptive fields as seen in primary visual cortex when learning on natural image patches. We demonstrated that the IP mechanism, while not being strictly necessary for this behavior, significantly speeds up the learning process. Moreover, the learned representations more closely matched the energy-efficient exponential distributions observed in cortical firing, which have both information maximizing and sparse coding properties. When comparing the learned filters in the network to those resulting from ICA, we found that our filters (a) provide a closer match to standard Gabor filters and (b) are automatically arranged on a smooth map. The learned filters are not as independent as those learned via ICA because significant correlations between neighboring units are introduced, which is biologically plausible, however.

Our simple model is able to learn Gabor-like receptive fields from natural images and arranges the filters into smooth maps. A number of previous models (both mechanistic and functional ones) have demonstrated similar results. Among them are models based on extensions to the self-organizing map framework [18], BCM-based models [5], topographic ICA [12], extensions to sparse coding approaches [29], and others. What distinguishes our model from these earlier ones, is that it utilizes an IP mechanism to obtain energy efficient coding, directly ensuring approximately exponential activity distributions in the networks' units. In addition, we demonstrated that the IP mechanism contributes to rapid learning in the network. Overall, our results suggest that IP may play an important role in the unsupervised learning of sensory representations in the cortex and it underscores the need to carefully study how different forms of neuronal plasticity may interact at the network level.

In the model of visual receptive field development we have used the simple Hebbian learning rule which was multiplied with a difference of Gaussian function that

modulated the sign of Hebbian learning (Hebbian vs. anti-Hebbian) based on a unit's distance to the most activated unit in the map. This implies that units close to the winning unit will strengthen their connections (long term potentiation, LTP) while far away units will weaken their connections (long term depression, LTD). Note that a qualitatively similar effect could be obtained by using a Bienenstock–Cooper–Munro (BCM) learning rule that has LTP and LTD components [5], combined with only an excitatory Gaussian neighborhood function. In future work we would like to explore such alternative learning schemes and also consider the combination with neural fields described by Wilson and Cowan-like dynamics. In addition we would like to construct hierarchical networks to model the development of receptive field properties in higher visual areas.

Acknowledgments

The authors thank Cornelius Weber, Erik Murphy-Chutorian, and three anonymous reviewers for comments on earlier drafts. This work was supported by the Hertie foundation.

References

[1] F. Attneave, Some informational aspects of visual perception, Psychol. Rev. 61 (1954) 183–193.
[2] R. Baddeley, L.F. Abbott, M.C. Booth, F. Sengpiel, T. Freeman, E.A. Wakeman, E.T. Rolls, Responses of neurons in primary and inferior temporal visual cortices to natural scenes, Proc. R. Soc. London B 264 (1998) 1775–1783.
[3] H.B. Barlow, Possible principles underlying the transformation of sensory messages, in: W.A. Rosenblith (Ed.), Sensory Communication, MIT Press, Cambridge, MA, 1961, pp. 217–234.
[4] A.J. Bell, T.J. Sejnowski, The independent components of scenes are edge filters, Vision Res. 37 (23) (1997) 3327–3338.
[5] L.N. Cooper, N. Intrator, B.S. Blais, H.Z. Shouval, Theory of Cortical Plasticity, World Scientific, London, 2004.
[6] N.S. Desai, L.C. Rutherford, G. Turrigiano, Plasticity in the intrinsic excitability of cortical pyramidal neurons, Nature Neurosci. 2 (6) (1999) 515–520.
[7] D. DeSieno, Adding a conscience to competitive learning, IEEE Proceedings of the International Conference on Neural Networks, vol. I, 1988, pp. 117–124.
[8] M.S. Falconbridge, R.L. Stamps, D.R. Badcock, A simple Hebbian/anti-Hebbian network learns the sparse, independent components of natural images, Neural Comput. 18 (2005) 415–429.
[9] P. Földiák, Forming sparse representation by local anti-hebbian learning, Biol. Cybern. 64 (1990) 165–170.
[10] P. Földiák, Sparse coding in the primate cortex, in: M.A. Arbib (Ed.), The Handbook of Brain Theory and Neural Networks, second ed., MIT Press, Cambridge, MA, 2002.
[11] A. Hyvärinen, P. Hoyer, Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces, Neural Comput. 12 (7) (2000) 1705–1720.
[12] A. Hyvärinen, P.O. Hoyer, M. Inki, Topographic independent component analysis, Neural Comput. 13 (2001) 1527–1558.
[13] S. Laughlin, A simple coding procedure enhances a neuron's information capacity, Z. Naturforsch 36 (1981) 910–912.
[14] P. Lennie, The cost of cortical computation, Curr. Biol. 13 (2003) 493–497.
[15] R. Linsker, From basic network principles to neural architecture: emergence of oriented columns, Proceedings of the National Academies of Sciences, vol. 83, 1986, pp. 8779–8783.
[16] R. Linsker, Self-Organization in a perceptual network, Computer 21 (3) (1988) 105–117.
[17] J. Lücke, C. von der Malsburg, Rapid processing and unsupervised learning in a model of the cortical macrocolumn, Neural Comput. 16 (3) (2004) 501–533.
[18] R. Miikkulainen, J. Bednar, Y. Choe, J. Sirosh, Computational Maps in the Visual Cortex, Springer, Berlin, 2005.
[19] K.D. Miller, A model for the development of simple cell receptive fields and the ordered arrangement of orientation columns through activity-dependent competition between ON- and OFF-center inputs, J. Neurosci. 14 (1) (1994) 409–441.
[20] J.P. Nadal, N. Parga, Nonlinear neurons in the low-noise limit:a factorial code maximizes information transfer, Network Comput. Neural Syst. 5 (4) (1994) 565–581.
[21] B.A. Olshausen, D.J. Field, Sparse coding with an overcomplete basis set: a strategy employed by V1?, Vision Res. 37 (23) (1997) 3311–3325.
[22] E.P. Simoncelli, B.A. Olshausen, Natural image statistics and neural representation, Annu. Rev. Neurosci. 24 (2001) 1193–1216.
[23] M. Stemmler, C. Koch, How voltage-dependent conductances can adapt to maximize the information encoded by neuronal firing rate, Nature Neurosci. 2 (1999) 521–527.
[24] J. Triesch, Synergies between intrinsic and synaptic plasticity in individual neurons, Adv. Neural Inf. Process. Syst. 17 (2005) 1417–1424.
[25] J. Triesch, A gradient rule for the plasticity of a neuron's intrinsic excitability, Proceedings of the International Conference on Artificial Neural Networks, 2005, pp. 65–70.
[26] J. Triesch, Synergies between intrinsic and synaptic plasticity mechanisms, Neural Comput. 2006, in press.
[27] J.H. van Hateren, A. van der Schaaf, Independent component filters of natural images comparsed with simple cells in primary visual cortex, Proc. R. Soc. London B 265 (1998) 359–366.
[28] B.T. Vincent, R.J. Baddeley, T. Troscianko, I.D. Gilchrist, Is the early visual system optimised to be energy efficient?, Network Comput. Neural Syst. 16 (2–3) (2005) 175–190.
[29] C. Weber, Self-organization of orientation maps, lateral connections, and dynamic receptive fields in the primary visual cortex, Proceedings of the International Conference on Artificial Neural Networks, 2001, pp. 1147–1152.
[30] W. Zhang, D.J. Linden, The other side of the engram: experience-dependent changes in neuronal intrinsic excitability, Nature Rev. Neurosci. 4 (2003) 885–900.

Nicholas Butko is a graduate student in the department of Cognitive Science at the University of California in San Diego. His research interest is developing novel algorithms to allow machines to perceive and act in real-world situations.

Jochen Triesch is Assistant Professor of Cognitive Science at UC San Diego and a Fellow of the Frankfurt Institute for Advanced Studies. His research interests span neural computation, human and machine vision, cognitive robotics, and models of human cognitive development.