A Network For Computing Topographic Infomax With An Overcomplete Basis: Correlates With The Neocortical Microcircuit

James Kozloski, Guillermo A. Cecchi, Charles C. Peck, and A. Ravishankar Rao
IBM T.J. Watson Research Center, Foreknown Heights, NY 10598

November 5, 2006

Abstract

We derive a novel neural multigrid, which employs local learning rules, simple activation functions, and feedback to maximize mutual information between inputs and outputs. The network is a modified *Linsker network* [1], which implements Bell and Sejnowski's infomax algorithm [2]. It differs from a Linsker network in its output layer, which can be larger than the input layer and thus overcomplete, and in its multigrid auxiliary layers, comprising feed-forward, lateral, and feedback connections, which create a topographic influence. Beyond its information theoretic grounding, the network demonstrates structural and functional correlates with the neocortical microcircuit. In particular, it succeeds in embedding an infomax solution in an ordered, 2-D topographic map.

1 Introduction

The primary visual cortex of primates and carnivores shows two distinct forms of self-organization: first, receptive fields organize to resemble edge-filters, and second, these filters are organized in a topographic map. The problem of topographic map formation can be characterized as one of order-embedding, in which a set of input vectors X is mapped onto a set of output vectors Y such that the partial ordering of outputs in their space, when embedded in a 2-D coordinate system, preserves the partial ordering of inputs in their space. An important additional (but often unstated) mapping objective is that the volume defined by Y into which points in X are mapped is maximized. This objective ignores the ordering of inputs and outputs and instead maximizes the mutual information between X and Y, $I(X;Y)$. We propose that visual cortex performs infomax using a neural multigrid, thus generating optimal edge-selective receptive fields and, at the same time, a topographic map.

Originally derived for multi-variate Gaussian inputs [3], infomax was then extended to accommodate input distributions of arbitrary shape [2]. Both derivations required critically sampled bases (wherein the number of outputs equals the number of inputs). While subsequent derivations [4] and related sparse-coding strategies [5] support overcomplete bases, none to our knowledge do so while generating a topographic map. Finally, algorithms that maximize the Shannon information rate subject to a topographic constraint [6, 7] generate ordered maps but, to our knowledge, do not achieve rates equal to what is achieved by standard infomax. Because the main operational principle of infomax is to make outputs uncorrelated, standard topographic mapping algorithms [8], which *induce* correlations between map neighbors, are inconsistent with infomax.

Here we present a network that performs infomax over an arbitrary input space, using critically sampled or overcomplete bases, while creating a topographic map. The network incorporates a novel neural multigrid, configured to estimate Fourier modes of a key infomax learning vector using feed-forward, lateral, and feedback connections. Strong spatial correlations are introduced by the configuration of the neural multigrid, which emphasizes low frequency modes of the learning vector.

As higher frequency modes are emphasized, infomax emerges and removes redundancy in the output without destroying global map structure. We observe that this network resembles the neocortical microcircuit in its structure, function, and principles of self-organization.

2 A Three-Stage Infomax Network

The network comprises three stages, which implement a modified Linsker network for information maximization [1]. Stage one selects a vector $\widehat{x}$ from the input ensemble, $\widehat{x} \in X$, and computes the input vector $x = q^{-1/2}(\widehat{x} - x_0)$, where $x_0 = \langle \widehat{x} \rangle$, $q = \langle (\widehat{x} - x_0)(\widehat{x} - x_0)' \rangle$, and $q^{-1/2}$ is the input whitening matrix. For results shown here, X was a set of image segments drawn randomly from a published set of natural images [9]. Stage two learns the input weight matrix C and computes $u \equiv Cx$. In addition, each stage two unit computes an element of the output vector y, $y_i = \sigma(u_i)$, where $\sigma(\cdot)$ denotes a nonlinear squashing function, here the logistic transfer function [2]: $y = 1/1 + e^{-(u+w_0)}$, where w_0 is an adaptive output bias vector, and $\Delta w_0 = \beta_{w_0}[1 - 2y]$.[1] The ensemble of all output vectors is then Y, and the objective to maximize $I(X;Y)$.

The outputs of stage three comprise a learning vector with the same dimension as u. When applied in stage two to learning C, this learning vector yields the anti-redundancy term, $(C')^{-1}$, of Bell and Sejnowski [2]. We refer to this vector (and its derivatives in the next section) as the *anti-redundancy learning vector*, hereafter denoted ψ. In a Linsker network, stage three comprises a single layer, wherein each unit i computes element ψ_i over a lateral weight matrix $\widehat{Q}$, whose elements undergo Hebbian learning according to $\Delta\widehat{Q} = \beta_Q[uu' - \widehat{Q}]$,[2] such that $\widehat{Q} \to Q \equiv \langle uu' \rangle$. For a given input presentation, lateral connections modify elements of an auxiliary vector v according to $v_t = v_{t-1} + u - \alpha\widehat{Q}v_{t-1}$. Regardless of initial v, and assuming $\widehat{Q} = Q$ and the scalar α is chosen so that v converges,[3] by Jacobi $\alpha v_\infty = Q^{-1}u$. Linsker showed that the infomax anti-redundancy term can be rewritten, $(C')^{-1} = Q^{-1}C\langle xx' \rangle$, and, by substitution: $(C')^{-1} = \alpha\langle v_\infty x' \rangle$. In practice, a finite number of iterations are sufficient to approximate αv_∞,[4] and therefore anti-redundancy learning for C_{ij} depends only on the locally computed element, $\psi_i = \alpha v_i$, and the local input, x_j. The infomax learning rule for the network is then $\Delta C = \beta_C[(\psi + 1 - 2y)x']$ [5] [1]. Each output of a Linsker network learns an expected infomax edge filter when trained using natural images, a fixed number of Jacobi iterations, and constant learning rates (Fig. 1A).

3 A Neural Multigrid Yields Topographic Infomax

We explored using multigrid methods to estimate the antiredundancy vector, ψ. Multigrid speeds convergence and accuracy of Jacobi iteration by decomposing it into iterative computations performed over a set of grids, each solving different Fourier modes of the problem [10]. The multigrid method we implement here in a neural network is nested iteration, though the network design can easily accommodate other multigrid methods such as "V-cycle" and "Full Multigrid" [10]. The set of grids over which the computation of ψ is performed, h_k, is enumerated by the set of wavelengths of the Fourier modes of the problem that each solves, for example $k \in \{2, 4, 8\}$. The iterative computation performed in each grid is similar to that in a Linsker network, now denoted

[1] Note that all learning rates in the network are constant, and denoted by β with subscript. We used $\beta_{w_0} = 0.0021$.

[2] We used $\beta_Q = 0.0007$.

[3] The convergence criterion of v is $0 < \alpha < 2/Q_+$, where Q_+ is the largest eigenvalue of Q [1]. In Linsker's network, α is computed by a nonlocal heuristic [3]. We have devised a dynamic, local computation of Q_+ based on power iteration: Let e represent an activity vector propagated through the lateral network $\widehat{Q}$ in the absence of the normal stage three forcing term, $v_t + u$, such that $e_t = -\alpha\widehat{Q}e_{t-1}$. Precalculating $\alpha = 1/\|e_t\|$ for each t ensures $\|e\| \to Q_+$ and $\alpha \to 1/Q_+$, thus satisfying the convergence criterion of v.

[4] For most problems, we found 4 Jacobi iterations to be sufficient.

[5] We used $\beta_C = 0.0021$.

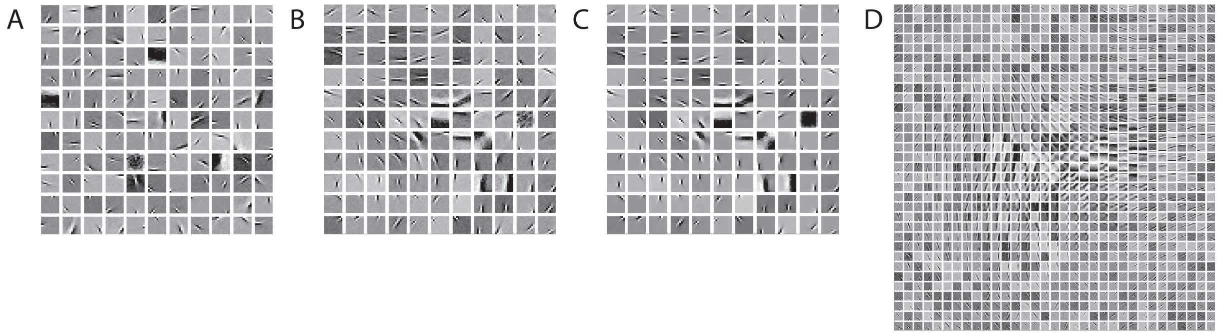

Figure 1: Learned weights displayed as receptive fields of simulated cortical units. A: Linsker network infomax; B: Scheduled multigrid, topographic infomax; C: Weights from C, modified only slightly by standard ICA, D: Weighted multigrid, overcomplete, topographic infomax.

$v_t^{h_n} = v_{t-1}^{h_n} + u^{h_n} - \alpha^{h_n}\widehat{Q}^{h_n}v_t^{h_n}, n \in k$. We chose neural multigrid wavelengths such that if the Linsker network and grid h_2 is an 11×11 layer, h_4 is a 5×5 layer, and h_8 a 2×2 layer.

Feed-forward connections propagate and restrict each u^{h_n} to each grid h_{2n}, such that $u^{h_{2n}} = S^{h_n}u^{h_n} \forall n \in k$, where S^{h_n} denotes the restriction operator (in our neural multigrid, a rectangular feed-forward weight matrix derived from a binomial filter). Q^{h_n} is restricted by $Q^{h_{2n}} = S^{h_n}Q^{h_n}S^{h_n\prime}$ [10]. In our network implementation, we derive $Q^{h_{2n}} = S^{h_n}\langle u^{h_n}u^{h_n\prime}\rangle S^{h_n\prime} = \langle S^{h_n}u^{h_n}u^{h_n\prime}S^{h_n\prime}\rangle = \langle u^{h_{2n}}u^{h_{2n}\prime}\rangle$. Thus, $Q_{h_{2n}}$ is computed locally in each grid by Hebbian learning over a lateral weight matrix $\widehat{Q}^{h_{2n}}$ such that $\widehat{Q}^{h_{2n}} \rightarrow Q^{h_{2n}} \equiv \langle u^{h_{2n}}u^{h_{2n}\prime}\rangle$. In a standard neural multigrid, feedback smoothly interpolates the result of coarse grid iteration, $\alpha^{h_{2n}}v^{h_{2n}}$, to the next finer grid, where it replaces v^{h_n} prior to Jacobi iteration within the finer grid: $v_0^{h_n} \leftarrow S^{h_n\prime}\alpha^{h_{2n}}v^{h_{2n}}$. In this way, higher frequency mode iteration refines the solution provided by lower frequency mode iteration. The process continues until $\alpha^{h_2}v^{h_2}$ is fed back to the Linsker network, where iteration begins with the multigrid feedback, $v_0 \leftarrow S^{h_2}\alpha^{h_2}v^{h_2}$, and proceeds to compute ψ.

Two separate topographic influences were explored using the neural multigrid. First, a scheduled neural multigrid iterates only at the two coarsest grids, with m inputs presented to the network in this configuration.[6] Units in the Linsker network learn based on a smooth anti-redundancy learning vector fed back from the coarsest grids and interpolated by every other multigrid layer. Finer layers in the neural multigrid are activated at intervals of m input presentations; the number of input presentations for which a grid h_n has been active is p^{h_n}, and ψ is computed in the partial multigrid as a linear combination of the feedback vectors from the two finest active grids, h_a and h_b, fed back through all intervening layers, $\psi = S\prod_{n=1}^{a} S^{h_n\prime}[\beta^{h_a}\alpha^{h_a}v^{h_a} + (1-\beta^{h_a})S^{h_b\prime}\alpha^{h_b}v^{h_b}]$, where $\beta^{h_a} = p^{h_a}/m \in [0,1]$. We explored a second topographic influence using a weighted neural multigrid, in which iteration proceeds in all grids as in a standard neural multigrid, but feedback from every grid is a linear combination of the result of iteration in that grid and the feedback received from the previous, coarser grid: $v_0^{h_n} \leftarrow S^{h_{2n}\prime}[\beta^{h_{2n}}\alpha^{h_{2n}}v^{h_{2n}} + (1-\beta^{h_{2n}})S^{h_{4n}\prime}\alpha^{h_{4n}}v^{h_{4n}}]$.[7]

Our aim was to create a phase-independent order embedding of outputs within a topographic map, as has been observed in visual cortex. Drawing upon Hyvärinen et al. [9], we reasoned that certain nonlinear transformations of the inputs to our neural multigrid would yield a phase independent embedding. We full-wave rectify multigrid inputs, such that $u^{h_2} = S|u|$. Given this transformation, feedback from the multigrid to the Linsker network also requires modification for anti-redundancy learning. At each unit i, the multigrid feedback vector is multiplied by ω_i such

[6] We used $m = 2,000,000$.

[7] We used $\beta^{h_2} = 0.01$ and $\beta^{h_4} = 0.1$.

that $v_{i0} \leftarrow \omega_i \sum_k S'_{ik} \alpha^{h_2} v_k^{h_2}$, where ω_i is 1 if the element u_i is positive, and -1 otherwise.

Scheduled incorporation of each grid of the neural multigrid resulted in an infomax solution embedded in a topographic map (Fig. 1B). We then "transplanted" the weight matrix, C, and bias vector, w_0, from this solution into a standard Bell and Sejnowski infomax algorithm. After presenting $500,000$ additional inputs to this system, the learned bases were modified only slightly (Fig. 1C), indicating that our algorithm had achieved topographic infomax.

Weighting the neural multigrid also yielded a topographic influence. For these experiments, we computed an anti-redundancy learning vector based on pooled outputs from nine separate critically sampled bases, each initially embedded in an 11×11 grid as above, then, for each coordinate in this grid, (x, y), into a single 33×33 *overcomplete* grid as follows: $x \leftarrow r + 3x$, $y \leftarrow s + 3y$, where (r, s) represents a unique pair of offsets applied to each 11×11 grid's corresponding coordinates, $r \in [0, 2]$, $s \in [0, 2]$. The restriction of the output of the co-embedded Linsker networks into the neural multigrid's first layer was performed with an S matrix scaled proportionally to accommodate the larger overcomplete input grid. Each lateral network included only connections between those units comprising a single Linsker network, and the overcomplete grid's lateral network thus comprised overlapping, periodic, lateral connections. The result of simulations using a weighted topographic influence and this overcomplete basis demonstrate that each critically sampled basis can be co-embedded into a single topographic map (Fig. 1D).

4 Comparisons with the Neocortical Microcircuit

We observe that the multilevel, multiscale networks described here for performing topographic infomax resembles the neocortical microcircuit in the following ways:

1) They computes the anti-redundancy learning vector by Jacobi iteration in a lateral network and apply this vector to Hebbian learning of input weights. A correlate of these properties in the neocortical microcircuit is dense lateral connectivity in layer 2/3, and feedback from 2/3 to 4.

2) Their dynamics are self-stabilized by the multiplicative gain, α, computed locally by power iteration. Correlates of this operation in the neocortical microcircuit include spontaneous activity and shunting inhibition.

3) They cast Jacobi iteration into an auxiliary neural multigrid. A correlate of this process is transient gap junctional coupling within the juvenile cortical microcircuit, which we propose as a means for juvenile layer 2/3 to emphasize low spatial frequency modes of the anti-redundancy learning vector during development.

4) They use full-wave rectification of outputs from the Linsker network to force the map into a phase-independent order-embedding of spatial frequency, retinal location, and orientation. These properties match those of primary visual cortex, and support a previously hypothesized role for complex cells in generating a topographic map [9], here in a network that maximizes information between inputs and the simple cell layer.

5) They can generate topographically mapped, overcomplete bases. The overcomplete simple-cell layer has network structural correlates with primate V1, comprising independent, overlapping lateral networks connected at regular, periodical intervals.

We conclude that a plausible explanation for the circuit architecture, functional properties, and development of visual cortex is that it implements a neural multigrid for information maximization.

References

[1] Linsker, R. (1997) "A local learning rule that enables information maximization for arbitrary input distributions," Neural Computation, 9:1661-1665.

[2] Bell, A. J. & Sejnowski, T. J. (1995) ”An information-maximisation approach to blind separation and blind deconvolution,” Neural Computation, 7:1129-1159.

[3] Linsker, R. (1992) ”Local synaptic learning rules suffice to maximise mutual information in a linear network,” Neural Computation, 4:691-702.

[4] Shriki, O., Sompolinsky H. & Lee D. D. (2000) ”An information maximization approach to overcomplete and recurrent representations”, 12th Conference on Neural Information Processing Systems, pp. 87-93.

[5] Olshausen, B. A. & Field, D. J. (1996) ”Sparse coding with an overcomplete basis set: A strategy employed by V1?,” Vision Research, 37:3311-3325.

[6] Linsker R. (1989) ”How to generate ordered maps by maximizing the mutual information between input and output signals”, Neural Computation, 1:402-411.

[7] Van Hulle, M. M. (1997) ”The formation of topographic maps that maximize the average mutual information of the output responses noiseless input signals,” Neural Computation, 9:595-606.

[8] Kohonen, T. (1997), Self-Organizing Maps, Berlin: Springer-Verlag.

[9] Hyvärinen, A., Hoyer, P. O. & Inki, M. (2001) ”Topographic independent component analysis,” Neural Computation, 13:1527-1528.

[10] Briggs W. L., Henson V. E. & McCormick S. F., A Multigrid Tutorial, Phildelphia, PA: Society for Industrial and Applied Mathematics.