Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada
July 16-21, 2006

Information Rate Maximization over a Resistive Grid

H. Koeppl

***Abstract*—The work presents the first results of the authors research on adaptive Cellular Neural Networks (CNN) based on a global information theoretic cost-function. It considers the simplest case of optimizing a resistive grid such that the Shannon information rate across the input-output boundaries of the grid is maximized. Besides its importance in information theory, information rate has been proven to be a useful concept for principal as well independent component analysis (PCA, ICA). In contrast to linear fully connected neural networks, resistive grids due to their local coupling can resemble models of physical media and are feasible for a VLSI implementation. Results for spatially invariant as well as for the spatially variant case are presented and their relation to principal subspace analysis (PSA) is outlined. Simulation results show the validity of the proposed results.**

I. Introduction

The big picture of this ongoing research is the study and design of local learning algorithms for the cell parameters $\boldsymbol{\alpha}$ of a CNN [1]. After the adaptation the collective dynamics of the cells should obey a global information theoretic optimality criterion J which in general is a function of the input and the output signal. For a planar lattice of cells the situation is illustrated in Fig. 1. This field of research touches upon methods of artificial intelligence such as multi-agent systems [2] and collective intelligence [3]. It is also related to artificial neural networks, and especially for this result, to the PCA and ICA neural networks [4]. Furthermore, synergies are expected from the fields of decentralized control [5] and smart matter research [6].

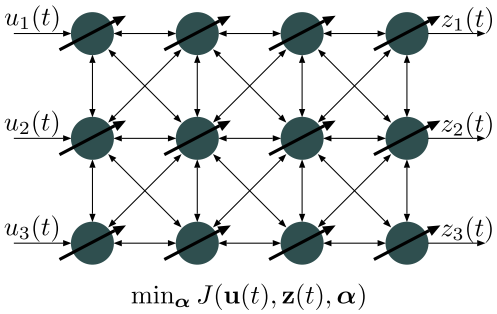
$u_1(t)$
$u_2(t)$
$u_3(t)$
$z_1(t)$
$z_2(t)$
$z_3(t)$
$\min_{\boldsymbol{\alpha}} J(\mathbf{u}(t), \mathbf{z}(t), \boldsymbol{\alpha})$

Fig. 1. Adaptive CNN, with local identical learning algorithms for each cell (indicated by an arrow); the CNN parameters $\boldsymbol{\alpha}$ are adjusted according to some global optimality criterion J.

In this work focus is put on the study of the simplest case where the lattice is given as a planar resistive grid. The objective of the network is to tune its parameters $\boldsymbol{\alpha}$ such that the information rate or the mutual information [7] between the input $\mathbf{u} \equiv [u_1, \ldots, u_N]^T$ and its in general distorted output signal $\mathbf{z} \equiv [z_1, \ldots, z_M]^T$ is maximized. It was shown [8] that for a fully connected one-layer feed-forward neural network with linear activation the optimal network configuration corresponding to this optimization problem is the PCA or PSA network. The exploding wiring complexity of fully connected networks results in serious problems for a VLSI implementation of such networks. Thus, a natural question to ask is whether it is possible to perform PCA or PSA with a network that has local connectivity but multiple layers. It is well known that networks with local connectivity such as CNNs are very suitable for a VLSI implementation [1]. On the other hand CNNs have been successfully applied to model distributed physical systems governed by partial differential equations. Thus, another naturally arising question is whether one can design physical distributed spatially-invariant or spatially-variant systems that process input signals such that their output corresponds to the result of a continuous PCA or PSA analysis.

Throughout the work it is assumed that the perturbation of $\mathbf{z}$ is given as $\mathbf{z} = \mathbf{y} + \boldsymbol{\epsilon}$ with the deterministic output of the grid $\mathbf{y} \equiv [y_1, \ldots, y_M]^T$ and the additive perturbation $\boldsymbol{\epsilon} \equiv [\epsilon_1, \ldots, \epsilon_M]^T$. In contrast to information theory which tries to maximize the information rate over a given communication channel subject to different channel coding schemes, our attempt is to adapt the channel, i.e., the parameters of the resistive grid, to maximize the information rate.

Initially in [9] the InfoMax principle, i.e., the principle of maximum information preservation has been proposed to model orientation selective cells in the mammalian visual cortex [10]. Based on these results the popular InfoMax ICA algorithm was developed in [11]. In [8] the application of this principle to linear feedforward networks is considered. While most of the subsequent work deals with the derivation of the InfoMax principle for the resistive grid and does not culminate in a local learning algorithm, section VI goes all the way to a local algorithm for a particular grid structure. The applied structure is based on results in [12].

The work is organized as follows. In section II the considered processing system is defined and its input-output (i/o) relation is given. Section III introduces the mutual information or information rate. Results for the spatially invariant resistive grid is given in section IV, while section V presents the result for the spatially invariant situation. A special case of a spatially variant network, which performs exact PCA is proposed in section VI. Simulation results are presented in section VII and section VIII draws the conclusions.

H. Koeppl is a Erwin Schrödinger postdoctoral fellow at the Department of Electrical Engineering and Computer Sciences, University of California Berkeley, CA 94720-1770, USA (email: heinz.koeppl@berkeley.edu).

0-7803-9490-9/06/$20.00/©2006 IEEE

II. The Resistive Grid

Following Fig. 1 we define a layer k of the resistive grid [13] as the collection of cells which are $k-1$ cells away from the input boundary. Thus, the cells at the input boundary resembles layer $k = 1$ while the output layer denotes layer $k = L$. The dynamics for the cell at position (l, k) at layer $1 < k \leq L$ of the resistive grid is described as

$$\dot{x}(l,k,t) = \sum_{r=-R}^{R} \sum_{s=-R}^{R} \bar{A}(l,k,r,s)x(l-r,k-s,t), \tag{1}$$

with the appropriate boundary and initial conditions and R the Moore neighborhood size. The output signal $\mathbf{y}$ at layer $k = L$ computes to $y_l(t) = x(l, L, t)$. For $k = 1$ we have

$$\begin{aligned}\dot{x}(l,1,t) = \sum_{r=-R}^{R} \sum_{s=-R}^{R} \bar{A}(l,1,r,s)x(l-r,1-s,t) \\ + \sum_{r=-R}^{R} \bar{B}(l,r)u(l-r,t).\end{aligned} \tag{2}$$

The following assumptions are made for the sake of clarity of the presentation. Only the nearest neighbor coupling $R = 1$ is considered and Dirichlet boundary conditions are assumed for the input and output boundary, while periodic (case I) as well as Dirichlet conditions (case II) are considered for the remaining two boundaries. Some of the subsequent results can be generalized by suspending these assumptions. Introducing the collection of cell states of layer k as $\mathbf{x}_k(t) = [\,x(1,k,t), \ldots, x(1,N,t)\,]^T$ and defining $\mathbf{x}(t) \equiv [\,\mathbf{x}_1^T(t), \ldots, \mathbf{x}_L^T(t)\,]^T$ the dynamics of the grid can be cast into the standard form of a linear control system

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) &= \mathbf{\Xi}\mathbf{x}(t),\end{aligned} \tag{3}$$

with $\mathbf{x}(0) = \mathbf{x}_0$. The involved matrices are block matrices of the following form

$$\mathbf{A} \equiv \begin{pmatrix} \mathbf{C}_1 & \mathbf{N}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{P}_2 & \mathbf{C}_2 & \mathbf{N}_2 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_3 & \mathbf{C}_3 & \mathbf{N}_3 & \cdots & \mathbf{0} \\ \vdots & & & & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{P}_L & \mathbf{C}_L \end{pmatrix}, \tag{4}$$

where the submatrices $\mathbf{P}_k$, $\mathbf{C}_k$ and $\mathbf{N}_k$ refers to the coupling matrices at layer k with the previous $(k-1)$, the current (k) and the next $(k+1)$ layer, respectively. The structure of these matrices are exemplified for $\mathbf{P}_k$ subsequently. The (p, q)-th element of the submatrix $\mathbf{P}_k$ is

$$(\mathbf{P}_k)_{pq} \equiv \begin{cases} \bar{A}(p,k,1,q-p) & \text{for} \quad |q-p| \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

for the case II and

$$(\mathbf{P}_k)_{pq} \equiv \begin{cases} \bar{A}(p,k,1,q-p) & \text{for} \quad N\text{mod}(|q-p|) \leq 1 \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

for the case I. The corresponding submatrices $\mathbf{C}_k$ and $\mathbf{N}_k$ are obtained by replacing the "1" in (5) and (6) by "0" and "-1", respectively. The $LN \times N$ input matrix $\mathbf{B}$ of (3) is $\mathbf{B} \equiv [\,\mathbf{\Theta}, \mathbf{0}\,]^T$ with the $N \times N$ submatrix

$$\Theta_{pq} \equiv \begin{cases} \bar{B}(p,q-p) & \text{for} \quad |q-p| \leq 1 \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

for the case II and

$$\Theta_{pq} \equiv \begin{cases} \bar{B}(p,q-p) & \text{for} \quad N\text{mod}(|q-p|) \leq 1 \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

for the case I. The $M \times LN$ output matrix $\mathbf{\Xi}$ of (3) becomes to $\mathbf{\Xi} \equiv [\,\mathbf{0}, \mathbf{\Gamma}\,]$ with the $M \times N$ auxiliary matrix $\mathbf{\Gamma} = [\mathbf{0}, \mathbf{I}_M]$ and with the M-dimensional unit matrix $\mathbf{I}_M$. This complication arising through the fact of different input and output dimensions is necessary to study situations where the resistive grid should perform a dimension reduction, i.e., a PCA. The setup for $M < N$ is shown in Fig. 2.

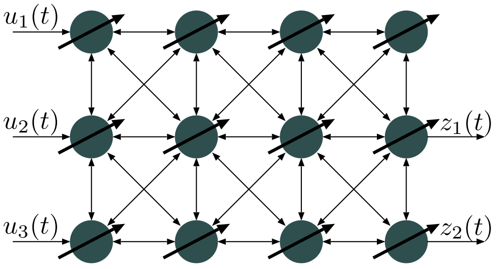

Fig. 2. Dimension reduction $M < N$ of the adaptive resistive grid, for the purpose of principal component analysis.

For $\mathbf{A}$ a Hurwitz matrix the asymptotic output $\mathbf{y} \equiv \lim_{t\to\infty} \mathbf{y}(t)$ for a constant input with $\mathbf{u}(t) = \mathbf{u}$ for all t computes to $\mathbf{y} = \mathbf{W}\mathbf{u}$ with the $M \times N$ weight matrix $\mathbf{W} \equiv -\mathbf{\Xi}\mathbf{A}^{-1}\mathbf{B}$. In terms of the coupling matrices $\mathbf{P}_k$, $\mathbf{C}_k$ and $\mathbf{N}_k$ the weight matrix $\mathbf{W}$ turns out to be

$$\mathbf{W} = (-1)^L \mathbf{\Gamma}\mathbf{S}_L^{-1}\mathbf{P}_L\mathbf{S}_{L-1}^{-1}\mathbf{P}_{L-1} \cdots \mathbf{P}_2\mathbf{S}_1^{-1}\mathbf{\Theta} \tag{9}$$

with the recursion

$$\mathbf{S}_k = \mathbf{C}_k - \mathbf{P}_k\mathbf{S}_{k-1}^{-1}\mathbf{N}_{k-1}, \tag{10}$$

with $\mathbf{S}_1 = \mathbf{C}_1$ and $k = 2, \ldots, L$. Interestingly, $\mathbf{S}_k$ is just the Schur complement [14] of the matrix

$$\begin{pmatrix} \mathbf{C}_k & \mathbf{P}_k \\ \mathbf{N}_{k-1} & \mathbf{S}_{k-1} \end{pmatrix}. \tag{11}$$

Using (9) together with the Schur complement (10) we are able to compute the asymptotic transfer matrix $\mathbf{W}$ of the resistive grid deploying only the coupling matrices.

The following two points should be noted. First, the matrix $\mathbf{W}$ will subsequently be used to compute the steady state response $\mathbf{y}$ to a constant input applied only a finite time T. This is only reasonable if we assume that the time constant of the dynamics of the resistive grid is much smaller than T. Thus under this assumption the input signal $\mathbf{u}(t)$ to the resistive grid can be any piecewise constant signal generated from a discrete-time sequence $\mathbf{u}[n]$ by a zero-th order hold operation

$$\mathbf{u}(t) = \mathbf{u}[n]\gamma(t), \quad \text{with} \quad n = \lfloor \frac{t}{T} \rfloor \in \mathbb{Z}, \tag{12}$$

where $\gamma(t)$ is a rectangular pulse of unit height and width T centered around time t. The second point to note is that, the goal of this research is to learn the coupling templates $A(k, l, r, s)$ and $B(k, r)$ during the processing of input data $\mathbf{u}(t)$. Thus, to put it precisely, the coupling templates are also a function of time t and (3) is time-variant linear system. By the assumption that the time constant for the adaptation dynamics is much larger than T the separate treatment of the processing dynamics and of the adaptation dynamics is justified.

III. Information Rate, Mutual Information

In the following we digress from the exact notation by not distinguishing between a random variable and its realization for the sake of conciseness. Furthermore all the following probabilistic quantities relate the input, the noise and output process at an arbitrary but same time point. As no quantities relate signals evaluated at different time points such as temporal correlations, the time arguments of the signals are not display subsequently.

The information rate or mutual information for continuous-valued random variables [7, p. 231] is defined as the difference between two entropies

$$I(\mathbf{z};\mathbf{u}) \equiv H(\mathbf{z}) - H(\mathbf{z}|\mathbf{u}), \tag{13}$$

where $H(\mathbf{z})$ denotes the entropy of the output signal $\mathbf{z}$ and $H(\mathbf{z}|\mathbf{u})$ denotes that part of the entropy of the output signal that is not caused by the input signal. Thus, in the current setup $H(\mathbf{z}|\mathbf{u})$ is the entropy of the additive perturbation $\boldsymbol{\epsilon}$. From (13) it becomes clear that for a $\boldsymbol{\epsilon}$ with given constant statistics the method of maximizing the mutual information $I(\mathbf{z};\mathbf{u})$ is equivalent to the maximum entropy method [15], i.e., maximizing $H(\mathbf{z})$. In terms of probability densities $I(\mathbf{z};\mathbf{u})$ can be written as

$$\begin{aligned} I(\mathbf{z};\mathbf{u}) &= \int_{-\infty}^{\infty} p(\mathbf{u},\mathbf{z}) \log \frac{p(\mathbf{u},\mathbf{z})}{p(\mathbf{u})p(\mathbf{z})} \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{z} \\ &= \int_{-\infty}^{\infty} p(\mathbf{z}|\mathbf{u})p(\mathbf{u}) \log \frac{p(\mathbf{z}|\mathbf{u})}{p(\mathbf{z})} \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{z}, \end{aligned} \tag{14}$$

where in the second equation the definition of the conditional distribution is applied. We assume that the input signal $\mathbf{u}$ as well as the additive output perturbation $\boldsymbol{\epsilon}$ are zero-mean Gaussian random processes with covariance matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$, respectively. Thus, the conditional density of $\mathbf{z}$ given the input $\mathbf{u}$ is

$$\begin{aligned} p(\mathbf{z}|\mathbf{u}) &= \frac{1}{(\sqrt{2\pi}\,)^M |\boldsymbol{\Sigma}|)^{\frac{1}{2}}} \\ &\times \exp\left[-\frac{1}{2}(\mathbf{z}-\mathbf{W}\mathbf{u})^T \boldsymbol{\Sigma}^{-1}(\mathbf{z}-\mathbf{W}\mathbf{u})\right], \end{aligned} \tag{15}$$

while the density of the input process $\mathbf{u}$ is just

$$p(\mathbf{u}) = \frac{1}{(\sqrt{2\pi}\,)^N |\boldsymbol{\Phi}|)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\mathbf{u}^T \boldsymbol{\Phi}^{-1}\mathbf{u}\right). \tag{16}$$

The distribution of the output signal $p(\mathbf{z})$ is obtained by marginalization

$$p(\mathbf{z}) = \int_{-\infty}^{\infty} p(\mathbf{z}|\mathbf{u})p(\mathbf{u})\mathrm{d}\mathbf{u}, \tag{17}$$

which yields

$$\begin{aligned} p(\mathbf{z}) = &\frac{1}{(\sqrt{2\pi}\,)^M |\mathbf{W}\boldsymbol{\Phi}\mathbf{W}^T + \boldsymbol{\Sigma}|^{\frac{1}{2}}} \\ &\times \exp\left[-\frac{1}{2}\mathbf{z}^T(\mathbf{W}\boldsymbol{\Phi}\mathbf{W}^T + \boldsymbol{\Sigma})^{-1}\mathbf{z}\right], \end{aligned} \tag{18}$$

Applying (15), (16) and (18) to the second equation in (14) gives

$$I(\mathbf{z};\mathbf{u}) = \frac{1}{2}\log\frac{|\mathbf{W}\boldsymbol{\Phi}\mathbf{W}^T + \boldsymbol{\Sigma}|}{|\boldsymbol{\Sigma}|}. \tag{19}$$

IV. Spatially Invariant Grid

For the analysis of the spatially invariant grid the following assumptions are made. Periodic boundary condition for the boundaries perpendicular to the input boundary (case I) and $N = M$ is assumed. The reason for this is that the setup allows for insightful simplifications through its rich algebraic structure. In the spatially invariant situation the entire resistive grid is characterized by its feedback template $\bar{A}(r, s)$ and its control template $\bar{B}(r)$, i.e., 12 real numbers. For convenience we associate with the feedback template $\bar{A}(r, s)$ a template matrix $\bar{\mathbf{A}}$ with the elements $\bar{A}_{kl} = \bar{A}(k-2, l-2)$. Correspondingly a template vector $\bar{B}_k = \bar{B}(k-2)$ is defined. For the matrices of (3) this results in a block-Töplitz matrix for

$$\mathbf{A} \equiv \begin{pmatrix} \mathbf{C} & \mathbf{N} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{P} & \mathbf{C} & \mathbf{N} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{P} & \mathbf{C} & \mathbf{N} & \cdots & \mathbf{0} \\ \vdots & & & & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{P} & \mathbf{C} \end{pmatrix}, \tag{20}$$

with the $N \times N$ circulant [16] submatrices

$$\begin{aligned} \mathbf{P} &= \operatorname{circ}(\bar{A}_{32}, \bar{A}_{33}, 0 \ldots, 0, \bar{A}_{31}) \\ \mathbf{C} &= \operatorname{circ}(\bar{A}_{22}, \bar{A}_{23}, 0 \ldots, 0, \bar{A}_{21}) \\ \mathbf{N} &= \operatorname{circ}(\bar{A}_{12}, \bar{A}_{13}, 0 \ldots, 0, \bar{A}_{11}). \end{aligned} \tag{21}$$

The submatrix $\boldsymbol{\Theta}$ of $\mathbf{B}$ is the circulant matrix

$$\boldsymbol{\Theta} = \operatorname{circ}(\bar{B}_2, \bar{B}_3, 0, \ldots, 0, \bar{B}_1) \tag{22}$$

and the submatrix $\boldsymbol{\Gamma}$ of $\boldsymbol{\Xi}$ is just the $N \times N$ identity matrix, which is the special circulant $\boldsymbol{\Gamma} = \operatorname{circ}(1, 0, \ldots, 0)$. The interesting property of circulant matrices is that all circulant share the same eigenvector system, i.e., all circulants get diagonalized by the same unitary matrix $\mathbf{U}$ with components

$$U_{kl} = \frac{1}{\sqrt{N}}\, e^{-\jmath\frac{2\pi(k-1)(l-1)}{N}}, \tag{23}$$

which is just the discrete Fourier transform matrix. A direct consequence of this fact is that products, sums and inverses of circulants are once again circulants. Thus, the set of circulant matrices form a group. Because of this, the result of the recursion in (10) and consequently the weight matrix in (9) are circulants. The eigenvalues of the matrices $\mathbf{P}$, $\mathbf{C}$ and $\mathbf{N}$ denoted as p_l, c_l and n_l with $l = 1, \ldots, N$, respectively are

$$\begin{aligned} p_l &= \bar{A}_{32} + \bar{A}_{33}\xi^{l} + \bar{A}_{31}\xi^{-l} \\ c_l &= \bar{A}_{22} + \bar{A}_{23}\xi^{l} + \bar{A}_{21}\xi^{-l} \\ n_l &= \bar{A}_{12} + \bar{A}_{13}\xi^{l} + \bar{A}_{11}\xi^{-l}, \end{aligned} \tag{24}$$

with $\xi \equiv e^{-\jmath\frac{2\pi}{N}}$. With (10) the eigenvalues s_l^k of the Schur complement $\mathbf{S}_k$ with $k = 1, \ldots, L$ are

$$\begin{aligned} s_l^1 &= \frac{1}{c_l} \\ s_l^2 &= c_l - \frac{p_l n_l}{c_l} \\ s_l^3 &= c_l - \cfrac{p_l n_l}{c_l - \cfrac{p_l n_l}{c_l}} \\ &\vdots \\ s_l^L &= c_l - \cfrac{p_l n_l}{c_l - \cfrac{p_l n_l}{c_l - \cfrac{p_l n_l}{c_l - \cdots}}}, \end{aligned} \tag{25}$$

written in terms of continued fractions. With (9) the eigenvalues w_l of the weight matrix $\mathbf{W}$ are

$$w_l = \frac{(p_l)^{L-1}\theta_l}{s_l^1 s_l^2 \cdots s_l^L}(-1)^L \tag{26}$$

where θ_l denotes the eigenvalues of the submatrix $\mathbf{\Theta}$. Applying the continued fractions of (25) to compute the product $s_l^1 s_l^2 \cdots s_l^L$ we obtain through inference from the intermediate steps

$$\begin{aligned} s_l^1 s_l^2 &= c_l^2 - p_l n_l \\ s_l^1 s_l^2 s_l^3 &= c_l^3 - 2c_l p_l n_l \\ s_l^1 s_l^2 s_l^3 s_l^4 &= c_l^4 - 3c_l^2 p_l n_l + (p_l n_l)^2 \\ s_l^1 s_l^2 s_l^3 s_l^4 s_l^5 &= c_l^5 - 4c_l^3 p_l n_l + 3c_l (p_l n_l)^2 \\ s_l^1 s_l^2 s_l^3 s_l^4 s_l^5 s_l^6 &= c_l^6 - 5c_l^4 p_l n_l + 6c_l^2 (p_l n_l)^2 - (p_l n_l)^3, \end{aligned} \tag{27}$$

that with $\beta(p_l, c_l, n_l) \equiv \prod_{k=1}^{L} s_l^k$

$$\beta(p_l, c_l, n_l) = \sum_{k=0}^{\lfloor \frac{L}{2} \rfloor} (-1)^k a(k, L) c_l^{L-2k} (p_l n_l)^k, \tag{28}$$

where $a(k, L)$ denotes the polynomial coefficients. The result allows for an expression of the eigenvalues of $\mathbf{W}$ of a resistive grid on a cylinder (periodic boundary conditions) without the recursive definition of the s_l^k as

$$w_l = \frac{(p_l)^{L-1}\theta_l}{\beta(p_l, c_l, n_l)}(-1)^L. \tag{29}$$

Based on (29) an expression for $I(\mathbf{z}; \mathbf{u})$ of (19) in the above style can be obtained if one additionally assumes that the input covariance matrix $\mathbf{\Phi}$ as well as the noise covariance matrix $\mathbf{\Sigma}$ are circulant matrices. An example of an input process with such properties would be the response of a spatially invariant resistive grid to a spatially white random vector with equal variance for each component. Thus, for a cylindrical setup and equal variance for each component the choice can be considered reasonable. Concerning $\mathbf{\Sigma}$, the most popular choice in many disciplines for the noise covariance is $\mathbf{\Sigma} = \sigma^2 \mathbf{I}$, which obviously is a circulant matrix.

Applying these assumptions to (19) and making use of the eigenspace representation of the involved matrices yields

$$\begin{aligned} I(\mathbf{z}; \mathbf{u}) &= \log|\mathbf{U}\mathbf{U}^{\dagger}| + \sum_{l=1}^{N} \log(1 + \frac{\phi_l}{\sigma^2} w_l^* w_l) \\ &= \sum_{l=1}^{N} \log(1 + \frac{\phi_l}{\sigma^2} |\frac{(p_l)^{L-1}\theta_l}{\beta(p_l, c_l, n_l)}|^2), \end{aligned} \tag{30}$$

where ϕ_l are the eigenvalues of $\mathbf{\Phi}$ and "$(\cdot)^{\dagger}$" and "$(\cdot)^{*}$" denote hermitian and complex conjugation, respectively. Viewing the expression (30) as $\propto \log(1+\text{SNR})$ with SNR the signal-to-noise-ratio, indicates that $I(\mathbf{z}; \mathbf{u})$ can be arbitrarily increased by scaling the signal component through a larger gain $w_l^* w_l$ of the resistive grid. Thus, without an additional constraint the InfoMax solution is not finite. A natural choice is to limit some norm of the coupling templates. Another one would be to limit some norm of the eigenvalues w_l directly. For simplicity the control template is set to $\bar{B}_{kl} = \delta_{kl}$ with the Kronecker delta δ_{kl} for the following analysis. The first of the above choices would yield a Lagrange function

$$\mathcal{L}(\bar{\mathbf{A}}, \lambda) = I(\mathbf{z}; \mathbf{u}) + \lambda(\text{Tr}(\bar{\mathbf{A}}^T \bar{\mathbf{A}}) - K), \tag{31}$$

with K is predetermined positive constant and λ the Lagrange multiplier. In addition to this constraint, a stability constraint has to be applied because the expression (9) is only valid for the system (3) to be stable.

The following approach aims to combine the two above constraints into one. The idea is to set the self-feedback template coefficient $\bar{A}_{22}$ to a negative constant which guarantees a stability margin under the norm constraint for the remaining template coefficients. This can be formalized using the Gerschgorin circle theorem [14]. It states that the eigenvalue of the matrix $\mathbf{A}$ in (4) lie within the union of circles described by η with

$$|\eta - A_{ll}| < \rho_l, \quad \text{with} \quad \rho_l = \sum_{k=1}^{LN} |A_{lk}| - |A_{ll}| \tag{32}$$

For the special case of (20) the set of conditions can be reduced to one condition in terms of the template coefficients

$$|\eta - \bar{A}_{22}| < \rho, \quad \text{with} \quad \rho = \sum_{l=1}^{3}\sum_{k=1}^{3} |\bar{A}_{lk}| - |\bar{A}_{22}|, \tag{33}$$

which is valid for rows of (20) describing the dynamics of layer $1 < k < L$. For the layers $k = 1$ and $k = L$ a tighter bound for ρ can be obtained, but subsequently the sufficient condition (33) is used. Interpreted in terms of stability we require that for a given negative value of $\bar{A}_{22}$ the interval of radius ρ should not contain positive numbers, i.e., we require that $\bar{A}_{22} + \rho < 0$. Thus,

$$\sum_{l=1}^{3}\sum_{k=1}^{3} |\bar{A}_{lk}| < 2|\bar{A}_{22}|, \tag{34}$$

which means that $\mathbf{A}$ is diagonally dominant. To be compatible with the constraint term in the Lagrange function (31) we use Hölders inequality

$$\sum_{k=1}^{L} |x_k y_k| \le \left(\sum_{k=1}^{L} |x_k|^p\right)^{\frac{1}{p}} \left(\sum_{k=1}^{L} |x_k|^q\right)^{\frac{1}{q}}, \tag{35}$$

with $\frac{1}{p} + \frac{1}{q} = 1$ by choosing $y_k = 1$ for $k = 1, \dots, L$ and $p = q = 2$, such that

$$\sum_{k=1}^{L} |x_k| \le \big(L \sum_{k=1}^{L} x_k^2\big)^{\frac{1}{2}} \tag{36}$$

Thus the condition

$$4 \sum_{l=1}^{3}\sum_{k=1}^{3} \bar{A}_{lk}^2 < \bar{A}_{22}^2 \tag{37}$$

is sufficient for (34) to hold. Therefore the constant in (31) can be chosen to be $K = \frac{\bar{A}_{22}^2}{4} - \varepsilon$, where $\varepsilon > 0$ is a small additional margin. The optimization problem for a given value of $\bar{A}_{22}$ then reads

$$\begin{aligned} &\max_{\bar{\mathbf{A}} \backslash \bar{A}_{22}} \sum_{l=1}^{N} \log(1 + \frac{\phi_l}{\sigma^2} \big| \frac{(p_l)^{L-1}}{\beta(p_l, c_l, n_l)} \big|^2) \\ &\text{subject to} \\ &\mathrm{Tr}(\bar{\mathbf{A}}^T \bar{\mathbf{A}}) = \big(\frac{\bar{A}_{22}}{2}\big)^2 - \varepsilon, \end{aligned} \tag{38}$$

where p_l, c_l and n_l are related to $\bar{\mathbf{A}}$ by (24). The notation $\bar{\mathbf{A}} \backslash \bar{A}_{22}$ should indicate that $\bar{A}_{22}$ is not in the set of optimization variables. A thorough simulation study of this optimization problem will is subject to an upcoming paper.

V. Spatially Variant Grid

For the spatially variant grids we are going to distinguish two situations. In the first, for each layer k the coupling template is spatially invariant, but is different for different layers. The second case considers the most general case where each coupling strength can be adjusted individually.

A. Spatially Invariant Layers

Similar to section IV we associate with the feedback template of (1) and (2) a spatially invariant 3×3 feedback matrix $\bar{\mathbf{A}}^k$ for the layer k. In contrast to (20) the matrix $\mathbf{A}$ now takes on the general form of (4) and is not Töplitz anymore. On the other hand the submatrices resemble the spatially invariant coupling inside a layer and are therefore either Töplitz (case II) or a circulant (case I). Once again we will assume case II boundary conditions because of its richer algebraic structure. The submatrices of (4) is thus defined as

$$\begin{aligned} \mathbf{P}_k &= \mathrm{circ}(\bar{A}_{32}^k, \bar{A}_{33}^k, 0 \dots, 0, \bar{A}_{31}^k). \\ \mathbf{C}_k &= \mathrm{circ}(\bar{A}_{22}^k, \bar{A}_{23}^k, 0 \dots, 0, \bar{A}_{21}^k) \\ \mathbf{N}_k &= \mathrm{circ}(\bar{A}_{12}^k, \bar{A}_{13}^k, 0 \dots, 0, \bar{A}_{11}^k) \end{aligned} \tag{39}$$

Correspondingly, it is assumed that the control template (1) is spatially invariant such that (22) remains valid. Denoting p_l^k, c_l^k and n_l^k as the l-th eigenvalues of the matrices $\mathbf{P}_k$, $\mathbf{C}_k$ and $\mathbf{N}_k$, respectively, the eigenvalues of the asymptotic transfer matrix read

$$w_l = \frac{\theta_l \prod_{k=2}^{L} p_l^k}{s_l^1 s_l^2 \cdots s_l^L} (-1)^L, \tag{40}$$

with the recursion $s_l^k = c_l^k - \frac{p_l^k n_l^{k-1}}{c_l^{k-1}}$. In terms of the maximization of the information rate the results of section IV can be generalized to

$$\begin{aligned} &\max_{\{\bar{\mathbf{A}}^1 \backslash \bar{A}_{22}^1, \dots, \bar{\mathbf{A}}^L \backslash \bar{A}_{22}^L\}} \sum_{l=1}^{N} \log(1 + \frac{\phi_l}{\sigma^2} \big| \frac{\prod_{k=2}^{L} p_l^k}{s_l^1 s_l^2 \cdots s_l^L} \big|^2) \\ &\text{subject to} \\ &\mathrm{Tr}((\bar{\mathbf{A}}^1)^T \bar{\mathbf{A}}^1) = \big(\frac{\bar{A}_{22}^1}{2}\big)^2 - \varepsilon \\ &\vdots \\ &\mathrm{Tr}((\bar{\mathbf{A}}^L)^T \bar{\mathbf{A}}^L) = \big(\frac{\bar{A}_{22}^L}{2}\big)^2 - \varepsilon, \end{aligned} \tag{41}$$

where the feedback template was set to $\bar{B}_{kl} = \delta_{kl}$.

B. The General Case

An important restriction of the above formalism for the spatially invariant grid and spatially invariant layers is, that no dimension reduction as illustrated in Fig. 3 can be incorporated. Even if a spatially invariant grid of rectangular shape with periodic boundary condition is used in combination with a output matrix $\mathbf{\Xi}$ of lower dimension $M < N$, the elegant eigenvalue formalism breaks down. This becomes evident from (9), because the $\mathbf{\Gamma}$ is no more a identity matrix and $\mathbf{W}$ is rectangular matrix.

To study this interesting problem of dimension reduction, i.e., of lossy compression, one has to go back to the equations (3), (4), (9) and (19). Obviously, no constraints for the lateral boundary conditions and on the geometry of the grid is imposed. Thus, the transfer matrix of the grid in Fig. 3 can still be cast into the form (9). In terms of maximizing the information rate the general problem now reads

$$\begin{aligned} &\max_{\mathbf{A}, \mathbf{\Theta}} \log |\mathbf{W}\mathbf{\Phi}\mathbf{W}^T\mathbf{\Sigma}^{-1} + \mathbf{I}| \\ &\text{subject to} \\ &\Re(\mathbf{A}) < 0 \\ &\mathrm{Tr}(\mathbf{W}^T\mathbf{W}) - \bar{K} = 0, \end{aligned} \tag{42}$$

in combination with (9) and (10), with $\bar{K}$ some predetermined constant. The variables $\mathbf{A}$ and $\mathbf{\Theta}$ are the representatives of the individual coupling parameters of the grid.

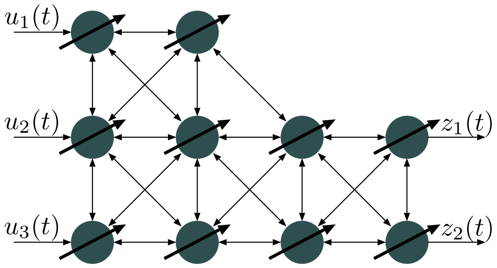

Fig. 3. Dimension reduction $M < N$ of the adaptive resistive grid, with different number of cells in the layers.

VI. Relation to Principal Component Analysis

The section is devoted to the relation between the InfoMax principle and the PCA as well as PSA. Using the identity $\log|\mathbf{X}| = \operatorname{Tr}\log(\mathbf{X})$ for a general positive definite matrix $\mathbf{X}$ and applying it to (19) yields

$$\begin{aligned} I(\mathbf{z};\mathbf{u}) &= |\mathbf{I} + \mathbf{W\Phi W}^T\mathbf{\Sigma}^{-1}| \\ &= \operatorname{Tr}\log(\mathbf{I} + \mathbf{W\Phi W}^T\mathbf{\Sigma}^{-1}) \\ &= \operatorname{Tr}\left[\mathbf{W\Phi W}^T\mathbf{\Sigma}^{-1} - \frac{1}{2}(\mathbf{W\Phi W}^T\mathbf{\Sigma}^{-1})^2 + \cdots\right] \\ &\approx J(\mathbf{W}) \equiv \operatorname{Tr}\left(\mathbf{W\Phi W}^T\mathbf{\Sigma}^{-1}\right), \end{aligned} \tag{43}$$

where a truncation of the power expansion of the logarithm at the first order is used. For $\mathbf{\Sigma}$ a diagonal matrix the last expression of (43) is just the weighted sum of the variances of the output signals y_k with $k = 1, \ldots, M$. This is the general cost-function for PCA [17], [4] and reduces to the cost-function of Oja's PSA algorithm for $\mathbf{\Sigma} = \mathbf{I}$. While for PCA $\mathbf{\Sigma}$ is a diagonal matrix with monotonically decreasing entries which was introduced to eliminate the ambiguity of the PSA in favor of PCA, in the InfoMax framework $\mathbf{\Sigma}$ has a physical, tangible interpretation. In PCA and PSA algorithms the matrix $\mathbf{W}$ should converge to a matrix with orthogonal rows which span the principle subspaces. This constraint with the necessary magnitude constraint for the network gain can be combined in

$$\mathbf{WW}^T = \mathbf{I}. \tag{44}$$

The gradient of the cost with respect to $\mathbf{W}$ [17] is

$$\frac{\partial J(\mathbf{W})}{\partial \mathbf{W}} = -\mathbf{W\Phi W}^T\mathbf{\Sigma}^{-1}\mathbf{W} + \mathbf{\Sigma}^{-1}\mathbf{W\Phi}, \tag{45}$$

where the constraint (44) was taken into account. One associates a gradient flow $\dot{\mathbf{W}} = -\mu\frac{\partial J(\mathbf{W})}{\partial \mathbf{W}}$, which performs gradient ascent on the cost function $J(\mathbf{W})$. Applying the stochastic approximation $\mathbf{\Phi} \approx \mathbf{uu}^T$ allows to decompose the expression of the matrix flow $\dot{\mathbf{W}}$ into flows of the $M \times 1$ column vectors of $\mathbf{W} = [\,\mathbf{w}_1, \ldots, \mathbf{w}_N\,]$

$$\dot{\mathbf{w}}_k = \mu\mathbf{y}(\hat{u}_k - u_k) = \mu e_k\mathbf{y}, \tag{46}$$

where additionally $\mathbf{\Sigma} = \mathbf{I}$ was assumed. The quantity $\hat{\mathbf{u}} \equiv [\,\hat{u}_1, \ldots, \hat{u}_N\,]^T$ is the reconstruction of the input in terms of the, in general lower dimensional, output $\hat{\mathbf{u}} = \mathbf{W}^T\mathbf{y}$, $e_k \equiv \hat{u}_k - u_k$ is the reconstruction error and $\mu > 0$ is the learning rate. According to the discussion in section II the learning rate has to be chosen small enough such that the adaptation and the processing dynamics can be considered to operate at largely different time scales.

Unfortunately, for the resistive grid one can not update the matrix $\mathbf{W}$ using (46), because it is a function of the coupling templates $\bar{A}(l, k, r, s)$ and $\bar{B}(l, r)$. This is in contrast to feedforward networks where the matrix of coupling strength is identical to its transfer matrix.

The interesting question posed here is whether there exists a recursive system, especially a resistive grid where a matrix of coupling strengths is identical to the asymptotic transfer matrix $\mathbf{W}$ of the system. The following approach is based on [12], where a very interesting multilayer recursive neural network is proposed. In this network the processing cells of all layers are recursively coupled to all cells of their neighboring layers except for the first layer. In terms of the processing topology the cells of the first layer, i.e., the input layer, do not receive the feedback signal from the second layer.

The difference to a planar resistive grid for layers $k > 1$ is that there is global coupling between layers, i.e, a cell at layer k is coupled to all cells in layers $k-1$ and $k+1$. In [12] an in-depth treatment of the PCA using such a network with two layers is performed. But for a network with two layers the proposed processing topology is just of the classically feedforward type, because the second layer, as it is the last layer, does not receive a feedback signal.

Subsequently it is shown that the topology can be changed such that the first layer receives a feedback signal from the second layer. The resulting network is depicted in Fig. 4 for the case of $N = 3$ and $M = 2$. The special features of the network is that the coupling strengths are symmetric and that there is no lateral coupling between the cells. In terms of the coupling matrices in (4) we obtain

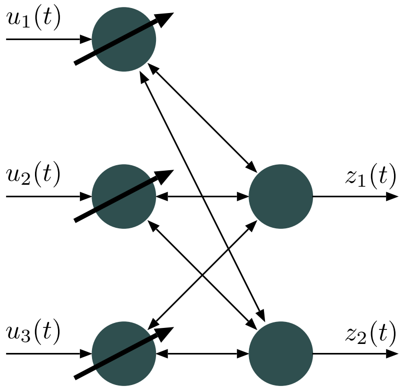

Fig. 4. A Two layer resistive grid which global coupling between the layers and no lateral, layer-internal coupling; output layer is not adaptive.

$$\begin{aligned} \mathbf{x}_1 &= \mathbf{C}_1\mathbf{x}_1 + \mathbf{N}_1\mathbf{x}_2 + \mathbf{\Theta u} \\ \mathbf{x}_2 &= \mathbf{C}_2\mathbf{x}_2 + \mathbf{P}_2\mathbf{x}_1 \end{aligned} \tag{47}$$

with $\mathbf{C}_1$, $\mathbf{C}_2$ and $\mathbf{\Theta}$ diagonal matrices and $\mathbf{P}_2 = \mathbf{N}_1^T$. For

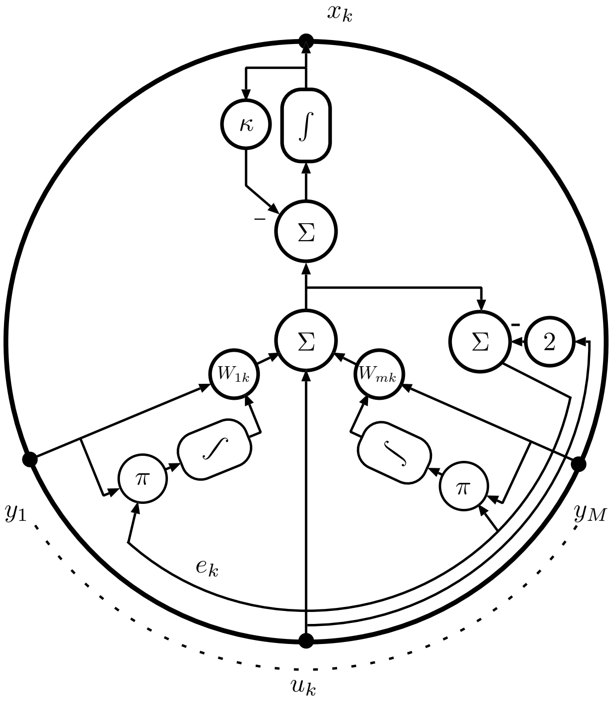

Fig. 5. Adaptive cell k for the resistive grid with global coupling between layers of Fig. 4; Although, there are only local learners for each cell, the network self-organizes to maximize the global PSA cost function $J(\mathbf{W})$ of (43) with $\mathbf{\Sigma} = \mathbf{I}$.

the recursively coupled higher layers $k > 1$ in [12] the self-feedback matrices were assumed to be $\mathbf{C}_1 = -\mathbf{I}_N$ and $\mathbf{C}_2 = -\mathbf{I}_M$. In the following we choose $\mathbf{C}_1 = -\kappa\mathbf{I}_N$, $\mathbf{C}_2 = -\kappa\mathbf{I}_M$ and $\mathbf{\Theta} = (\kappa^2 - 1)\mathbf{I}_N$ with $\kappa > 1$. Applying the expressions (9) and (10) to this setting gives

$$\mathbf{W} = (\kappa\mathbf{I} - \frac{\mathbf{P}_2\mathbf{P}_2^T}{\kappa})^{-1}\frac{1}{\kappa}\mathbf{P}_2\mathbf{\Theta}. \tag{48}$$

As there is global coupling between the layers, $\mathbf{P}_2$ is a general matrix without any band-structure. Thus, the matrix can be forced to obey $\mathbf{P}_2\mathbf{P}_2^T = \mathbf{I}$. With this the asymptotic transfer matrix becomes

$$\mathbf{W} = \mathbf{P}_2. \tag{49}$$

From (48) it is clear that for $\kappa = 1$ the matrix to be inverted is singular and the corresponding dynamical system (3) gets unstable.

Thus, for this specific network the matrix of coupling strength is identical to the asymptotic transfer matrix. The big advantage of this structure, compared to general recursive topologies is that no inverse matrix as in (9) is involved.

With this slight modification the network has been made fully recursive and stable. From (46) one can conclude that, subject to the global PCA cost function, each cell can adapt its coupling strength based on the signal locally available to the cell. Such an adaptive cell is shown in Fig. 5. The cell architecture can be generalized to perform PCA instead of PSA by choosing $\mathbf{\Sigma}$ to be the diagonal matrix with monotonically decreasing entries. The learning algorithm remains local with the difference that each weight in cell k would now receive a different reconstruction error signal for its adaptation.

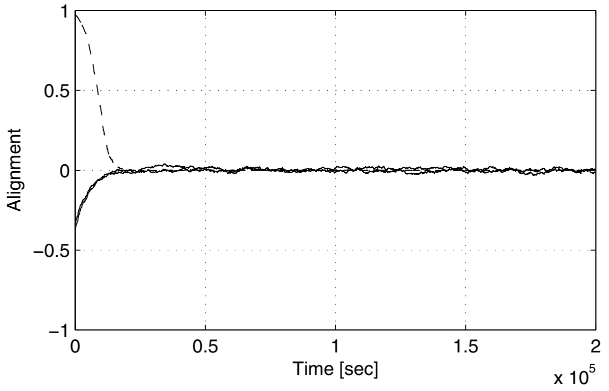

Fig. 6. Learning curve for the principal subspace network of Fig. 4; extraction of two principal subspaces from a 3-dimensional random input; mutual alignment of the two strongest eigenvector (dashed), alignment of the extracted subspace to the offline computed third eigenvector (solid).

VII. Simulation Results

The behavior of the special resistive grid for PSA introduced in section VI is illustrated with the following toy example. Consider the network in Fig. 4 with a 3-dimensional random process as input and the first two principal components as its output. The input process is generated by correlating a discrete-time white Gaussian random process with variances $\sigma_1^2 = 1$, $\sigma_2^2 = 0.6$ and $\sigma_3^2 = 0.1$ by a randomly generated orthogonal matrix to yield a nondiagonal input covariance matrix $\mathbf{\Phi}$. This process is transformed to a continuous-time piecewise constant signal through a zero-th order hold operation and is then applied to the continuous-time Simulink model of the resistive grid. The time constant in (12) for the hold operation is $T = 20$ sec. The stability margin parameter κ is chosen to be $\kappa = 2$, while the learning rate is conservatively set to $\mu = 3 \times 10^{-4}$. In Fig. 6 the alignment, i.e., the cosine of the angle between two vectors, is shown over time. The mutual alignment of the two vectors spanning the principal subspace as well as their alignment to the offline computed third eigenvector of the input covariance matrix $\mathbf{\Phi}$ is depicted. A short section of the time evolution of one input component $u_1(t)$ and its internal reconstruction $\hat{u}_1(t)$ is given in Fig. 7.

VIII. Conclusions

First results of the research on adaptive Cellular Neural Networks are presented. Focus is put on the analysis of a simple linear resistive grid. A compact expression for the information rate over the resistive grid for the spatially invariant case is derived in terms of the eigenvalues of the asymptotic transfer matrix and the spectrum of the input correlation sequence. Furthermore, the optimization problem with its magnitude constraint as well as its stability constraint is formalized. For resistive grids where the coupling templates vary for different layers but where the template

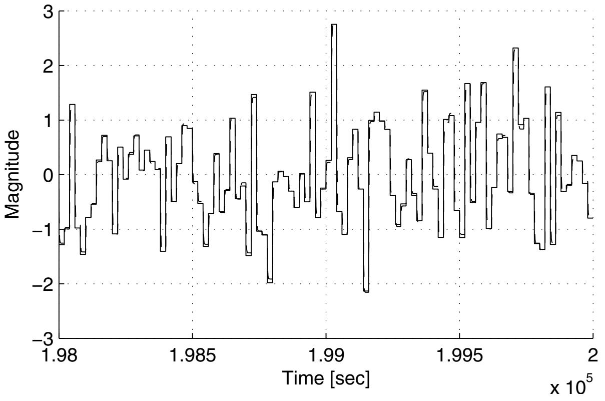

Fig. 7. Short section of the time evolution of one component $u_1(t)$ (solid) of the vectorial input process and its network reconstruction $\hat{u}_1(t)$ (dashed) after adaptation; one clearly observes the piecewise constant nature of the input sequence using a zero-th order hold and the sufficiently fast processing dynamics of the network.

inside one layer is spatially invariant similarly compact expression as for the spatially invariant case are derived. For a special two-layered fully connected resistive grid we are able to compute the gradient flow of the weight matrix without involving a inverse of a matrix. This allows for the design of local learners inside each cell which requires the locally available signals inside a cell only. One approach to overcome the exploding wiring complexity for large input dimension of this fully connected grid is to performs local PSA or PCA by splitting the input vector into multiple vectors of smaller dimension. In this case the networks has to be fully connected for the smaller dimension. Many of the obtained results can be generalized to a multilayer resistive grid with a 2-dimensional (visual) input.

References

[1] L. O. Chua, *CNN: A Paradigm for Complexity*, ser. World Scientific Series on Nonlinear Science. World Scientific, 1998.

[2] Y. Chang, T. Ho, and L. Kaelbling, "All learning is local: Multi-agent learning in global reward games," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Morgan Kaufmann, 2004.

[3] D. Wolpert and K. Tumer, "An introduction to collective intelligence," NASA Ames, Research Center, Tech. Rep. NASA-ARC-IC-99-63, 1999.

[4] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*. John Wiley, 2002.

[5] R. D'Andrea and G. E. Dullerud, "Distributed control design for spatially interconnected systems," *IEEE Transaction on Automatic Control*, vol. 48, no. 9, pp. 1478–1495, 2003.

[6] T. Hogg and B. Huberman, "Controlling smart matter," *Smart Materials and Structures*, vol. 7, pp. R1–R14, 1998.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley, 1991.

[8] R. Linsker, "An application of the principle of maximum information preservation to linear systems," in *Advances in Neural Information Processing Systems 1*, D. S. Touretzky, Ed. Morgan Kaufmann, 1989, pp. 186–194.

[9] ——, "Self-organization in a perceptual network," *IEEE Computer*, vol. 21, no. 3, pp. 105–117, 1988.

[10] ——, "From basic network principles to neural architecture: Emergence of spatial-opponent cells," in *Proc. Natl. Acad. Sci.*, vol. 83, 1986, pp. 7508–7512.

[11] A. J. Bell and T. J. Sejnowski, "An information maximisation approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[12] L. Xu, "Least mean square error reconstruction principle for self-organizing neural networks," *Neural Networks*, vol. 6, pp. 627–648, 1993.

[13] B. E. Shi and L. O. Chua, "Resistive grid image filtering: Input/output analysis via the CNN framework," *IEEE Transactions on Circuits and Systems–I: Fundamental Theory and Applications*, vol. 39, no. 7, pp. 531–548, 1992.

[14] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*. Philadelphia, USA: Society for Industrial and Applied Mathematics, 2000.

[15] D. S. Sivia, *Data Analysis, A Bayesian Tutorial*. Oxford University Press, 1996.

[16] P. J. Davis, *Circulant Matrices*. John Wiley, 1979.

[17] R. Brockett, "Dynamical systems that learn subspaces," in *Mathematical System Theory: The Influence of R.E. Kalman*, A. C. Antoulas, Ed. Berlin: Springer-Verlag, 1991, pp. 579–592.