Discrimination Networks for Maximum Selection

Brijnesh J. Jain and Fritz Wysotzki
Department of Computer Science
Technical University Berlin
Germany
email:{bjj—wysotzki}@cs.tu-berlin.de

Contact:

Brijnesh J. Jain
Department of Computer Science, Sekr. FR 5-8
Technical University Berlin
Franklinstr. 28/29
D-10587 Berlin
Germany

phone : +49-30-314-23938
fax : +49-30-314-24913
e-mail: bjj@cs.tu-berlin.de

Preferred section: Mathematical and Computational Analysis

Running title: Discrimination Networks

Discrimination Networks for Maximum Selection

Abstract: We construct a novel discrimination network using differentiating units for maximum selection. In contrast to traditional competitive architectures like MAXNET the discrimination network does not only signal the winning unit, but also provides information about its evidence. In particular, we show that a discrimination network converges to a stable state within finite time and derive three characteristics: (P_1) intensity normalization, (P_2) contrast enhancement, and (P_3) evidential response. In order to improve the accuracy of the evidential response we incorporate distributed redundancy into the network. This leads to a system which is not only robust against failure of single units and noisy data, but also enables us to sharpen the focus on the problem given in terms of a more accurate evidential response. The proposed discrimination network can be regarded as a connectionist model for competitive learning by evidence.

Keywords: competitive learning, contrast enhancement, distributed redundancy, dynamical systems, evidential response, maximum selection, winner-takes-all networks.

1 Introduction

One way to deal with maximum selection from a set of inputs within a connectionist framework are *winner-take-all* (WTA) networks, Feldman & Ballard (1982). The operation of these networks is a mode of contrast enhancement and pattern normalization where only the unit with the highest activation fires and all other units in the network are inhibited after some settling time. An example of a common competitive architecture to select the maximum or minimum from a set of data is MAXNET Lippman (1987). Other techniques to pick a maximum can be found in Feldman & Ballard (1982), Hopfield & Tank (1986), Kohonen (1984), and Lippman (1988). In Majani, Erlanson, & Abu-Mostafa (1989) maximum selection is generalized to k-winners-take-all networks which identify the largest k of n real numbers. More general results on the dynamics of competitive models are given in Hirsch (1989) and Lemon & Vijaya Kumar (1989). WTA networks are analyzed in Ermentrout (1992), Fang, Cohen, & Kincaid (1996), Hahnloser (1998). In Xie, Hahnloser, & Seung (2000) the dynamics of the winner-take-all competition between given groups of neurons is studied. Wersing, Beyn, & Ritter (2001) examines the behavior of a more general recurrent network with piecewise linear transfer functions. Results on the computational power of WTA nets can be found in Maass (2000a), Maass (2000b).

WTA networks for maximum selection suffer from two severe drawbacks: lack of *evidential response* and lack of *fault tolerance*. Both features, evidential response and fault tolerance, are regarded to be important properties of neural networks (see Haykin (1999) for a more detailed discussion).

Evidential response is a capability of neural networks which arises in the context of pattern classification and recognition to provide information about the confidence in the decision made. Following the extreme selection principle *winner takes all* inhibitory WTA networks are designed to signal the particular unit to select, but are not capable to indicate the level of confidence of the selected pattern. This information may be used to arrange for ambiguous decisions and thereby improve the classification and pattern recognition performance of the network.

In addition WTA networks are not fault tolerant, since they are not robust to failure of single units or noisy data. If a unit or its connecting links are damaged, performance is impaired or faulty and the network behaves unreliable. In classification tasks or prototype detection, failure of one WTA unit means loss of the whole category. Lack of fault tolerance can be simply removed by introducing *distributed*

redundancy. Nevertheless, due to the lack of evidential response the relationship between the quality of an evidential response and distributed redundancy remains unclear.

The main goal of our contribution is to propose a competitive *discrimination network* for maximum selection with evidential response. In an in-depth analysis we prove that the discrimination network converges within finite time to an equilibrium point and satisfies the following properties: (P_1) intensity normalization, (P_2) contrast enhancement, and (P_3) evidential response. In addition we show that incorporating redundancy into the network improves evidential responses. This result provides new insights into the benefits of distributed redundancy.

Properties (P_1) and (P_2) describe the evolution of the proposed network until convergence to an equilibrium point. Property (P_2) is well known for shunting networks and WTA networks (see e.g. (Grossberg, 1988; Levine, 2000). We restate (P_2) for discrimination networks using a straightforward proof. Property (P_3) is the key result of this paper. It enables us to extract information about the evidence for the selected unit. The evidential response is an approximation of a canonical discrimination measure δ, namely the difference of the maximum value of a set of inputs and the average of this set. In connection with property (P_3) a new feature of distributed redundancy emerges. It does not only lead to a robust and fault tolerance system, but also enables us to control the focus on the problem given. The sharpness of the focus can be assessed by the accuracy with which the network approximates the canonical discrimination measure δ.

To prove properties (P_1)-(P_3) and convergence within finite time to an equilibrium point we pursue the following approach: We consider a competitive linear neural architecture associated to the discrimination network. The dynamics of the simplified linear system is described by a system of coupled differential equations for which we will present its solution in closed form. Due to the linearity the associated linear system diverges. But the solution facilitates direct access to examine characteristic features of the linear network's behavior, in particular (P_1)-(P_3). Finally, we show that the nonlinear discrimination network preserves properties (P_1)-(P_3) and prove convergence to an equilibrium point within finite time.

Discrimination networks give rise to a modified version of simple competitive learning. As inhibitory WTA networks are a neural network implementation of competitive learning, the proposed discrimination network is a connectionist framework of *evidential learning*. In evidential learning the extreme selection principle of *winner takes all* is replaced by an activity-related soft selection principle. We introduce

evidential learning in Section 5.

The rest of this contribution is organized as follows: Section 2 introduces the architecture of a discrimination network. In Section 3 we analyse the linear model associated to a discrimination network and derive the properties (P_1)-(P_3). Section 4 transfers properties (P_1)-(P_3) to discrimination networks and proves its convergence to an equilibrium state. Section 5 shows discrimination network as a connectionist framework for competitive evidential learning. Finally, Section 6 concludes this contribution.

Notation: Let $\mathbb{R}$ be the field of real numbers and $\mathbb{R}_+$ be the set of non-negative real numbers. We denote vectors by bold letters (e.g. $\boldsymbol{x}$). A line spanned by a vector $\boldsymbol{x}$ is occasionally written in the form $\mathbb{R}\boldsymbol{x}$. For any vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ we set $\bar{x} := \frac{1}{n}\sum_i x_i$ and $\bar{\boldsymbol{x}} := (\bar{x}, \ldots, \bar{x})$. Furthermore, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ the scalar product is denoted by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ and the Euclidean distance by $||\boldsymbol{x} - \boldsymbol{y}||$.

2 Maximum selection with a discrimination network

In this Section we describe a discrimination network model with the objective to discriminate input patterns.

We call units which are able to recognize local minima of their activations *differentiating units*. A *discrimination network* consists of fully connected differentiating units with negative connective weights. The units switch off their outputs as soon as their activations arrive at a minimum. For time-discrete systems a differentiating unit can be modeled by a feedback loop using a unit-delay operator, whose output is delayed with respect to the input by one time unit (see Haykin, 1999, Section 1.5 for a discussion about unit-delay operators).

The dynamics of a continuous discrimination network is defined by the following pair of equations:

$$\dot{x}_i(t) = -w \cdot \sum_{\substack{j=1 \\ j \neq i}}^{n} y_j(t) \tag{1}$$

$$y_i(t) = x_i(t) \cdot f\Big(\dot{x}_i(t)\Big) \tag{2}$$

where $x_i(t)$ is the activation or the *short-term-memory* STM trace [1] of unit i, $w_{ij} = -w < 0$ represents the strength of the synapse connecting unit i and unit j, and $y_i(t)$ is the output signal. To recognize the turning point of $x_i(t)$ the output signal applies the function f, which is of the form

$$f\left(\dot{x}_i(t)\right) = \begin{cases} 0 & : \quad \dot{x}_i(t) \geq 0 \\ 1 & : \quad \text{otherwise} \end{cases} \tag{3}$$

Let the input patterns to discriminate be given by a tuple of n numbers $\boldsymbol{a} = (a_1, \ldots, a_n)$. To select a maximum value a_{i_0} $(1 \leq i_0 \leq n)$ construct a discrimination network with n units. Thus each unit represents one pattern (one number of the tuple). Set the initial activation $x_i(0)$ of unit i proportional to value a_i. During evolution of the network the units compete among each other until convergence to an equilibrium point. The network converges when the activation of a unit arrives at a minimum for the first time. The minimum is unique and therefore corresponds to the global minimum (Prop. 3.5). The winning unit is the one first arriving at a local minimum of its activation (Lemma 3.4). It is exactly the unit with highest initial activation, i.e. the unit representing the maximum value of the input patterns. Thus, to identify the winning unit and to signal the selected maximum value an appropriate readout device has to be wrapped around the discrimination system. Transferring the activation of the winning unit to the readout device does not only signal the winner but also provides us information about the evidence for the selected maximum value. The evidence for the selection is an approximation of a canonical discrimination measure $\delta^{i_0}_{\boldsymbol{a}} := a_{i_0} - (a_1 + \cdots + a_n)/n$ which is the difference of the maximum value a_{i_0} $(1 \leq i_0 \leq n)$ and the average value of the numbers $a_1, \ldots, a_n$. To improve the evidential response by means of a more accurate approximate value of $\delta^{i_0}_{\boldsymbol{a}}$, we introduce distributed redundancy into the discrimination network. Thus, incorporating distributed redundancy leads to a system which is robust against failure of single units and also sharpens the focus on the problem.

3 The associated linear model

Intensity normalization (P_1) is a property of a network where the total amount of activation converges to a constant value which is independent from the initial ac-

[1] The term *short-term memory* is used in the sense of Grossberg (e.g. Grossberg, 1988).

tivation. A network enhances contrasts (P_2) if it enlarges the difference between activations while preserving their proportions. Evidential response (P_3) in the context of maximum selection provides us information about the level of confidence in the decision made. Here the evidence that a particular unit represents the maximum of a set of inputs is given with respect to the difference of the initial activation of that unit and the average of all initial activations.

In order to show (P_1)-(P_3) we first consider a time-continuous linear network associated to the discrimination network. If (P_1)-(P_3) hold for an associated linear model, then we can prove that (P_1)-(P_3) also hold for a discrimination network (see Section 4), since the nonlinear network and its associated linear system are related as follows: In the associated linear system the activation of all units are decreasing first. Then after a certain period the activation of units with initial activation above average start increasing. In addition the linear system preserves the order of the units with respect of their activations. In the nonlinear system only units with decreasing activation fire while units with increasing activation remain quiescent. Thus the nonlinear system behaves like its associated linear system until the activation of the first unit –the one with highest initial activation– starts increasing.

Section 3.1 introduces the linear model associated to the nonlinear discrimination network. In Section 3.2 we give an exact solution of the underlying differential equations describing the associated linear system. Knowledge of the solution establishes a sound basis for uncovering various characteristic features of the system. In particular, features (P_1)-(P_3) are derived in Section 3.3.

3.1 The associated linear model

The linear model associated with the discrimination network is a competitive neural network with linear transfer function of the form

$$\dot{x}_i(t) = -w \cdot \sum_{\substack{j=1 \\ j \neq i}}^{n} h\Big(x_j(t)\Big), \qquad x_{0i} := x_i(0) \tag{4}$$

where $x_i(t)$ is the activation or the *short-term-memory* STM trace of unit i, and $-w < 0$ is the synaptic weight of the connection of unit i and unit j. The function $h(x) = kx$ is a linear transfer function with gain $k > 0$. For our analysis it is sufficient to assume $k = 1$.

Throughout Section 3 we consider the more general linear neural network

$$\dot{x}_i(t) = -dx_i(t) - w \sum_{\substack{j=1 \\ j \neq i}}^{n} h\Big(x_j(t)\Big) + I_i, \qquad x_{0i} := x_i(0) \tag{5}$$

where $d \geq 0$ is the self-decay rate and I_i an external input. Clearly, (4) is the associated homogeneous system of (5) with $d = 0$.

In vectorial notation eqn. (5) is written

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{W} h\Big(\boldsymbol{x}(t)\Big) + \boldsymbol{I}, \qquad \boldsymbol{x}(0) = \boldsymbol{x_0} \tag{6}$$

with connectivity-matrix

$$\boldsymbol{W} = \begin{pmatrix} -d & -w & \cdots & -w \\ -w & -d & & \vdots \\ \vdots & & \ddots & -w \\ -w & \cdots & -w & -d \end{pmatrix} \tag{7}$$

3.2 Solution of the linear system

To solve (6), we start with the algebraic part of the problem and compute the eigenvalues and an eigenbasis of the connectivity-matrix $\boldsymbol{W}$.

Lemma 3.1 *Let $\boldsymbol{W}$ be a symmetric $n \times n$ matrix of the form*

$$\boldsymbol{W} = \begin{pmatrix} a & b & \cdots & b \\ b & a & & \vdots \\ \vdots & & \ddots & b \\ b & \cdots & b & a \end{pmatrix} \tag{8}$$

Then

$$\boldsymbol{v_1} = (-1, 1, 0, \ldots, 0)^T; \ldots; \boldsymbol{v_{n-1}} = (-1, 0, \ldots, 0, 1)^T; \boldsymbol{v_n} = (1, \ldots, 1)^T$$

is an eigenbasis of $\boldsymbol{W}$ *with corresponding eigenvalues*

$$\gamma_1 = \cdots = \gamma_{n-1} = a - b, \quad \gamma_n = a + (n-1)b.$$

Proof: Appendix A, p. 25. □

In the following let

$$\lambda_1 := w - d$$
$$\lambda_2 := -(n-1)w - d$$

denote the distinct eigenvalues of W with $\boldsymbol{v_1}, \ldots, \boldsymbol{v_{n-1}} \in \text{Eig}(\boldsymbol{W}, \lambda_1)$ and $\boldsymbol{v_n} \in \text{Eig}(\boldsymbol{W}, \lambda_2)$. Lemma 3.2 summarizes general statements we will occasionally refer to in our analysis.

Lemma 3.2 *Let* $\boldsymbol{x} \in \mathbb{R}^n$. *Then the following properties are satisfied*

(1) $\mathbb{R}^n = Eig(\boldsymbol{W}, \lambda_1) \oplus Eig(\boldsymbol{W}, \lambda_2)$
(2) $\bar{\boldsymbol{x}} \in Eig(\boldsymbol{W}, \lambda_2)$
(3) $\boldsymbol{x} - \bar{\boldsymbol{x}} \in Eig(\boldsymbol{W}, \lambda_1)$

where $\lambda_1 := w - d$ *and* $\lambda_2 := -(n-1)w - d$.

Proof: Appendix A, p. 25. □

Now we are able to present an exact solution of the inhomogeneous differential equation modeling the dynamics of our system.

Proposition 3.1 *Let* $w > d$, $\bar{\boldsymbol{x}}_{\boldsymbol{0}} = \frac{1}{n}\langle \boldsymbol{x_0}, \boldsymbol{v_n} \rangle \boldsymbol{v_n}$, *and* $\bar{\boldsymbol{I}} = \frac{1}{n}\langle \boldsymbol{I}, \boldsymbol{v_n} \rangle \boldsymbol{v_n}$. *The solution of the initial value problem defined in (6) is given by*

$$\boldsymbol{x}(t) = \left(\boldsymbol{x_0} - \bar{\boldsymbol{x}}_{\boldsymbol{0}} + \frac{\boldsymbol{I} - \bar{\boldsymbol{I}}}{\lambda_1} \right) e^{\lambda_1 t} + \left(\bar{\boldsymbol{x}}_{\boldsymbol{0}} + \frac{\bar{\boldsymbol{I}}}{\lambda_2} \right) e^{\lambda_2 t} - \frac{\boldsymbol{I} - \bar{\boldsymbol{I}}}{\lambda_1} - \frac{\bar{\boldsymbol{I}}}{\lambda_2}$$

Proof: Appendix A, p. 25. □

Before leaving this subsection we state the solution of the associated homogenous system of (6). The solution follows directly from Prop. 3.1 substituting $\boldsymbol{I} := \boldsymbol{0}$.

Corollary 3.1 *The solution of the associated homogeneous equation* $\dot{\boldsymbol{x}}(t) = \boldsymbol{W}\boldsymbol{x}(t)$ *with initial condition* $\boldsymbol{x_0} := \boldsymbol{x}(0)$ *is of the form*

$$\boldsymbol{x_h}(t) = (\boldsymbol{x_0} - \bar{\boldsymbol{x}}_{\boldsymbol{0}})e^{\lambda_1 t} + \bar{\boldsymbol{x}}_{\boldsymbol{0}}e^{\lambda_2 t}.$$

3.3 Analysis of the linear system

In the preceding section we introduced a linear model and presented an exact solution of its underlying dynamical system. The solution is the starting point of our examination of the network's behavior. In the subsequent paragraphs we derive the properties (P_1)-(P_3).

3.3.1 (P_1) Intensity normalization

A network normalizes its total intensity

$$\chi(t) = \sum_{i=1}^{n} x_i(t) \tag{9}$$

if $\chi(t)$ converges to a constant value, which is independent from the initial activation $\boldsymbol{x_0}$.

In the homogeneous case intensity normalization refers to a process shifting the total intensity $\chi(t)$ of the system to 0 as time t tends to infinity. In the inhomogeneous case the normalized intensity depends on the external input $\boldsymbol{I}$ and the parameters d and w. In both cases the intensity of the network converges to a constant, which is independent from the initial activation.

Proposition 3.2 (Intensity normalization) *Let* $w > d$ *and* $\boldsymbol{x}(t)$ *be the solution of (6). Then*

$$\lim_{t\to\infty} \sum_{i=1}^{n} x_i(t) = \frac{\bar{I}}{\lambda_2}$$

Proof: Appendix A, p. 26. □

Figure 1 gives an example of intensity normalization performed by a homogeneous linear neural network with $d < w$ consisting of two units. The *stable subspace* $E_s = \mathrm{Eig}(\boldsymbol{W}, \lambda_2)$ and the *unstable subspace* $E_u = \mathrm{Eig}(\boldsymbol{W}, \lambda_1)$ are defined by the equations

$x_1 - x_2 = 0$ and $x_1 + x_2 = 0$, respectively. Trajectories with $\boldsymbol{x_0} = \bar{\boldsymbol{x}}_{\boldsymbol{0}} \in E_s$ converge along E_s to the stable equilibrium $\mathbf{0}$ in direction of positive and negative multiples of the eigenvector $\boldsymbol{v_n} = \boldsymbol{v_2}$, and those with $\boldsymbol{x_0} = \mathbf{0} \in E_u$ diverge on the unstable hyperplane (line) defined by E_u. All other trajectories as the one shown in Figure 1 are superpositions of these motions.

The transformation T projects each point of the 2-dimensional plane orthogonally to the stable line E_s. With the time the projection $T\Big(\boldsymbol{x}(t)\Big)$ asymptotically converges along the stable line E_s to the origin $\mathbf{0}$, and therefore the trajectory $\boldsymbol{x}(t)$ asymptotically converges to the unstable subspace E_u. Intensity normalization can be seen in two ways: (1) Since $\boldsymbol{x}(t)$ asymptotically converges to the unstable line E_u defined by all points $\boldsymbol{x} = (x_1, x_2)$ with $x_1 + x_2 = 0$, the total intensity $\chi(t) = x_1(t) + x_2(t)$ asymptotically converges to 0. (2) Since the projected trajectory $T\Big(\boldsymbol{x}(t)\Big)$ converges to 0, the components $x_1(t)$ and $x_2(t)$ of $\boldsymbol{x}(t)$, and therefore $\chi(t)$ converge to 0.

The projection T reveals that an unstable system like the linear model bears an inherent stable behavior. Here this inherent stable behavior is regarded as intensity normalization.

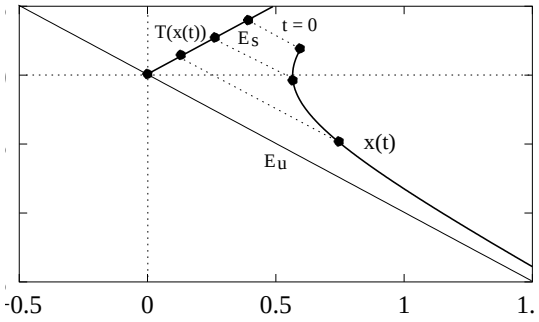

Fig. 1. Intensity normalization

3.3.2 (P_2) Contrast enhancement

Assume that the activations $x_i(t)$ of each unit i represent different brightness values. Then the magnitude of the difference

$$c_{ij}(t) = \Big| x_i(t) - x_j(t) \Big| \tag{10}$$

can be interpreted as the contrast between unit i and unit j. Then a network performs contrast enhancement if it satisfies

$$(CE_1) \quad x_{0i} < x_{0j} \Rightarrow c_{ij}(t_1) < c_{ij}(t_2)$$

$$(CE_2) \quad c_{kl}^{ij} = \frac{c_{ij}(t)}{c_{kl}(t)} \equiv \text{const.}$$

for all $i, j, k, l \in \{1, \ldots, n\}$ and $0 \leq t_1 < t_2$. The first property (CE_1) tells us that the network enlarges the contrasts and the second property (CE_2) says that the proportions of contrasts remain unchanged.

In the following we show that during evolution the linear network performs contrast enhancement.

To avoid copious distinction of cases and to keep the following statements and their proofs simple, we assume $x_{01} > x_{02} > \cdots > x_{0n}$ without restriction. Following statements and proofs can be adapted easily for cases with $x_{0i} = x_{0j}$ for $i \neq j$. Furthermore we assume $I_1 \geq I_2 \geq \cdots \geq I_n$. The next Lemma shows that the network preserves the order of the activations $x_i(t)$.

Lemma 3.3 *Let $x_{0i} < x_{0j}$. If $I_i \leq I_j$, then*

$$x_i(t) < x_j(t) \tag{11}$$

for $t \geq 0$.

Proof: Appendix A, p. 26. □

The proof of Lemma 3.3 reveals that the contrast $c_{ij}(t)$ between any pair of units i and j is monotonically increasing. This is exactly property CE_1.

Proposition 3.3 (Contrast enhancement – CE_1) *Let $x_{0i} < x_{0j}$. Then*

$$c_{ij}(t_1) < c_{ij}(t_2) \tag{12}$$

for $0 \leq t_1 < t_2$

Proof: Follows from the proof of Lemma 3.3. □

Figure 2 shows an example how a network with seven units enlarges contrasts according to CE_1. The initial activation $\boldsymbol{x_0} = \boldsymbol{x}(0)$ of the units is $x_{01} = 0.7$,

$x_{02} = 0.6, \ldots, x_{07} = 0.1$ and the weights are given by $w = 0.4$, $d = 0.2$. The vertical axis represents activation and the horizontal axis time. The different graphs show the STM traces of each unit as a function of time. Contrast enhancement increases the distance between two STM traces with time. To complete the proof that the

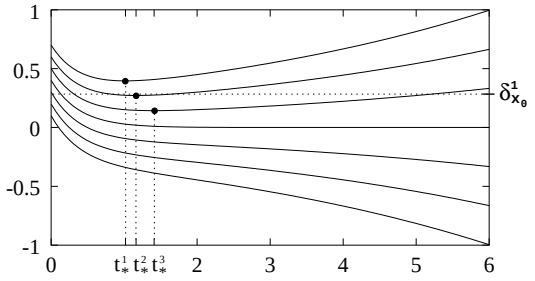

Fig. 2. STM traces of a seven unit linear network

linear model performs contrast enhancement we have to show that the proportions $c_{kl}^{ij}(t)$ of the contrasts remain unchanged.

Proposition 3.4 (Contrast enhancement – CE_2**)** *Let* $\boldsymbol{I} = \alpha \cdot \boldsymbol{x_0}$ *with* $\alpha \in \mathbb{R}$. *For all units* $1 \leq i, j, k, l \leq n$ *with* $k \neq l$ *the following identity holds:*

$$c_{kl}^{ij}(t) = \left| \frac{x_i(t) - x_j(t)}{x_k(t) - x_l(t)} \right| \equiv \left| \frac{x_{0i} - x_{0j}}{x_{0k} - x_{0l}} \right|.$$

Proof: Appendix A, p. 27. □

3.3.3 (P_3) Evidential Response

Evidential response is a capability of networks to provide information not only about which particular pattern to select, but also about the evidence for the choice made. In this section we show that at specific points of time the system signals the evidence for favorite units in terms of approximating a canonical discrimination measure. Throughout this subsection we assume $x_{01} > \cdots > x_{0n}$ and $I_{01} \geq \cdots \geq I_n$.

Property (P_3) is based on the assumption that the difference

$$\delta_{\boldsymbol{x_0}}^i := x_{0i} - \bar{x}_0$$

of the initial activation x_{0i} of unit i and the average of all initial activations $\bar{x}_0$ is considered to be a canonical measure, which provides information about the evidence

or the level of confidence. The value $\delta^i_{\boldsymbol{x_0}}$ signals how evident or how obvious choice i is in comparison to all other possible choices.

Let us call any unit i with initial activation above average ($x_{0i} > \bar{x}_0$) *dominating unit* and all other units with initial activation equal to or below average ($x_{0i} \leq \bar{x}_0$) *subdued unit*. A network provides *evidential responses* with respect to $\delta^i_{\boldsymbol{x_0}}$, if the the STM trace of dominating units i best approximate the desired response $\delta^i_{\boldsymbol{x_0}}$ at their minima.

For the seven unit linear model considered in Figure 2, the average initial activation is $\bar{x}_0 = (0.7 + \cdots + 0.1)/7 = 0.4$. Since $x_{01} = 0.7$, $x_{02} = 0.6$, and $x_{01} = 0.5$ are larger than $\bar{x}_0$, units 1-3 are dominating units. Similarly, from $x_{0i} \leq \bar{x}_0$ for $4 \leq i \leq 7$ follows that units 4-7 are subdued units.

Dominating and subdued units differ in their STM traces. The STM traces of dominating units have a unique global minimum while the activations of subdued units are monotonically decreasing (see Figure 2). The global minima constitute the best approximation of the measure $\delta_{\boldsymbol{x_0}}$. For this reason we first compute the global minima of dominating units (Prop. 3.5) and then show that they are the best possible approximation of $\delta_{\boldsymbol{x_0}}$ (Corollary 3.2).

Proposition 3.5 *Let $w > d$. Furthermore set*

$$\boldsymbol{u} := (\boldsymbol{I} - \bar{\boldsymbol{I}})/\lambda_1 + \bar{\boldsymbol{I}}/\lambda_2, \quad \boldsymbol{y} := \lambda_1(\boldsymbol{x_0} - \bar{\boldsymbol{x}}_{\boldsymbol{0}}) + \boldsymbol{I} - \bar{\boldsymbol{I}}, \quad \boldsymbol{z} := -\lambda_2 \bar{\boldsymbol{x}}_{\boldsymbol{0}} - \bar{\boldsymbol{I}}.$$

If the following conditions are satisfied

$$(1)\ x_{0i} > \bar{x}_0 \qquad (2)\ I_i \geq \bar{I} > 0 \qquad (3)\ z_i > 0,$$

then $x_i(t)$ has a global minimum at

$$t^i_* = \frac{1}{\lambda_2 - \lambda_1} \ln\left(\frac{y_i}{z_i}\right) \quad \textit{with} \quad x_i(t^i_*) = \frac{\lambda_2 - \lambda_1}{\lambda_1 \lambda_2} y_i^{\frac{\lambda_2}{\lambda_2 - \lambda_1}} z_i^{\frac{\lambda_1}{\lambda_1 - \lambda_2}} - u_i.$$

Proof: Appendix A, p. 27. □

Note, that condition (1) restricts the statement to dominating units. Condition (2) requires that the external input of dominating unit is above average $\bar{I}$ and condition (3) bounds the external input. Next we show in Corollary 3.2 that the STM traces

of dominating units i most exactly approximate the canonical measure $\delta^i_{\boldsymbol{x_0}}$ at their global minima.

Corollary 3.2 *Under the same assumption as in Prop. 3.5, the global minimum $x_i(t^i_*)$ of a dominating unit i is bounded from below by the canonical measure $\delta^i_{\boldsymbol{x_0}} = x_{0i} - \bar{x}_0$.*

Proof: Follows from Prop. 3.5 together with the inequality $x_i(t) > \delta^i_{\boldsymbol{x_0}}$ for all $i \in \{1, \ldots, n\}$ with $x_{i0} > \bar{x}_0$. □

Of course Corollary 3.2 is of no use, if the lower bound $\delta^i_{\boldsymbol{x_0}}$ is only a rough estimate of the global minimum $x_i(t^i_*)$. It is Prop. 3.6 which rejects this objection and leads to the most important result of our contribution. If we increase the number of units, then in the limit the global minimum $x_i(t)^i_*$ of a dominating unit i converges to the canonical measure $\delta^i_{\boldsymbol{x_0}}$.

Proposition 3.6 *Let $w > d$. Consider the conditions of Prop. 3.5. Then*

$$\lim_{n\to\infty} x_i(t^i_*) = x_{0i} - \bar{x}_0.$$

Proof: Appendix A, p. 28. □

Thus, the network does not only signal the favorite choices, but also generates information about how evident the choices are in comparison to all possible choices. The evidential responses are approximations of the canonical measure $\delta_{\boldsymbol{x_0}}$ and their accuracy can be controlled by adjusting the number of units in the network in a suitable manner. This observation suggests that one incorporates distributed redundancy to control the accuracy of the approximation of $\delta_{\boldsymbol{x_0}}$ generated by the discrimination network (see Section 4).

Figure 2 illustrates the evidential responses of the seven unit linear model considered previously. The global minima of the STM traces of dominating units 1, 2, 3 are labeled by filled circles at times t^1_*, t^2_*, and t^3_*. The dotted horizontal line shows the desired evidential response $\delta^1_{\boldsymbol{x_0}}$ for the winning unit 1 only. We see that $\delta^1_{\boldsymbol{x_0}}$ is a lower bound of $x_1(t)$ which is best approximated at its minimum $x_1(t^1_*)$.

In addition, the network extracts further characteristic information about the input patterns. From Prop. 3.5 we know that $x_i(t^i_*) \approx x_{0i} - \bar{x}_0$. Therefore the system is able to approximate the average $\bar{x}_0$, if it is possible to reconstruct the initial activation

$\boldsymbol{x_0}$. Furthermore, if $1 > x_{01} > \ldots > x_{0n} > 0$, then it can be shown easily that $(x_1(t_*^1))^2 > \sum_i (x_i - \bar{x})^2/n$. Thus the activation of the winning unit at t_*^1 provides an upper bound of the standard deviation. This knowledge provides further information about the evidence for a particular choice.

We conclude this section with an important auxiliary result. Lemma 3.4 proves that the chronological order of STM traces arriving at their global minima is determined by the order of the initial activations (see also Figure 2). This result provides a basis to *employ differentiating units* to signal the winner and to identify the point of time when the STM trace of the winner arrives at its global minimum. The ability of differentiating units to identify local minima gives us direct access to a sensible evidential response in terms of an approximation of $\delta_{\boldsymbol{x_0}}$.

Lemma 3.4 *Consider the assumptions in Prop. 3.5. Let* $x_{0i} > x_{0j} > \bar{x}_0$ *and* $I_i > I_j > \bar{I}$. *If* $\lambda_1 = w - d > 0$, *then*

$$t_*^i < t_*^j.$$

Proof: Appendix A, p. 29. □

4 Analysis of the nonlinear discrimination network

Now we are able to show that the nonlinear discrimination network behaves as follows: Given an initial activation $\boldsymbol{x_0}$ the network normalizes its intensity (P_1) and enhances contrasts (P_2) until it converges to an equilibrium point within finite time. After convergence the activations of dominating units provide evidential responses (P_3) with respect to the canonical measure $\delta_{\boldsymbol{x_0}}^i = x_{0i} - \bar{x_0}$.

For convenience we use the following notations: We denote by index i_* the most dominating (winning) unit. The most dominating unit is the unit with maximal initial activation. Index $j \neq i_*$ is reserved for the remaining dominating units, index k for subdued units, and index i for any unit without further specification of the category it belongs to. For the winner i_* we simply write t_* and $\delta_{\boldsymbol{x_0}}^*$ instead of $t_*^{i_*}$ and $\delta_{\boldsymbol{x_0}}^{i_*}$, respectively.

From Lemma 3.3 and 3.4 we know, that the STM trace of the most dominating unit i_* wins the competition and first arrives at its minimum at time t_* (see also Figure

2). According to (2) the output signal of the winner i_* is given by

$$y_{i_*}(t_*) = x_{i_*}(t_*) \cdot f\Big(\dot{x}_{i_*}(t_*)\Big).$$

Since $x_{i_*}(t_*)$ is a global minimum of $x_{i_*}(t)$ (Prop. 3.5), the derivative $\dot{x}_{i_*}(t)$ vanishes at t_*. From the dynamical system (2) and (3) of the discrimination network follows that the output signal $y_{i_*}(t_*)$ of unit i_* is zero. The zero output of the winner removes the strongest inhibitory input from all units $i \neq i_*$ and leads to an increase of their activations after a period of continuous decay. As soon as their differentiating unit detects a turning point in its STM trace, it switches off its output signal. Since any unit receives a zero input from all other units, the change of activation remains zero and thus the discrimination network has converged to an equilibrium point (see Figure 3).

After convergence the activation of the winner is given by Prop. 3.5 with $d = 0$ and $\mathbf{I} = \mathbf{0}$. From Corollary 3.2 we know that the winner i_* best approximates the desired response $\delta^*_{\boldsymbol{x_0}}$ at the minimum $x_{i_*}(t_*)$ of its STM trace. The activation $x_{i_*}(t_*)$ may be regarded as an evidential response approximating a natural level of confidence given by $\delta^*_{\boldsymbol{x_0}} = x_{0i_*} - \bar{x}_0$. After convergence the activation $x_j(t_*)$ of the remaining dominating units j are slightly larger than their best possible approximation $x_j(t^j_*)$ of $\delta^j_{\boldsymbol{x_0}}$. This is due to the following facts:

(1) The STM trace of the winner i_* arrives first at its minimum at t_*. Then we have $t_* < t^j_*$.
(2) The STM traces $x_j(t)$ of the dominating units $j \neq i_*$ are monotonically decreasing during the period $0 \leq t < t_*$. Taking into account that $t_* < t^j_*$ holds, we obtain $x_j(t_*) > x_j(t^j_*)$.
(3) According to Corollary 3.2 the value $\delta^j_{\boldsymbol{x_0}}$ is a lower bound of $x_j(t)$. Then from fact (2) finally follows $x_j(t_*) > x_j(t^j_*) > \delta^j_{\boldsymbol{x_0}}$.

The accuracy with which the activation $x_j(t_*)$ of a dominating unit j approximates the desired evidential response $\delta^j_{\boldsymbol{x_0}}$ degrades, if the choice j is less evident relative to the choice i_*, i.e. if $x_{0j} - \bar{x}_0$ is significantly smaller than $x_{0i_*} - \bar{x}_0$. In this sense the discrimination network generates more accurate evidential responses for dominating units $j \neq i_*$ for which x_{0j} is close to x_{0i_*}. The accuracy of the evidential responses can be improved by incorporating distributed redundancy into the network, as we shall see shortly.

Since the discrimination network behaves identical to a linear network until the activation of the winning unit becomes minimal at t_*, it also has properties (P_1)-

(P_3) during the time interval $[0, t_*]$. Thus after imposing an initial activation $\boldsymbol{x_0}$ the total intensity $\chi(t) \geq 0$ of the network continuously decreases (P_1) and enhances contrasts (P_2) until t_*. At t_* the activations of dominating units are the best possible approximations of the canonical measure $\delta^i_{\boldsymbol{x_0}} = x_{0i} - \bar{x}_0$ the network can achieve (P_3). Prop. 4.1 summarizes the statements mentioned.

Proposition 4.1 *Consider the dynamical system given by (1) and (2). For a given initial activation $x_{01} > \cdots > x_{0n}$ the discrimination network converges at time t^1_* to an equilibrium point satisfying properties (P_1)-(P_3) during the time interval $[0, t^1_*]$.*

Proof: Appendix B, p. 29. □

Differentiating units are responsible for bounded activation and convergence to an equilibrium point within finite time. After convergence the activation of dominating units j approximate the desired evidence $\delta^j_{\boldsymbol{x_0}}$. Thus, *a network using differentiating units is capable to signal the best choices and provides information about the evidence for the selection made.*

Figure 3 shows the STM traces of a discrimination network consisting of seven units. The units are initialized as in the example given in Figure 2. Here the connective weight is given by $w = 0.2$. In a computer simulation we approximated the dynamics of the proposed discrimination network given by the iterative system

$$x_i(t+1) = x_i(t) - \tau w \sum_{j \neq i} y_j(t) \tag{13}$$

$$y_i(t+1) = x_i(t) f\Big(x_i(t) - x_i(t-1)\Big) \tag{14}$$

using a time constant $\tau = 0.01$. The vertical axis represents the activation of the STM traces and the horizontal axis the time. The different graphs show how the STM traces of each unit proceed as a function of the time. Since the connective weight w is smaller than in the example given in Figure 2 the STM trace of unit 1 requires more time to arrive at its minimum. At $t_* = t^1_*$ the system has converged. After convergence the activation $x_1(t)$ of the winner is equal to the activation $x_1(t_*)$ of unit 1 in the associated linear model. The activations $x_j(t)$ of the remaining dominating units are slightly larger than the activations $x_j(t^j_*)$ of the same units in the linear case.

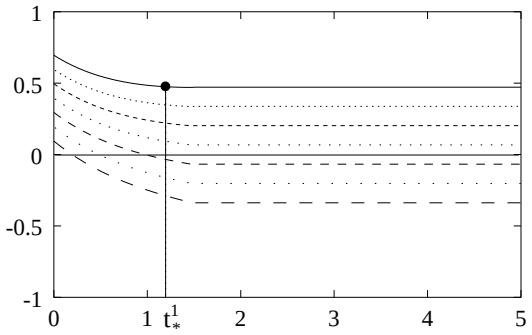

Fig. 3. STM traces of a seven unit discrimination network

Incorporating distributed redundancy: In order to control the accuracy of the approximation of $\delta_{\boldsymbol{x_0}}$ generated by the discrimination network we increase the number of units by means of distributed redundancy

The network can be enlarged without changing the behavior of the STM traces by the following procedure. Increase the number n of units by adding artificial units without changing the average activation $\bar{x}_0$. This can be achieved by connecting a fixed number k of copies of each unit to the net where the activations of the copies are slightly perturbed by a random noise. This enlarges the system to kn fully connected units consisting of n groups each with k units of nearly equal initial activation. If the noise is bounded by an interval $[-\varepsilon, +\varepsilon]$ with expectation 0 and small bound $\varepsilon \ll \min\{x_{0i} \mid 1 \leq i \leq n\}$ compared to the initial activation of the original units, then the expectation of $\bar{x}_0$ of the enlarged system corresponds to the average of the original system. Now let x_{0i} be the initial activation of a representative of group i. Then at t_*^i the activations of group i retrieve a better averaged approximation of $x_{0i} - \bar{x}_0$ than unit i of the original system (Prop. 3.6). From Prop. 3.5 we can conclude $\lim_{n\to\infty} t_*^1 = 0$.

Consequently, enlarging the network (1) improves the accuracy of an evidential response with respect to the canonical measure $\delta_{\boldsymbol{x_0}}$ and (2) accelerates the decision process. To offer a physiological interpretation for this observation, let us explain the accuracy of an evidential response as a focus setting. Then in a physiological manner *increasing distributed redundancy does not only lead to a system which is robust against failure of single units, but also sharpens the focus on the problem given.*

5 Evidential Learning with Discrimination Networks

In this section we show that the proposed discrimination network can be regarded as a neural network model for competitive learning by evidence.

In its simplest form a competitive learning neural network consists of a single layer of inhibitory connected output units $C = \{c_1, \ldots, c_n\}$. With each unit c_i a model $\boldsymbol{y_i} \in Y = \{\boldsymbol{y_1}, \ldots \boldsymbol{y_n}\} \subseteq \mathbb{R}^d$ is associated. Each output unit c_i is fully connected to a set of m input units x_j where d is the dimension of the input space. Given a set $X = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_p}\} \subseteq \mathbb{R}^d$ of data points, the competitive learning algorithm moves the models $\boldsymbol{y_i} \in Y$ to centers of the clusters in the input data by using a *competitive learning rule*. Once the competitive learning algorithm has converged, it can be applied as a pattern classifier where an input pattern is assigned to the class represented by the winner of the competition.

Evidential learning adopts the perception that models, which are more similar to the current input $\boldsymbol{x}$ should be moved closer to $\boldsymbol{x}$ and models, which are too dissimilar to $\boldsymbol{x}$ remain unchanged. Figure 4 outlines the basic procedure of competitive learning by evidence.

Evidential learning differs from simple competitive learning only in the particular choice of the learning rates. Models $\boldsymbol{y_i}$ associated with dominating units c_i are updated by using the following learning rule

$$\boldsymbol{y_i} \quad \leftarrow \quad \boldsymbol{y_i} + \eta(t) \cdot \delta^i_{\boldsymbol{x_0}} \cdot \Big(\boldsymbol{x} - \boldsymbol{y_i}\Big) \tag{15}$$

where $\eta(t) \in [0, 1]$ is a monotonically decreasing control parameter and $\delta^i_{\boldsymbol{x_0}}$ is the evidence

$$\delta^i_{\boldsymbol{x_0}} = \langle \boldsymbol{x}, \boldsymbol{y_i} \rangle - \frac{1}{n} \sum_{j=1}^{n} \langle \boldsymbol{x}, \boldsymbol{y_j} \rangle. \tag{16}$$

According to the update rule (15) evidential learning moves models $\boldsymbol{y_i}$ associated with dominating units c_i closer towards the current input data $\boldsymbol{x}$. The learning rate $\eta_i(t)$ controls how close a dominating model $\boldsymbol{y_i}$ is moved towards $\boldsymbol{x}$. In the extreme case where $\eta(t) = 0$, no updating takes place. On the other hand, if $\eta(t) = 1$ the evidential learning rule (15) moves dominating models $\boldsymbol{y_i}$ with higher evidence closer to $\boldsymbol{x}$ where the evidence for a dominating unit c_i is given by $\delta^i_{\boldsymbol{x_0}}$.

The evidence for a dominating unit c_i reflects the relative level of confidence that a

given data point $\boldsymbol{x}$ belongs to the category represented by c_i, relative with respect to all categories. According to that confidence the associated model $\boldsymbol{y_i}$ is moved toward $\boldsymbol{x}$ to make unit c_i more likely to dominate on that input in the future. After convergence of evidential learning unseen data can be categorized indicating the level of confidence.

Evidential Learning Algorithm:

Let $X = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_p}\} \subseteq \mathbb{R}^d$ be a set of data points and $Y = \{\boldsymbol{y_1}, \ldots \boldsymbol{y_n}\} \subseteq \mathbb{R}^d$ a set of models.

(1) Initialize $Y = \{\boldsymbol{y_1}, \ldots \boldsymbol{y_n}\}$.

(2) **Repeat**

(a) Randomly select a data point $\boldsymbol{x} \in X$.

(b) Find all dominating units c_i for which

$$\langle \boldsymbol{x}, \boldsymbol{y_i} \rangle > \frac{1}{n} \sum_{j=1}^{n} \langle \boldsymbol{x}, \boldsymbol{y_j} \rangle$$

(c) Adjust the associated dominating models $\boldsymbol{y_i}$ by using the update formula

$$\boldsymbol{y_i} \leftarrow \boldsymbol{y_i} + \eta(t) \cdot \delta^i_{\boldsymbol{x_0}} \cdot (\boldsymbol{x} - \boldsymbol{y_i})$$

(d) Decrease the learning rate $\eta(t)$.

until no noticeable changes in Y are observed.

(3) Output cluster centers $Y = \{\boldsymbol{y_1}, \ldots \boldsymbol{y_n}\}$.

Fig. 4. Outline of evidential learning.

In simple competitive learning, it does not matter how the winner-take-all character for maximum selection is implemented. In a practical setting one prefers a simple search algorithm for finding the maximum. In a neural network implementation the winner-take-all circuit is usually implemented by a (digital or analog) inhibitory WTA network like MAXNET (Lippman (1987)).

In a similar manner as in simple competitive learning the selection of dominating units together with their evidential responses is algorithmically trivial. However, as MAXNET is a neural network implementation of the winner-takes-all circuit modeling simple competitive learning the neural network implementation of evidential

learning is based on the proposed discrimination network. The evidential learning network consists of the discrimination network where each unit c_i is fully connected to a set of inputs x_j via excitatory connections with weight y_{ij}. Applying a given data point $\boldsymbol{x}$ to the input units of the evidential learning networks gives an initial activation $x_{0i} = \langle \boldsymbol{x}, \boldsymbol{y_i} \rangle$ of each unit c_i of the discrimination network. Then the discrimination network evolves by normalizing the total intensity (P_1) and enhancing contrasts (P_2) until convergence to an equilibrium state. After convergence the activation of dominating units c_i is approximately

$$x_{0i} - \bar{x}_0 = \langle \boldsymbol{x}, \boldsymbol{y_i} \rangle - \frac{1}{n} \sum_j \langle \boldsymbol{x}, \boldsymbol{y_j} \rangle$$

where $\bar{x}_0 = 1/n \cdot \sum_j \langle \boldsymbol{x}, \boldsymbol{y_j} \rangle$ is the average initial activation.

6 Conclusion

In this paper we introduced and rigorously analyzed a discrimination network consisting of fully connected differentiating units competing among each other for the most evidential response. In contrast to traditional WTA architectures for maximum selection the discrimination network does not only signal the winning unit, but also generates an evidential response. Furthermore, the role of distributed redundancy in discrimination networks is not only restricted to provide a robust and fault tolerant system as in classical competitive network approaches. Distributed redundancy plays also an important role in focusing on the maximum selection problem. Discriminant networks are a neural network implementation of evidential learning, a modified version of simple competitive learning replacing the extreme selection principle of *winner takes all* by an activity related soft selection principle.

The mathematical analysis revealed that at the beginning the network stores information provided in form of an initial activation in the short-term-memory. This information is processed by normalizing the total intensity (P_1) and enhancing contrasts (P_2). During this process the network generates information about the evidence for the winning unit by approximating the canonical measure $\delta_{\boldsymbol{x_0}}$ (P_3). In addition characteristic information about the input patterns is extracted, particularly an approximation of the average and an upper bound of the standard deviation of the initial activation. Both characteristic values are incorporated in the information signaling the evidence for a choice. The accuracy of the evidential response with

respect to the canonical measure δ_{x_0} can be controlled by distributed redundancy. The system converges to an equilibrium point when the activation of the winning unit arrives at its minimum.

References

Ermentrout, B. (1992). Complex Dynamics in Winner-Take-All Neural Nets with Slow Inhibition. *Neural Networks*, 5, 415-431.

Fang, Y., Cohen, M.A., & Kincaid, T.G. (1996) Dynamics of a Winner-Take-All Neural Network. *Neural Networks*, 9, 1141-1154.

Feldman, J.A., & Ballard, D.H. (1982). Connectionist Models and their Properties. *Cognitive Science*, 6, 205-254.

Grossberg, S. (1988). Nonlinear Neural Networks: Principles, Mechanisms, and Architectures. *Neural Networks*, 1, 17-61.

Hahnloser, R. (1998). On the Piecewise Analysis of Networks of Linear Threshold Neurons. *Neural Networks*, 11, 691-697.

Haykin, S. (1999). Neural Networks. Prentice Hall, Inc, 2nd edition.

Hirsch, M.W. (1989). Convergent Activation Dynamics in Continuous Time Networks. *Neural Networks*, 2, 331-349.

Hopfield, J.J., & Tank, D.W. (1986). Computing with Neural Circuits: A Model. *Science*, 223, 625-633.

Kohonen, T. (1984). Self-Organization and Associative Memory. Springer-Verlag, Berlin.

Lemon, M., & Vijaya Kumar, B.V.K. (1989). Emulating the Dynamics for a Class of Laterally Inhibited Neural Networks. *Neural Networks*, 2, 193-214.

Levine, D.S. (2000). Introduction to Neural and Cognitive Modeling. Lawrence Erlbaum Associates, New Jersey.

Lippman, R.P. (1987). An Introduction to Computing with Neural Nets. *IEEE ASSP Magazine*, April, 4-22.

Lippman, R.P., Gold, B., & Malpass, M.L. (1988). A Comparison of Hamming and Hopfield Neural Nets for pattern classification. *MIT Lincoln Laboratory Technical Report*, TR-769.

Maass, W. (2000). On the computational power of winner-take-all. *Neural Computation*, 12, (11), 2519-2536.

Maass, W. (2000). Neural Computation with Winner-Take-All as the only Nonlinear Operation. In Solla, S.A., Leen, T.K.,& Müller, K.R. (Eds.), *Advances in Information Processing Systems 12* (pp. 293-299), MIT Press, Cambridge.

Majani, E., Erlanson, R., & Abu-Mostafa, Y. (1989). On the k-winner-take-all network. *Advances in Neural Information Processing Systems*, 1, 634-642.

Wersing, H., Beyn, W.-J, & Ritter, H. (2001). Dynamical Stability Conditions for Recurrent Neural Networks with Unsaturating Piecewise Linear Transfer Functions. *Neural Computation*, 13, 1811-1825.

Xie, X., Hahnloser, R., & Seung, S. (2000). Learning Winner-Take-All competition between Groups of Neurons in Lateral Inhibitory Networks. In Leen, T.K., Dietterich, T.G., & Tresp, V. (Eds.), *Advances in Neural Information Processing Systems 13* (pp. 350–356). MIT Press, Cambridge.

A Proofs of Section 3.1

Proof of Lemma 3.1:
Check $\boldsymbol{W}\boldsymbol{v_i} = \gamma_i \boldsymbol{v_i}$ for all $i \in \{1, \ldots, n\}$ to show that $\boldsymbol{v_i}$ are eigenvectors of $\boldsymbol{W}$ with eigenvalues γ_i. Furthermore the eigenvectors $\boldsymbol{v_1}, \ldots, \boldsymbol{v_n}$ are linearly independent: Let $\boldsymbol{e_1}, \ldots, \boldsymbol{e_{n-1}}$ be the canonical basis of $\mathbb{R}^{n-1}$. The linear mapping

$$F : \mathbb{R}^{n-1} \to \mathbb{R}^n, \qquad (x_1, \ldots, x_{n-1}) \mapsto \left(-\sum_{i=1}^{n-1} x_i, x_1, \ldots, x_{n-1} \right)$$

is injective with $F(\boldsymbol{e_i}) = \boldsymbol{v_i}$ for all $i \in \{1, \ldots, n-1\}$. Consequently $\boldsymbol{v_1}, \ldots, \boldsymbol{v_{n-1}}$ are linearly independent. Since $\boldsymbol{W}$ is symmetric, $\mathbb{R}^n$ is the orthogonal sum of the eigenspaces of $\boldsymbol{W}$ belonging to its distinct eigenvalues. Hence $\boldsymbol{v_1}, \ldots, \boldsymbol{v_n}$ are linearly independent and therefore form a basis of $\mathbb{R}^n$. □

Proof of Lemma 3.2:
(1) Since $\boldsymbol{W}$ is symmetric, $\mathbb{R}^n$ is the orthogonal sum of the eigenspaces $\mathrm{Eig}(\boldsymbol{W}, \lambda_1)$ and $\mathrm{Eig}(\boldsymbol{W}, \lambda_2)$. (2) By definition we have $\bar{\boldsymbol{x}} = (\bar{x}, \ldots, \bar{x})$. Thus $\bar{\boldsymbol{x}} = \bar{x} \cdot \boldsymbol{v_n} \in \mathrm{Eig}(\boldsymbol{W}, \lambda_2)$. (3) From

$$\langle \boldsymbol{x} - \bar{\boldsymbol{x}}, \boldsymbol{v_n} \rangle = \sum_{i=1}^{n} (x_i - \bar{x}) \cdot 1 = \sum_{i=1}^{n} x_i - n\bar{x} = 0$$

follows $\boldsymbol{x} - \bar{\boldsymbol{x}}$ is orthogonal to $\mathrm{Eig}(\boldsymbol{W}, \lambda_2) = \mathbb{R}\boldsymbol{v_n}$. Hence $\boldsymbol{x} - \bar{\boldsymbol{x}} \in \mathrm{Eig}(\boldsymbol{W}, \lambda_1)$. □

Proof of Proposition 3.1:
The proof is a straightforward matter of differentiating the solution and plugging the derivative into the differential equation to check whether it is satisfied:

$$\dot{\boldsymbol{x}}(t) = \lambda_1 \left(\boldsymbol{x_0} - \bar{\boldsymbol{x}}_{\boldsymbol{0}} + \frac{\boldsymbol{I} - \bar{\boldsymbol{I}}}{\lambda_1} \right) e^{\lambda_1 t} + \lambda_2 \left(\bar{\boldsymbol{x}}_{\boldsymbol{0}} + \frac{\bar{\boldsymbol{I}}}{\lambda_2} \right) e^{\lambda_2 t} = \boldsymbol{W}\boldsymbol{x}(t) + \boldsymbol{I}.$$

Let us put

$$\boldsymbol{a} := \left(\boldsymbol{x_0} + \frac{\boldsymbol{I}}{\lambda_1} \right) e^{\lambda_1 t}, \quad \boldsymbol{b} := \left(\bar{\boldsymbol{x}}_{\boldsymbol{0}} + \frac{\bar{\boldsymbol{I}}}{\lambda_2} \right) e^{\lambda_2 t}, \quad \boldsymbol{c} := \frac{\boldsymbol{I}}{\lambda_1}, \quad \boldsymbol{d} := \frac{\bar{\boldsymbol{I}}}{\lambda_2}.$$

Then, since $\bar{\boldsymbol{b}} = \boldsymbol{b}$ and $\bar{\boldsymbol{d}} = \boldsymbol{d}$, we find

$$\boldsymbol{x}(t) = \boldsymbol{a} - \bar{\boldsymbol{a}} + \bar{\boldsymbol{b}} - (\boldsymbol{c} - \bar{\boldsymbol{c}}) - \bar{\boldsymbol{d}} \qquad \text{and} \qquad \dot{\boldsymbol{x}}(t) = \lambda_1(\boldsymbol{a} - \bar{\boldsymbol{a}}) + \lambda_2\bar{\boldsymbol{b}}.$$

By Lemma 3.2 we have

$$\boldsymbol{W}\boldsymbol{x} = \lambda_1(\boldsymbol{a} - \bar{\boldsymbol{a}}) + \lambda_2\bar{\boldsymbol{b}} - \lambda_1(\boldsymbol{c} - \bar{\boldsymbol{c}}) - \lambda_2\bar{\boldsymbol{d}}.$$

Since $-\lambda_1(\boldsymbol{c} - \bar{\boldsymbol{c}}) - \lambda_2\bar{\boldsymbol{d}} = -\boldsymbol{I} + \bar{\boldsymbol{I}} - \bar{\boldsymbol{I}} = -\boldsymbol{I}$, we obtain $\dot{\boldsymbol{x}}(t) = \boldsymbol{W}\boldsymbol{x}(t) + \boldsymbol{I}$. □

Proof of Proposition 3.2:
We have to show $\lim_{t\to\infty}\langle \boldsymbol{v_n}, \boldsymbol{x}(t)\rangle = \bar{\boldsymbol{I}}/\lambda_2$ where $\boldsymbol{v_n} = (1, \ldots, 1) \in \mathrm{Eig}(\boldsymbol{W}, \lambda_2)$ is the n-th eigenvector of $\boldsymbol{W}$. Let us put

$$\boldsymbol{a} := \boldsymbol{x_0} + \frac{\boldsymbol{I}}{\lambda_1}, \qquad \boldsymbol{b} := \bar{\boldsymbol{x}}_{\boldsymbol{0}} + \frac{\bar{\boldsymbol{I}}}{\lambda_2}, \qquad \boldsymbol{c} := \frac{\boldsymbol{I}}{\lambda_1}, \qquad \boldsymbol{d} := \frac{\bar{\boldsymbol{I}}}{\lambda_2}.$$

Then the solution of eqn. (6) is of the form

$$\boldsymbol{x}(t) = (\boldsymbol{a} - \bar{\boldsymbol{a}})e^{\lambda_1 t} + \boldsymbol{b}e^{\lambda_2 t} - (\boldsymbol{c} - \bar{\boldsymbol{c}}) - \boldsymbol{d}.$$

where $\boldsymbol{a} - \bar{\boldsymbol{a}}$ and $\boldsymbol{c} - \bar{\boldsymbol{c}}$ are vectors of the eigenspace $\mathrm{Eig}(\boldsymbol{W}, \lambda_1)$. Hence, from Lemma 3.2 follows $\langle \boldsymbol{v_n}, \boldsymbol{a} - \bar{\boldsymbol{a}}\rangle = \langle \boldsymbol{v_n}, \boldsymbol{c} - \bar{\boldsymbol{c}}\rangle = 0$. We have

$$\langle \boldsymbol{v_n}, \boldsymbol{x}(t)\rangle = \langle \boldsymbol{v_n}, \boldsymbol{b}\rangle e^{\lambda_2 t} - \langle \boldsymbol{v_n}, \boldsymbol{d}\rangle.$$

The assumption follows from $\lim_{t\to\infty} e^{\lambda_2 t} = 0$ together with $\bar{\boldsymbol{d}} = \boldsymbol{d}$. □

Proof of Lemma 3.3:
Assume $\lambda_1 \neq 0$. Using the solutions $\boldsymbol{x}(t)$ given in Prop. 3.1, we see

$$\begin{aligned} x_i(t) - x_j(t) &= \left(x_{0i} - x_{0j} + \frac{I_i - I_j}{\lambda_1}\right) e^{\lambda_1 t} - \frac{I_i - I_j}{\lambda_1} \\ &= (x_{0i} - x_{0j})e^{\lambda_1 t} + \frac{I_i - I_j}{\lambda_1}(e^{\lambda_1 t} - 1). \end{aligned}$$

By assumption $x_{0i} - x_{0j} < 0$ and $I_i - I_j \leq 0$. Since $\lambda_1 > 0$, we have $\frac{I_i - I_j}{\lambda_1} \leq 0$ and $e^{\lambda_1 t} - 1 \geq 0$ for $t \geq 0$. Putting all these inequalities together gives $x_i(t) - x_j(t) < 0$ for $t \geq 0$. □

Proof of Proposition 3.4:
Since $k \neq l$ and by assumption $x_{0k} \neq x_{0l}$, the function c_{kl}^{ij} is well-defined. Let $\lambda_1 \neq 0$. Like in the proof of the previous Lemma, subtracting the ith and jth component of the solution $\boldsymbol{x}(t)$ (see Prop. 3.1) leads to

$$\begin{aligned} x_i(t) - x_j(t) &= \left(x_{0i} - x_{0j} + \frac{I_i - I_j}{\lambda_1}\right) e^{\lambda_1 t} - \frac{I_i - I_j}{\lambda_1} \\ &= \left(x_{0i} - x_{0j} + \frac{\alpha x_{0i} - \alpha x_{0j}}{\lambda_1}\right) e^{\lambda_1 t} - \frac{\alpha x_{0i} - \alpha x_{0j}}{\lambda_1} \\ &= (x_{0i} - x_{0j}) \frac{(\lambda_1 + \alpha) e^{\lambda_1 t} - \alpha}{\lambda_1}. \end{aligned}$$

Clearly, $\kappa(t) := ((\lambda_1 + \alpha)e^{\lambda_1 t} - \alpha)/\lambda_1$ is independent of unit i and j. Hence,

$$c_{kl}^{ij}(t) = \left|\frac{x_i(t) - x_j(t)}{x_k(t) - x_l(t)}\right| = \left|\frac{(x_{0i} - x_{0j})\kappa(t)}{(x_{0k} - x_{0l})\kappa(t)}\right| = \left|\frac{x_{0i} - x_{0j}}{x_{0k} - x_{0l}}\right|.$$

□

Proof of Proposition 3.5:
Since $w > d$, we find $\lambda_1 \neq 0, \lambda_2 \neq 0$. Then by Prop. 3.1 the solution is of the form

$$\begin{aligned} x_i(t) &= \left(x_{0i} - \bar{x}_0 + \frac{I_i - \bar{I}}{\lambda_1}\right) e^{\lambda_1 t} + \left(\bar{x}_0 + \frac{\bar{I}}{\lambda_2}\right) e^{\lambda_2 t} - \frac{I_i - \bar{I}}{\lambda_1} - \frac{\bar{I}}{\lambda_2} \\ &= \frac{1}{\lambda_1} y_i e^{\lambda_1 t} - \frac{1}{\lambda_2} z_i e^{\lambda_2 t} - u_i. \end{aligned}$$

Differentiating $x_i(t)$ and equating with 0, we find

$$\dot{x}_i(t) = y_i e^{\lambda_1 t} - z_i e^{\lambda_2 t} = 0.$$

By assumption $z_i \neq 0$, so that

$$e^{(\lambda_2 - \lambda_1)t} = \frac{y_i}{z_i}.$$

Applying ln(.) to both sides yields

$$(\lambda_2 - \lambda_1)t = \ln\left(\frac{y_i}{z_i}\right).$$

Note that the term on the right side of the equation is defined, since $y_i/z_i > 0$. Dividing by $\lambda_2 - \lambda_1 = -nw \neq 0$ gives t_*^i. Since $\dot{x}_i(t) = 0$ is a necessary but not sufficient condition for a local extremum, we have to check the second derivative:

$$\ddot{x}_i(t) = \lambda_1^2 \left(x_{0i} - \bar{x}_0 + \frac{I_i - \bar{I}}{\lambda_1} \right) e^{\lambda_1 t} + \lambda_2^2 \left(\bar{x}_0 + \frac{\bar{I}}{\lambda_2} \right) e^{\lambda_2 t}.$$

Now with condition (1) and (2), we obtain $\dot{x}_i(t) > 0$ for $t \in \mathbb{R}$. Hence, $x_i(t)$ has a local minimum at t_*^i. Clearly, t_*^i is the global minimum, since $x_i(t)$ is continuously differentiable and $\dot{x}_i(t)$ has its unique zero at t_*^i.

To conclude the proof, we compute the value of $x_i(t)$ at t_*^i. Plugging $\ln(y_i/z_i)/(\lambda_2 - \lambda_1)$ into $x_i(t)$ yields

$$\begin{aligned} x_i(t_*^i) &= \frac{y_i}{\lambda_1} \exp\left(\frac{\lambda_1}{\lambda_2 - \lambda_1} \ln\left(\frac{y_i}{z_i} \right) \right) - \frac{z_i}{\lambda_2} \exp\left(\frac{\lambda_2}{\lambda_2 - \lambda_1} \ln\left(\frac{y_i}{z_i} \right) \right) - u_i \\ &= \frac{\lambda_2 - \lambda_1}{\lambda_1 \lambda_2} y_i^{\lambda_2/(\lambda_2 - \lambda_1)} z_i^{\lambda_1/(\lambda_1 - \lambda_2)} - u_i. \end{aligned}$$

□

Proof of Proposition 3.6:
From Prop. 3.5 we know

$$x_i(t_*^i) = \frac{\lambda_2 - \lambda_1}{\lambda_1 \lambda_2} y_i^{\lambda_2/(\lambda_2 - \lambda_1)} z_i^{\lambda_1/(\lambda_1 - \lambda_2)} - u_i.$$

Let us check the convergence for each term separately. First consider

$$\lim_{n \to \infty} \frac{\lambda_2 - \lambda_1}{\lambda_1 \lambda_2} = \lim_{n \to \infty} \frac{1}{\lambda_1} - \frac{1}{\lambda_2} = \frac{1}{\lambda_1}.$$

From $\lambda_2/(\lambda_2 - \lambda_1) = ((n-1)w + d)/nw$ and $\lim_{n \to \infty} \lambda_2/(\lambda_2 - \lambda_1) = 1$ follows

$$\lim_{n \to \infty} y_i^{\lambda_2/(\lambda_2 - \lambda_1)} = y_i.$$

Note, that $\lambda_1/(\lambda_1 - \lambda_2) = (w - d)/nw = (1 - d)/n$ and $0 \leq d < 1$. Then

$$1 = z_i^0 \leq z_i^{(1-d)/n} \leq (nw\bar{x})^{1/n}.$$

Since $n^{1/n} \to 1$, we have $(nw\bar{x})^{1/n} \to 1$ and therefore $z_i^{(1-d)/n} \to 1$ for $n \to \infty$. For the last term u_i of $x_i(t_*^i)$, we see

$$\lim_{n\to\infty} u_i = \lim_{n\to\infty} \frac{I_i - \bar{I}}{\lambda_1} + \frac{\bar{I}}{\lambda_2} = \frac{I_i - \bar{I}}{\lambda_1}.$$

Finally, putting all these limits together, we obtain

$$\lim_{n\to\infty} x_i(t_*^i) = \frac{y_i}{\lambda_1} - \frac{I_i - \bar{I}}{\lambda_1} = \frac{\lambda_1(x_{0i} - \bar{x}_0) + I_i - \bar{I}}{\lambda_1} - \frac{I_i - \bar{I}}{\lambda_1} = x_{0i} - \bar{x}_0.$$

□

Proof of Lemma 3.4:
First note that $z := z_i = z_j$ for all $1 \le i, j \le n$. By assumption

$$\begin{aligned} y_i - y_j &= \lambda_1(x_{0i} - \bar{x}_0) + I_i - \bar{I} - (\lambda_1(x_{0j} - \bar{x}_0) + I_j - \bar{I}) \\ &= \lambda_1(x_{0i} - x_{0j}) + I_i - I_j > 0 \end{aligned}$$

holds. Hence, $y_i/y_j > 1$. Using $\lambda_1 = w - d$ and $\lambda_2 = -(n-1)w - d$ with $w > d \ge 0$ we obtain

$$\begin{aligned} t_*^i - t_*^j &= \frac{1}{\lambda_2 - \lambda_1}\left(\ln\left(\frac{y_i}{z}\right) - \ln\left(\frac{y_j}{z}\right)\right) \\ &= -\frac{1}{nw}\ln\left(\frac{y_i}{y_j}\right) < 0. \end{aligned}$$

This proves the assertion. □

B Proofs of Section 4

Proof of Proposition 4.1:
By Lemma 3.3 the continuous system preserves the order of activations, i.e. $x_1(t) > \cdots > x_n(t)$ for $t \ge 0$. From Lemma 3.4 we know, that unit 1 arrives first at its minimum at time t_*^1. If we choose w sufficiently small both statements hold for the corresponding time-discrete system.

Now assume that unit 1 arrives at its minimum at time $t_* = t_*^1$. To show that the system converges we have to show that $\dot{\boldsymbol{x}}(t_*) = 0$. The proof is divided into three parts: Part 1 shows that $\dot{x}_1(t_*) = 0$. In the 2nd part we show that $\dot{x}_i(t_*) = 0$ for all units i with positive activation $x_i(t_*) \geq 0$. Finally, the third part asserts that $\dot{x}_j(t_*) = 0$ for all units j with negative activation $x_j(t_*) < 0$.

(1) *Proof:* $\dot{x}_1(t_*) = 0$

Follows from the fact that $x_1(t_*)$ is a minimum.

(2) *Proof:* $\dot{x}_i(t_*) = 0$ for all units i with $x_i(t_*) \geq 0$

From the first part of the proof we find that

$$\dot{x}_1(t_*) = -w \cdot \sum_{k=2}^{n} y_k(t_*) = 0$$

Thus we have $\sum_{k=2}^{n} y_k(t_*) = 0$. Now observe that

$$\dot{x}_i(t_*) = -w \cdot \sum_{\substack{k=2 \\ k \neq i}}^{n} y_k(t_*) = -w \cdot \underbrace{\sum_{k=2}^{n} y_k(t_*)}_{=0} + w \cdot y_i(t_*) = w \cdot y_i(t_*).$$

From $w > 0$ and $y_i(t_*) \geq 0$ directly follows that $\dot{x}_i(t_*) \geq 0$. But then we have $y_i(t_*) = 0$ and therefore $\dot{x}_i(t_*) = 0$.

(3) *Proof:* $\dot{x}_j(t_*) = 0$ for all units j with $x_j(t_*) < 0$

Let index i refer to units with $x_i(t_*) \geq 0$ and index k to units with $x_k(t_*) < 0$. We have

$$\dot{x}_j(t_*) = -w \cdot \underbrace{\sum_{i}^{n} y_i(t_*)}_{=0} - w \cdot \sum_{k \neq j} y_k(t_*) = -w \cdot \sum_{k \neq j} y_k(t_*) > 0.$$

With a similar argumentation as in the 2nd part of this proof we conclude $\dot{x}_j(t_*) = 0$.

Putting parts (1)-(3) together we have shown $\dot{\boldsymbol{x}}(t_*) = 0$. This proves convergence of the system. □