Fixed Points in Two–Neuron Discrete Time Recurrent Networks: Stability and Bifurcation Considerations

Peter Tiňo*

Department of Computer Science and Engineering

Slovak Technical University

Ilkovicova 3, 812 19 Bratislava, Slovakia

Email: `tino@decef.elf.stuba.sk`

Bill G. Horne and **C. Lee Giles**†

NEC Research Institute

4 Independence Way

Princeton, NJ 08540

Email: {horne,giles}@research.nj.nec.com

Technical Report

UMIACS-TR-95-51 and CS-TR-3461

Institute for Advanced Computer Studies

University of Maryland

College Park, MD 20742

*Currently with NEC Research Institute, 4 Independence Way, Princeton, NJ 08540, Email: `tino@research.nj.nec.com`

†Also with UMIACS, University of Maryland, College Park, MD 20742

ABSTRACT

The position, number and stability types of fixed points of a two–neuron recurrent network with nonzero weights are investigated. Using simple geometrical arguments in the space of derivatives of the sigmoid transfer function with respect to the weighted sum of neuron inputs, we partition the network state space into several regions corresponding to stability types of the fixed points. If the neurons have the same mutual interaction pattern, i.e. they either mutually inhibit or mutually excite themselves, a lower bound on the rate of convergence of the attractive fixed points towards the saturation values, as the absolute values of weights on the self–loops grow, is given. The role of weights in location of fixed points is explored through an intuitively appealing characterization of neurons according to their inhibition/excitation performance in the network. In particular, each neuron can be of one of the four types: greedy, enthusiastic, altruistic or depressed. Both with and without the external inhibition/excitation sources, we investigate the position and number of fixed points according to character of the neurons. When both neurons self-excite themselves and have the same mutual interaction pattern, the mechanism of creation of a new attractive fixed point is shown to be that of saddle node bifurcation.

1 Introduction

In this contribution we address the issues concerning fixed points of discrete–time recurrent neural networks consisting of two neurons. Nonzero weights are assumed. As pointed out in [3], because of the interest in associative memory applications, a great deal of previous work has focused on the question of how to constrain the weights of the recurrent networks so that they exhibit only fixed points (no oscillatory dynamics) [6]. In this context, it is desirable that all fixed points are attractive. Recently, Jin, Nikifiruk and Gupta [16] reported new results on the absolute stability for a rather general class of recurrent neural networks. Conditions under which all fixed points of the network are attractive were determined by the weight matrix of the network.

However, there are many applications where oscillatory dynamics of recurrent networks is desirable. For example, when trained to act as a finite state machine ([7], [9] [11], [12], [17], [19], [21], [22]), the network has to induce a stable representation of state transitions associated with each input symbol of the machine. A transition may have a character of a loop (do not move from the current state when the symbol x is presented), or a cycle (when repeatedly presenting the same input, we eventually return to the state where we have started). As reported in [5], [17], and [18], loops and cycles associated with an input symbol x are usually represented as attractive fixed points and periodic orbits respectively of the underlying dynamical system corresponding to the input x. In this respect, one can look at the training process from the point of view of bifurcation analysis. The network solves the task of finite state machine simulation by location of point and periodic attractors and shaping their respective basins of attraction [8]. Before training, the connection weights are set to small random values and as a consequence, the network has only one attractor basin. This implies that the network must undergo several bifurcations [10].

In [18], a preliminary analysis of the two–neuron recurrent network is given. Under some some specific conditions on weight values, the number, position and stability types of fixed points of the underlying dynamical systems are analyzed and bifurcation mechanism is clarified. The most typical bifurcation responsible for the creation of a new fixed point is the saddle node bifurcation.

Typically, studies of the asymptotic behaviour of recurrent neural networks usually assume some form of a structure in the weight matrix describing connectivity pattern among recurrent neurons. For example, symmetric connectivity and absence of self-interactions enabled Hopfield [14] to interpret the network as a physical system having energy minima in attractive fixed points of the network. These rather strict conditions were weakened in [6], where a more easily satisfied conditions are formulated. Blum and Wang [4] globally analyze networks with nonsymmetrical connectivity patterns of special types. In case of two recurrent neurons with sigmoidal activation function $g(\ell) = 1/(1 + e^{-\ell})$, they give results for weight matrices with diagonal elements equal to zero[1].

This paper presents a generalization of the results presented in [18]. A similar approach to determining the number and position of fixed points in continuous–time recurrent neural networks can be found in [3].

In section 3, the network state space is partitioned into several regions corresponding to stability types of the fixed points. This is done by first exploring the space of derivatives of the sigmoid transfer function with respect to the weighted sum of neuron inputs. Then, the structure is transformed into the space of neuron activations.

It was proved by Hirsh [13], that when all the weights in a recurrent network with exclusively

[1] In such a case the recurrent network is shown to have only one fixed point and no "genuine" periodic orbits (of period greater than one)

self-exciting (or exclusively self-inhibiting) neurons are multiplied by larger and larger positive number (neural gain), attractive fixed points tend to saturated activation values. Due to the analysis in section 3, in case of two–neuron network, under the assumption that the neurons have the same mutual interaction pattern[2], we give a lower bound on the rate of convergence of the attractive fixed points towards the saturation values as the absolute values of weights on the self–loops grow.

In section 4 the position and the number of fixed points is discussed. The role of weights in location of fixed points is investigated through an intuitively appealing characterization of neurons according to their inhibition/excitation performance in the network. For example, we view a neuron as a greedy one, if it self–excites itself, but inhibits the other neuron; an enthusiastic neuron excites both itself and the other neuron; etc...

In the context of greedy and enthusiastic neurons, the saddle node bifurcation, as a mechanism responsible for creation of a new attractive fixed point, is described in section 5.

Section 2 briefly introduces some necessary concepts from the theory of discrete time dynamical systems.

2 Dynamical systems

A discrete-time dynamical system can be represented as the iteration of a (differentiable) function $f: A \to A$ $(A \subseteq \Re^n)$, i.e.

$$x_{t+1} = f(x_t), \quad t \in \mathbf{N}, \tag{1}$$

where $\mathbf{N}$ denotes the set of all natural numbers. For each $x \in A$, the iteration (1) generates a sequence of distinct points defining the orbit, or trajectory of x under f. Hence, the orbit of x under f is the set $\{f^m(x) |\ m \geq 0\}$. For $m \geq 1$, f^m is the composition of f with itself m times. f^0 is defined to be the identity map on A.

A point $x_* \in A$ is called a *fixed point of* f, if $f^m(x_*) = x_*$, for all $m \in \mathbf{N}$.

Fixed points can be classified according to the behaviour of the orbits of points in their vicinity. A fixed point x_* is said to be asymptotically stable (or an *attractive point of* f), if there exists a neighborhood $O(x_*)$ of x_*, such that $\lim_{m\to\infty} f^m(x) = x_*$, for all $x \in O(x_*)$. As m increases, trajectories of points near to an asymptotically stable fixed point tend to it.

A fixed point x_* of f is asymptotically stable only if for each eigenvalue λ of $Df(x_*)$, the Jacobian of f at x_*, $|\lambda| < 1$ holds. The eigenvalues of $Df(x_*)$ govern whether or not the map f in a vicinity of x_* has contracting or expanding directions. Eigenvalues larger in absolute value than one lead to expansion, whereas eigenvalues smaller than one lead to contraction. If all the eigenvalues of $Df(x_*)$ are outside the unit circle, x_* is a *repulsive point*, or repellor. All points from a neighborhood of a repellor move away from it as m increases. If some eigenvalues of $Df(x_*)$ are inside and some are outside the unit circle, x_* is said to be a *saddle point*.

3 Qualitative analysis

The iterative map under consideration can be written as follows:

$$(x_{n+1}, y_{n+1}) = (g(ax_n + by_n + t_1), g(cx_n + dy_n + t_2)), \tag{2}$$

[2] they either mutually inhibit or mutually excite themselves

where $(x_n, y_n) \in (0,1)^2$ is the state of the network at the time step n, $a,b,c,d \in \Re \setminus \{0\}$ and $t_1, t_2 \in \Re$ are weights and bias terms respectively. g is a sigmoid function $g(\ell) = 1/(1+e^{-\ell})$. Since the neuron activations x_n and y_n are positive, signs of the weights determine the type of the corresponding connections: a *connection* is *exciting* and *inhibiting* if its weight is positive and negative respectively.

The aim of this section is to partition the state space $(0,1)^2$ of neurons' activations into several regions according to stability types of fixed points of (2).

Define the map $\phi : (0,1)^2 \to (0,1/4]^2$ as

$$\phi(x,y) = (x(1-x), y(1-y)). \tag{3}$$

Let $F(u,v)$ be a function $F : \Re^2 \to \Re^2$. The sets

$$\{(u,v)|F(u,v) < 0\}, \quad \{(u,v)|F(u,v) = 0\} \quad \text{and} \quad \{(u,v)|F(u,v) > 0\}$$

are denoted by F^-, F^0 and F^+ respectively.

Theorem 1: *If $bc > 0$, then all attractive fixed points (x,y) of (2) satisfy*

$$\phi(x,y) \in \left(0, \frac{1}{|a|}\right) \times \left(0, \frac{1}{|d|}\right).$$

Proof: Any fixed point (x,y) of (2) satisfies

$$(x,y) = (g(ax+by+t_1), g(cx+dy+t_2)). \tag{4}$$

Jacobian $J(x,y)$ of (2) in (x,y) is given by

$$\begin{pmatrix} aG_1(x,y) & bG_1(x,y) \\ cG_2(x,y) & dG_2(x,y) \end{pmatrix},$$

where $G_1(x,y) = g'(ax+by+t_1)$ and $G_2(x,y) = g'(cx+dy+t_2)$. Since $g'(p) = g(p)(1-g(p))$, considering (4) and (3) we have

$$(G_1(x,y), G_2(x,y)) = \phi(x,y). \tag{5}$$

The eigenvalues of J are[3]

$$\lambda_{1,2} = \frac{aG_1 + dG_2 \pm \sqrt{\mathcal{D}(G_1,G_2)}}{2},$$

where

$$\mathcal{D}(G_1,G_2) = (aG_1 - dG_2)^2 + 4G_1G_2bc. \tag{6}$$

Assume $a,d > 0$, i.e both neurons self-excite themselves. Then $\mathcal{D}^+, \alpha^+ \supseteq (0,\infty)^2$, where

$$\alpha(G_1,G_2) = aG_1 + dG_2. \tag{7}$$

Since G_1, G_2 can only be positive, it follows that to identify possible values of G_1 and G_2 so that $|\lambda_{1,2}| < 1$, it is sufficient to solve the inequality $aG_1 + dG_2 + \sqrt{\mathcal{D}(G_1,G_2)} < 2$, or equivalently

$$2 - aG_1 - dG_2 > \sqrt{\mathcal{D}(G_1,G_2)}. \tag{8}$$

[3] to simplify the notation, the identification (x,y) of a fixed point in which (2) is linearized is omitted

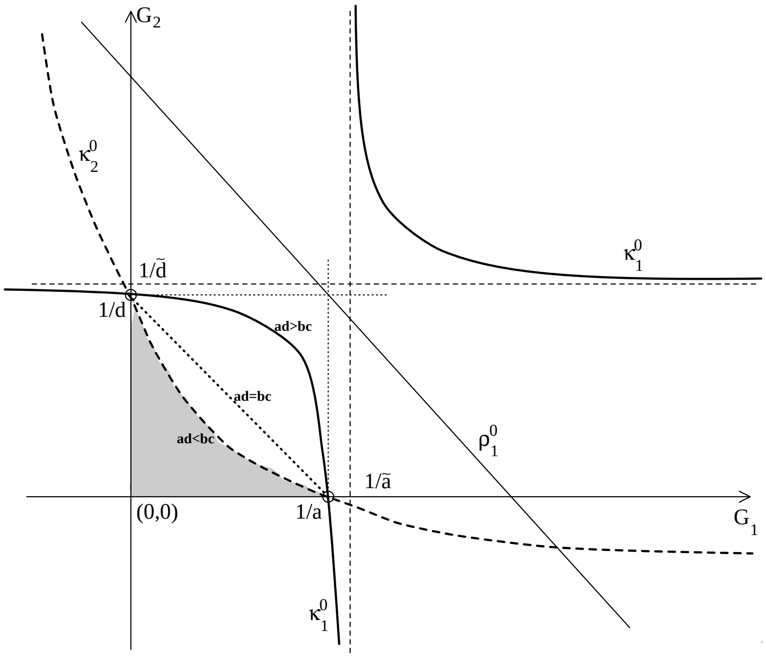

Figure 1: An illustration for the proof of Theorem 1. $a, d > 0,\ bc > 0$. All $(G_1, G_2) \in (0, 1/4]^2$ below the left branch (if $ad \geq bc$), or between the branches of κ_1^0 (if $ad < bc$) correspond to the attractive fixed points. Different line styles are associated with the cases $ad > bc$, $ad = bc$ and $ad < bc$, namely, the solid, dotted and dashed lines respectively.

Consider only (G_1, G_2) such that[4]

$$\rho_1(G_1, G_2) = aG_1 + dG_2 - 2 < 0. \tag{9}$$

All $(G_1, G_2) \in \rho_1^+$ (above ρ_1^0) lead to at least one eigenvalue of J greater than 1. Squaring both sides of (8) we arrive at

$$\kappa_1(G_1, G_2) = (ad - bc)G_1G_2 - aG_1 - dG_2 + 1 > 0. \tag{10}$$

If $ad \neq bc$, the "border" curve κ_1^0 is a hyperbola

$$G_2 = \frac{1}{\tilde{d}} + \frac{B}{G_1 - \frac{1}{\tilde{a}}}, \tag{11}$$

where

$$\tilde{a} = \frac{ad - bc}{d} = a - \frac{bc}{d}, \quad \tilde{d} = \frac{ad - bc}{a} = d - \frac{bc}{a}, \quad \text{and} \quad B = \frac{bc}{(ad - bc)^2}.$$

It is easy to check that $(1/a, 0), (0, 1/d) \in \kappa_1^0$.

If $ad = bc$, κ_1^0 is a line passing through the points $(0, 1/d)$ and $(1/a, 0)$ (see figure 1).

A fixed point (x, y) of (2) is attractive only if $(G_1, G_2) = \phi(x, y) \in \kappa_1^+ \cap \rho_1^-$, where the map ϕ is defined by (3). A necessary (not sufficient) condition for (x, y) to be attractive reads[5]

$$\phi(x, y) \in \left(0, \frac{1}{a}\right) \times \left(0, \frac{1}{d}\right).$$

[4] (G_1, G_2) lying under the line $\rho_1^0 : aG_1 + dG_2 = 2$.

[5] If $ad > bc$, then $0 < \tilde{a} < a$ and $0 < \tilde{d} < d$. $(G_1, G_2) \in \kappa_1^+$ lie under the "left branch" and above the "right branch" of κ_1^0. It is easy to see that since we are confined to ρ_1^- (below the line ρ_1^0), only (G_1, G_2) under the "left branch" of κ_1^0 will be considered. Indeed, ρ_1^0 is a decreasing line going through $(1/a, 1/d)$ and so it never intersects the right branch of κ_1^0. If $ad < bc$, then $\tilde{a}, \tilde{d} < 0$ and $(G_1, G_2) \in \kappa_1^+$ lie between the two branches of κ_1^0.

Consider now the case of self-inhibiting neurons, i.e. $a, d < 0$. Since $\alpha^- \supseteq (0, \infty)^2$, in order to identify possible values of (G_1, G_2) such that $|\lambda_{1,2}| < 1$, it is sufficient to solve the inequality $aG_1 + dG_2 - \sqrt{\mathcal{D}(G_1, G_2)} > -2$, or equivalently

$$2 + aG_1 + dG_2 > \sqrt{\mathcal{D}(G_1, G_2)}. \tag{12}$$

Analogically to the previous case, we shall consider only (G_1, G_2) such that[6]

$$\rho_2(G_1, G_2) = aG_1 + dG_2 + 2 > 0. \tag{13}$$

$(G_1, G_2) \in \rho_2^-$ (above ρ_2^0) lead to at least one eigenvalue of J greater than 1.

Squaring both sides of (12) we arrive at

$$\kappa_2(G_1, G_2) = (ad - bc)G_1G_2 + aG_1 + dG_2 + 1 > 0, \tag{14}$$

which is equivalent to

$$((-a)(-d) - bc)G_1G_2 - (-a)G_1 - (-d)G_2 + 1 > 0. \tag{15}$$

Further analysis equals the analysis from the previous case $(a, d > 0)$ with $a, \tilde{a}, d$ and $\tilde{d}$ replaced by $|a|, |a| - bc/|d|, |d|$ and $|d| - bc/|a|$ respectively.

If $ad \neq bc$, the "border" curve κ_2^0 is a hyperbola

$$G_2 = \frac{-1}{\tilde{d}} + \frac{B}{G_1 + \frac{1}{\tilde{a}}} \tag{16}$$

with $(-1/a, 0), (0, -1/d) \in \kappa_2^0$.

If $ad = bc$, κ_2^0 is a line passing through $(0, -1/d)$ and $(-1/a, 0)$.

A fixed point (x, y) of (2) is attractive only if $(G_1, G_2) = \phi(x, y) \in \kappa_2^+ \cap \rho_2^+$. (G_1, G_2) corresponding to attractive fixed points of (2) must lie in $(0, 1/|a|) \times (0, 1/|d|)$.

Finally, consider the case when one of the neurons has a self-excitation link, while the other self-inhibits itself. Without a loss of generality assume that $a > 0$ and $d < 0$. Assume further that[7] $(G_1, G_2) \in \alpha^+ \cup \alpha^0$. It is sufficient to solve the inequality (8). The relevant[8] (G_1, G_2) lie in $\kappa_1^+ \cap \rho_1^- \cap (\alpha^+ \cup \alpha^0)$ (figure 2). From $a > 0, d < 0$ it follows that $ad - bc$ is negative, and $0 < a < \tilde{a}$, $\tilde{d} < d < 0$.

For $(G_1, G_2) \in \alpha^-$ the relevant (G_1, G_2) lie in $\kappa_2^+ \cap \rho_2^+$.

It can be easily seen that κ_1^0 and κ_2^0 intersect on the line α^0 in $(1/\tilde{a}, 1/a) \times (1/|\tilde{d}|, 1/|d|)$ (see figure 2).

Added together, a fixed point (x, y) of (2) is attractive only if

$$\phi(x, y) \in [\kappa_1^+ \cap \rho_1^- \cap (\alpha^+ \cup \alpha^0)] \cup [\kappa_2^+ \cap \rho_2^+ \cap \alpha^-].$$

In particular, if (x, y) is attractive, then $\phi(x, y)$ must lie in $(0, 1/a) \times (0, 1/|d|)$. Examination of the case $a < 0, d > 0$ in the same way leads to a conclusion that all attractive fixed points of (2) have their corresponding (G_1, G_2) in $(0, 1/|a|) \times (0, 1/d)$. ∎

[6] (G_1, G_2) lying under the line $\rho_2^0 : -aG_1 - dG_2 = 2$.

[7] (G_1, G_2) such that $aG_1 + dG_2$ is nonnegative lie under or on the line $\alpha^0 : G_2 = aG_1/|d|$.

[8] (G_1, G_2) that correspond to fixed points in which both the eigenvalues of J have the absolute value less than one

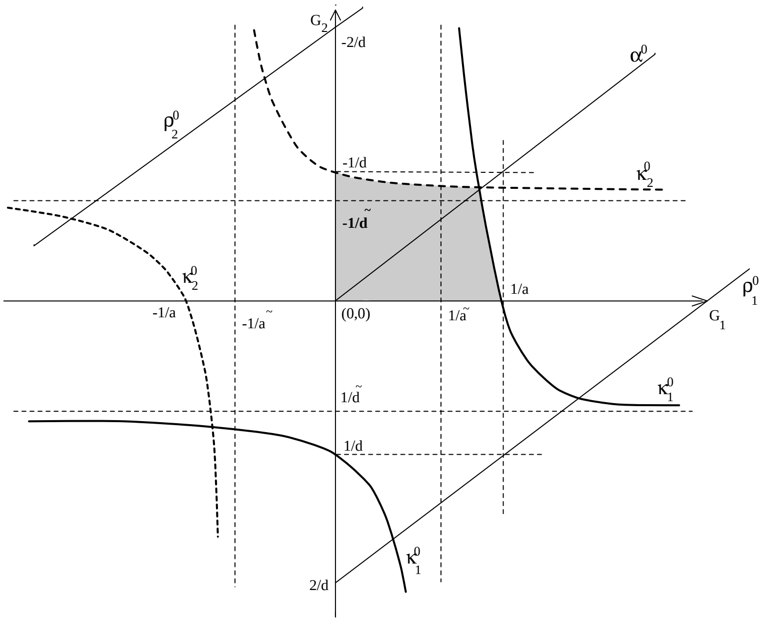

Figure 2: An illustration for the proof of Theorem 1. $a > 0, d < 0, bc > 0$. All $(G_1, G_2) \in (0, 1/4]^2$ below and on the line α^0, and between the two branches of κ_1^0 (solid line) correspond to the attractive fixed points. So do all $(G_1, G_2) \in (0, 1/4]^2$ above α^0, between the two branches of κ_2^0 (dashed line).

Theorem 2: *Assume $bc < 0$. Suppose $ad > 0$, or $ad < 0$ with $|ad| \leq |bc|/2$. Then each fixed point (x, y) of (2) such that*[9]

$$\phi(x,y) \in \left(0, \frac{1}{|a|}\right) \times \left(0, \frac{1}{|\tilde{d}|}\right) \cup \left(0, \frac{1}{|\tilde{a}|}\right) \times \left(0, \frac{1}{|d|}\right)$$

is attractive. In particular, all fixed points (x, y) for which

$$\phi(x,y) \in \left(0, \frac{1}{|\tilde{a}|}\right) \times \left(0, \frac{1}{|\tilde{d}|}\right)$$

are attractive

Proof: $\mathcal{D}(G_1, G_2)$ is no longer exclusively positive. It follows from analytic geometry (see for example [2]) that $\mathcal{D}(G_1, G_2) = 0$ defines either a single point or two lines (that can collide into one, or disappear). Since $(aG_1 - dG_2)^2 \geq 0$, $\mathcal{D}(G_1, G_2) = 0$ is satisfied only by those (G_1, G_2) for which $G_1 G_2 \geq 0$. Furthermore, $\mathcal{D}(0, 0) = 0$. Hence, $\mathcal{D}^0$ is either a single point – the origin, or a pair of increasing lines (that may be the same) passing through the origin.

Assume $a, d > 0$. Since $\mathcal{D}(1/a, 1/d) = 4bc/ad < 0$ and $\mathcal{D}(1/a, 0) = \mathcal{D}(0, 1/d) = 1 > 0$, the point $(1/a, 1/d)$ is always in[10] $\mathcal{D}^-$, while $(1/a, 0), (0, 1/d) \in \mathcal{D}^+$.

First, we shall examine the case when $\mathcal{D}(G_1, G_2)$ is negative. From

$$|\lambda_{1,2}|^2 = \frac{(aG_1 + dG_2)^2 + |\mathcal{D}|}{4} = G_1 G_2 (ad - bc)$$

[9] Recall that $\tilde{a}$ and $\tilde{d}$ denote $a - bc/d$ and $d - bc/a$ respectively.

[10] $\mathcal{D}^-$ is a nonempty region

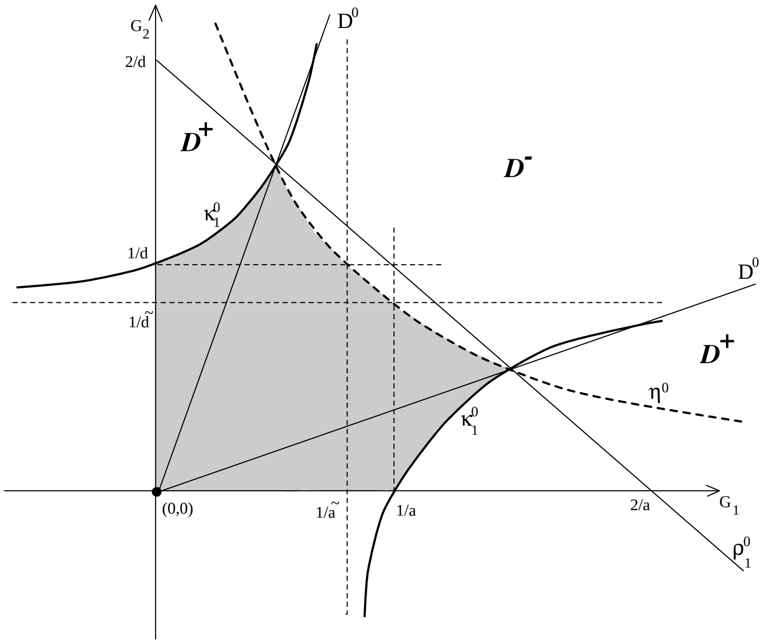

Figure 3: An illustration for the proof of Theorem 2. $a, d > 0$, $bc < 0$. All $(G_1, G_2) \in \mathcal{D}^-$ and below the right branch of η^0 (dashed line) correspond to the attractive fixed points. So do $(G_1, G_2) \in (0, 1/4]^2$ in $\mathcal{D}^+$ between the two branches of κ_1^0.

it follows, that $(G_1, G_2) \in \mathcal{D}^-$, for which $|\lambda_{1,2}| < 1$, lie in $\mathcal{D}^- \cap \eta^-$ (figure 3), where

$$\eta(G_1, G_2) = (ad - bc)G_1 G_2 - 1. \tag{17}$$

It is easy to show that

$$\left(\frac{1}{a}, \frac{1}{\tilde{d}}\right), \left(\frac{1}{\tilde{a}}, \frac{1}{d}\right) \in \eta^0, \quad \frac{1}{\tilde{a}} < \frac{1}{a}, \quad \frac{1}{\tilde{d}} < \frac{1}{d}, \quad \text{and} \quad \left(\frac{1}{a}, \frac{1}{\tilde{d}}\right), \left(\frac{1}{\tilde{a}}, \frac{1}{d}\right) \in \mathcal{D}^-.$$

Turn now to the case $\mathcal{D}(G_1, G_2) > 0$. Using the technique from the previous proof we conclude that the relevant (G_1, G_2) lie in[11] $\mathcal{D}^+ \cap \rho_1^- \cap \kappa_1^+$.

$\mathcal{D}^0, \rho_1^0, \kappa_1^0$ and η^0 intersect in two points as suggested by figure 3. To see this, note that for the points on η^0, it holds

$$G_1 G_2 = \frac{1}{ad - bc},$$

and for all $(G_1, G_2) \in \mathcal{D}^0 \cap \eta^0$ we have

$$(aG_1 + dG_2)^2 = 4. \tag{18}$$

For $G_1, G_2 > 0$, (18) defines the line ρ_1^0.

Similarly, for (G_1, G_2) lying on κ_1^0 and η^0, it holds

$$aG_1 + dG_2 = 2,$$

which is the definition of the line ρ_1^0. κ_1^0 and η^0 are monotonically increasing and decreasing respectively, and there is exactly one intersection point of the right branch of η^0 with each of the two branches of κ_1^0.

[11] under the line ρ_1^0 and between the two branches of hyperbola κ_1^0. Note that $0 < a < \tilde{a}$, $0 < d < \tilde{d}$.

For $(G_1, G_2) \in \mathcal{D}^0$,

$$|\lambda_{1,2}| = \frac{aG_1 + dG_2}{2},$$

and the relevant (G_1, G_2) are from $\mathcal{D}^0 \cap \rho_1^-$.

In summary, if $a, d > 0$, each fixed point (x, y) of (2) such that $\phi(x, y) = (G_1, G_2)$ is from

$$\phi(x, y) \in \left(0, \frac{1}{a}\right) \times \left(0, \frac{1}{\tilde{d}}\right) \cup \left(0, \frac{1}{\tilde{a}}\right) \times \left(0, \frac{1}{d}\right)$$

is attractive.

Assume $a, d < 0$. This case is identical to the case $a, d > 0$ examined above, with $a, \tilde{a}, d, \tilde{d}, \rho_1^-$ and κ_1^+ replaced by $|a|, |\tilde{a}|, |d|, |\tilde{d}|, \rho_2^+$ and κ_2^+ respectively.

First, note that $\mathcal{D}^0$ is the same as before, since

$$(aG_1 - dG_2)^2 = (|a|G_1 - |d|G_2)^2.$$

Furthermore, $ad - bc = |a||d| - bc$ and so $(G_1, G_2) \in \mathcal{D}^-$, for which $|\lambda_{1,2}| < 1$, lie in $\mathcal{D}^- \cap \eta^-$. Again, it directly follows that

$$\left(\frac{1}{|a|}, \frac{1}{|\tilde{d}|}\right), \left(\frac{1}{|\tilde{a}|}, \frac{1}{|d|}\right) \in \eta^0, \quad \frac{1}{|\tilde{a}|} < \frac{1}{|a|}, \quad \frac{1}{|\tilde{d}|} < \frac{1}{|d|}, \quad \text{and} \quad \left(\frac{1}{|a|}, \frac{1}{|\tilde{d}|}\right), \left(\frac{1}{|\tilde{a}|}, \frac{1}{|d|}\right) \in \mathcal{D}^-.$$

For $\mathcal{D}^+$ the relevant (G_1, G_2) lie in $\mathcal{D}^+ \cap \rho_2^+ \cap \kappa_2^+$.

All $(G_1, G_2) \in \mathcal{D}^0 \cap \rho_2^+$ lead to $|\lambda_{1,2}| < 1$. Hence, if $a, d < 0$, every fixed point (x, y) of (2) such that

$$\phi(x, y) \in \left(0, \frac{1}{|a|}\right) \times \left(0, \frac{1}{|\tilde{d}|}\right) \cup \left(0, \frac{1}{|\tilde{a}|}\right) \times \left(0, \frac{1}{|d|}\right)$$

is attractive.

Finally, consider the case $a > 0, d < 0$. The case $a < 0, d > 0$ would be treated in exactly the same manner.

Assume $\mathcal{D}^-$ is a nonempty region. Then, $ad > bc$ must hold and

$$\left(\frac{1}{a}, \frac{1}{|d|}\right) \in \mathcal{D}^-.$$

This can be easily seen, since for $ad < bc$ we would have

$$\mathcal{D}(G_1 G_2) = (aG_1 - dG_2)^2 + 4G_1G_2bc = (aG_1 + dG_2)^2 + 4G_1G_2(bc - ad) \geq 0$$

and $\mathcal{D}^-$ would not be a nonempty region. The sign of

$$\mathcal{D}\left(\frac{1}{a}, \frac{1}{|d|}\right) = 4\left(1 + \frac{bc}{a|d|}\right)$$

is determined by the sign of $a|d| + bc = bc - ad < 0$.

$(G_1, G_2) \in \mathcal{D}^-$, for which $|\lambda_{1,2}| < 1$, lie in $\mathcal{D}^- \cap \eta^-$ and

$$\left(\frac{1}{a}, \frac{1}{\tilde{d}}\right), \left(\frac{1}{|\tilde{a}|}, \frac{1}{|d|}\right) \in \eta^0.$$

Note that $\tilde{d} \geq |d|$ and $|\tilde{a}| \geq a$ only if $2a|d| \leq |bc|$.

Only those $(G_1, G_2) \in \mathcal{D}^0$ are taken into account for which $|aG_1 + dG_2| < 2$. This is true for all[12] $(G_1, G_2) \in \mathcal{D}^0 \cap \rho_1^- \cap \rho_2^+$ (figure 4).

[12] (G_1, G_2) between the lines ρ_1^0 and ρ_2^0.

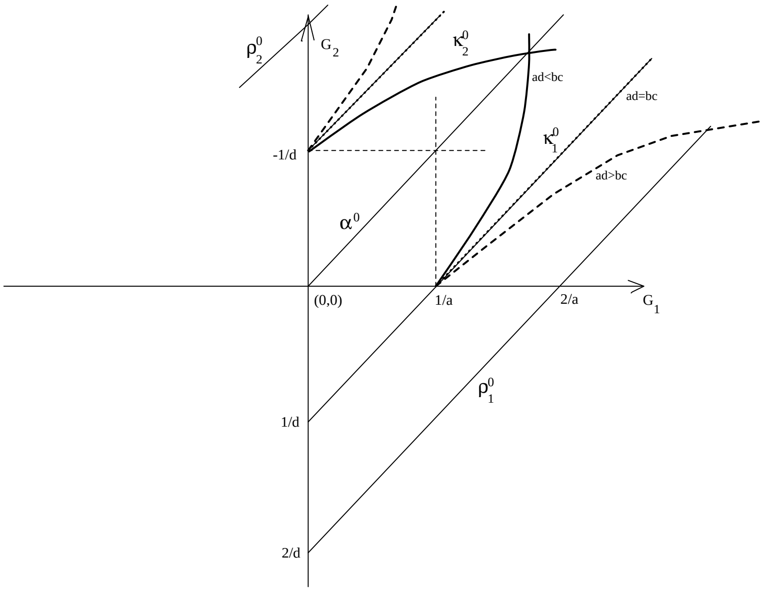

Figure 4: An illustration for the proof of Theorem 2, when $\mathcal{D} > 0$. $a > 0, d < 0,\ bc < 0$. If $ad - bc < 0$, the branches of κ_1^0 and κ_2^0 intersect on the line α^0. As $ad - bc$ grows, the meeting point moves up on the line α^0. When $ad = bc$, the branches deform into the lines and as $ad - bc > 0$ grows further, the two branches move towards the axis $G_1 = 0, G_2 = 0$.

If $\mathcal{D}(G_1, G_2) > 0$, the inequalities to be solved depend on the sign of $aG_1 + dG_2$. Following the same reasoning as in the proof of Theorem 1, we conclude that the relevant (G_1, G_2) lie in

$$[\kappa_1^+ \cap \rho_1^- \cap (\alpha^+ \cup \alpha^0)] \cup [\kappa_2^+ \cap \rho_2^+ \cap \alpha^-].$$

When $ad - bc < 0$, the branches of κ_1^0 and κ_2^0 intersect on the line α^0 in $(1/a, \infty) \times (1/|d|, \infty)$. As $ad - bc$ grows, the meeting point moves up on the line α^0. When $ad = bc$, the branches deform into the lines and as $ad - bc > 0$ grows further, the two branches move towards the axis $G_1 = 0, G_2 = 0$ (figure 4). ∎

In the proof of Theorem 1, we have seen that if $a > 0, d < 0$ and $bc > 0$, then all $(G_1, G_2) \in (0, 1/\tilde{a}) \times (0, 1/|\tilde{d}|)$ potentially correspond to attractive fixed points of (2) (figure 2). In the proof of the last Theorem it was shown that when $a > 0, d < 0, bc < 0$, if $2a|d| \geq |bc|$, then $(1/a, 1/|d|)$ is on or under the right branch of η^0 and each $(G_1, G_2) \in (0, 1/a) \times (0, 1/|d|)$ potentially corresponds to an attractive fixed point of (2). Hence, the following Theorem can be formulated:

Theorem 3: *If $ad < 0$ and*

- *$bc > 0$, then every fixed point (x, y) of (2) such that*

$$\phi(x, y) \in \left(0, \frac{1}{|\tilde{a}|}\right) \times \left(0, \frac{1}{|\tilde{d}|}\right)$$

  is attractive.

- *$bc < 0$ with $|ad| \geq |bc|/2$, then each fixed point (x, y) of (2) satisfying*

$$\phi(x, y) \in \left(0, \frac{1}{|a|}\right) \times \left(0, \frac{1}{|d|}\right)$$

is attractive.

Now, we transform our results into the (x, y)–space of neuron activations. For $u > 4$, define

$$\Delta(u) = \frac{1}{2}\sqrt{1 - \frac{4}{u}}.$$

In Theorems 1, 2 and 3 a structure reflecting stability types of the fixed points of (2) was introduced into the (G_1, G_2)–space. The region $(0, 1/4]^2$ in (G_1, G_2)–plane corresponds to four regions

$$\left(0, \frac{1}{2}\right]^2, \quad \left(0, \frac{1}{2}\right] \times \left[\frac{1}{2}, 1\right), \quad \left[\frac{1}{2}, 1\right) \times \left(0, \frac{1}{2}\right], \quad \text{and} \quad \left[\frac{1}{2}, 1\right)^2.$$

in the (x, y)–space. In particular, for each $(G_1, G_2) \in (0, 1/4]^2$, under the map ϕ, there are four preimages

$$(x, y) = \phi^{-1}(G_1, G_2) = \left\{ \left(\frac{1}{2} \pm \Delta\left(\frac{1}{G_1}\right), \frac{1}{2} \pm \Delta\left(\frac{1}{G_2}\right) \right) \right\}. \tag{19}$$

Results formulated in Theorems 1, 2 and 3 can now be stated for the space of activations of recurrent neurons.

For $\alpha > 4, \delta > 4$, the regions of the (x, y)–space

$$\left(0, \frac{1}{2} - \Delta(\alpha)\right) \times \left(0, \frac{1}{2} - \Delta(\delta)\right),$$

$$\left(\frac{1}{2} - \Delta(\alpha), \frac{1}{2}\right] \times \left(0, \frac{1}{2} - \Delta(\delta)\right) \cup \left(0, \frac{1}{2} - \Delta(\alpha)\right) \times \left(\frac{1}{2} - \Delta(\delta), \frac{1}{2}\right]$$

and

$$\left(\frac{1}{2} - \Delta(\alpha), \frac{1}{2}\right] \times \left(\frac{1}{2} - \Delta(\delta), \frac{1}{2}\right]$$

are denoted by $R^A_{00}(\alpha, \delta)$, $R^S_{00}(\alpha, \delta)$ and $R^R_{00}(\alpha, \delta)$ respectively. Regions symmetrical to $R^A_{00}(\alpha, \delta)$, $R^S_{00}(\alpha, \delta)$ and $R^R_{00}(\alpha, \delta)$ with respect to the line $x = 1/2$ are denoted by $R^A_{10}(\alpha, \delta)$, $R^S_{10}(\alpha, \delta)$ and $R^R_{10}(\alpha, \delta)$ respectively. Similarly, let $R^A_{01}(\alpha, \delta)$, $R^S_{01}(\alpha, \delta)$ and $R^R_{01}(\alpha, \delta)$ denote the regions symmetrical to $R^A_{00}(\alpha, \delta)$, $R^S_{00}(\alpha, \delta)$ and $R^R_{00}(\alpha, \delta)$ with respect to the line $y = 1/2$. Finally, $R^A_{11}(\alpha, \delta)$, $R^S_{11}(\alpha, \delta)$ and $R^R_{11}(\alpha, \delta)$ denote regions that are symmetrical to $R^A_{01}(\alpha, \delta)$, $R^S_{01}(\alpha, \delta)$ and $R^R_{01}(\alpha, \delta)$ with respect to the line $x = 1/2$ (figure 5).

Corollary 1: *If* $bc > 0, |a| > 4, |d| > 4$, *then all attractive fixed points of (2) lie in* $\bigcup_{i \in \mathcal{I}} R^A_i(|a|, |d|)$, *where* $\mathcal{I}$ *is the index set* $\mathcal{I} = \{00, 10, 01, 11\}$.

Corollary 2: *If* $bc < 0$, $ad < 0$, $|a| > 4$, $|d| > 4$ *and* $|ad| \geq |bc|/2$, *then all fixed points of (2) lying in* $\bigcup_{i \in \mathcal{I}} R^A_i(|a|, |d|)$, $\mathcal{I} = \{00, 10, 01, 11\}$ *are attractive.*

Corollary 3: *If* $|\tilde{a}|, |\tilde{d}| > 4$ *and one of the following conditions is satisfied*

- $bc > 0$ *and* $ad < 0$
- $bc < 0$ *and* $ad > 0$

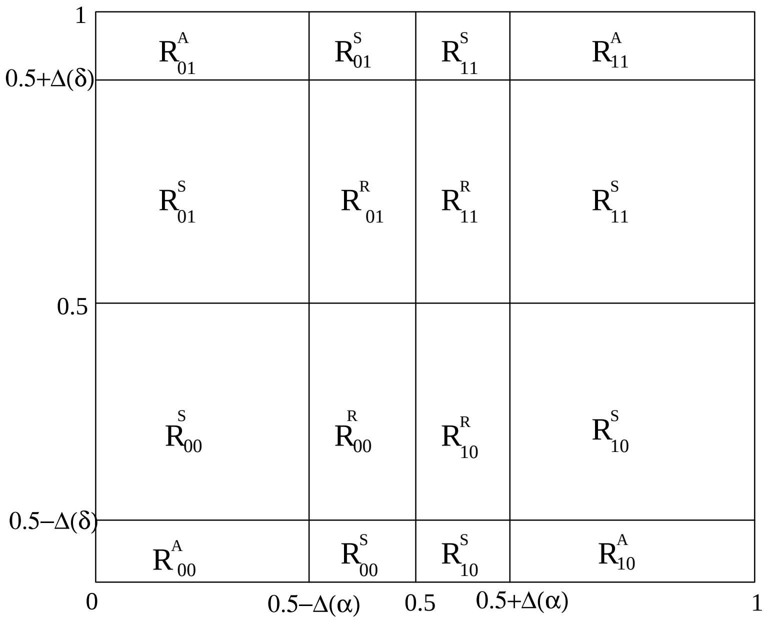

Figure 5: Partitioning of the network state space according to stability types of the fixed points.

- $bc < 0, ad < 0$ *and* $|ad| \leq |bc|/2$

then all fixed points of (2) lying in $\bigcup_{i\in\mathcal{I}} R_i^A(|\tilde{a}|, |\tilde{d}|)$, $\mathcal{I} = \{00, 10, 01, 11\}$ *are attractive.*

For an insight into a bifurcation mechanism (explored in section 5) by which attractive fixed points of (2) are created (or dismissed), it is useful to have an idea where other types of fixed points can lie. For the case when both neurons are either self-exciting, or self-inhibiting ($ad > 0$), and their mutual interaction is of the same character $(bc > 0)$, we have the following theorem:

Theorem 4: *Suppose* $ad > 0, bc > 0, |a| > 4, |d| > 4$. *Then the following can be said about the fixed points of (2):*

- *attractive points can lie only in* $\bigcup_{i\in\mathcal{I}} R_i^A(|a|, |d|)$, $\mathcal{I} = \{00, 10, 01, 11\}$.
- *if* $ad \geq bc/2$, *then all fixed points in* $\bigcup_{i\in\mathcal{I}} R_i^S(|a|, |d|)$ *are saddle points; repulsive points can lie only in* $\bigcup_{i\in\mathcal{I}} R_i^R(|a|, |d|)$.
- *if* $|ad - bc| < 4min\{|a|, |d|\}$, *then there are no repellors.*

Proof: Regions for attractive fixed points result from Corollary 1.

Consider first the case $a, d > 0$. A fixed point (x, y) of (2) is a saddle if $|\lambda_2| < 1$ and $|\lambda_1| = \lambda_1 > 1$.

Assume $ad > bc$. Then

$$0 < \sqrt{(aG_1 + dG_2)^2 - 4G_1G_2(ad - bc)} = \sqrt{\mathcal{D}(G_1, G_2)} < aG_1 + dG_2.$$

It follows that if $aG_1+dG_2<2$, i.e. $(G_1,G_2)\in\rho_1^-$, $0<aG_1+dG_2-\sqrt{\mathcal{D}(G_1,G_2)}<2$ holds and $0<\lambda_2<1$.

For $(G_1,G_2)\in\rho_1^0\cup\rho_1^+$, we solve the inequality $aG_1+dG_2-\sqrt{\mathcal{D}(G_1,G_2)}<2$, that is satisfied by (G_1,G_2) from $\kappa_1^-\cap(\rho_1^0\cup\rho_1^+)$.

It can be seen (figure 1) that in all fixed points (x,y) of (2) with

$$\phi(x,y)\in\left(0,\frac{1}{4}\right]\times\left(0,min\left\{\frac{1}{\tilde{d}},\frac{1}{4}\right\}\right]\cup\left(0,min\left\{\frac{1}{\tilde{a}},\frac{1}{4}\right\}\right]\times\left(0,\frac{1}{4}\right],$$

the eigenvalue $\lambda_2>0$ is less than 1. This is certainly true for all (x,y) such that $\phi(x,y)\in(0,1/4]\times(0,1/d)\cup(0,1/a)\times(0,1/4]$. In particular, the preimages of $(G_1,G_2)\in(1/a,1/4]\times(0,1/d)\cup(0,1/a)\times(1/d,1/4]$ under ϕ define the region $\bigcup_{i\in\mathcal{I}}R_i^S(a,d)$ where only saddle fixed points of (2) can lie.

Fixed points (x,y) whose images under ϕ lie in $\kappa_1^+\cap\rho_1^+$ are repellors. No (G_1,G_2) can lie in that region, if $\tilde{a},\tilde{d}\le 4$, that is, if $d(a-4)\le bc$ and $a(d-4)\le bc$, which is equivalent to $max\{a(d-4),d(a-4)\}\le bc$.

In the case $ad=bc$, we have $\sqrt{\mathcal{D}(G_1,G_2)}=aG_1+dG_2$ and so $\lambda_2=0$. Hence, there are no repelling points if $ad=bc$.

Assume $ad<bc$. Then $\sqrt{\mathcal{D}(G_1,G_2)}>aG_1+dG_2$, which implies that λ_2 is negative. It follows that the inequality to be solved is $aG_1+dG_2-\sqrt{\mathcal{D}(G_1,G_2)}>-2$. It is satisfied by (G_1,G_2) from κ_2^+. If $2ad\ge bc$, for the coefficients of κ_2^0 we have $|\tilde{a}|\le a$ and $|\tilde{d}|\le d$.

Fixed points (x,y) with

$$\phi(x,y)\in\left(0,\frac{1}{4}\right]\times\left(0,min\left\{\frac{1}{|\tilde{d}|},\frac{1}{4}\right\}\right]\cup\left(0,min\left\{\frac{1}{|\tilde{a}|},\frac{1}{4}\right\}\right]\times\left(0,\frac{1}{4}\right],$$

have $|\lambda_2|$ less than 1. If $2ad\ge bc$, this is true for all (x,y) such that $\phi(x,y)\in(0,1/4]\times(0,1/d)\cup(0,1/a)\times(0,1/4]$ and the preimages of $(G_1,G_2)\in(1/a,1/4]\times(0,1/d)\cup(0,1/a)\times(1/d,1/4]$ under ϕ define the region $\bigcup_{i\in\mathcal{I}}R_i^S(a,d)$ where only saddle fixed points of (2) can lie.

There are no repellors if $|\tilde{a}|,|\tilde{d}|\le 4$, that is, if $min\{a(d+4),d(a+4)\}\ge bc$.

If we examined the case $a,d<0$ in the same spirit as the case $a,d>0$ we would conclude that

- if $ad>bc$, in all fixed points (x,y) of (2) with

$$\phi(x,y)\in\left(0,\frac{1}{4}\right]\times\left(0,min\left\{\frac{1}{|\tilde{d}|},\frac{1}{4}\right\}\right]\cup\left(0,min\left\{\frac{1}{|\tilde{a}|},\frac{1}{4}\right\}\right]\times\left(0,\frac{1}{4}\right],$$

  $|\lambda_1|<1$. Surely, this is true for all (x,y) such that $\phi(x,y)\in(0,1/4]\times(0,1/|d|)\cup(0,1/|a|)\times(0,1/4]$. The preimages of $(G_1,G_2)\in(1/|a|,1/4]\times(0,1/|d|)\cup(0,1/|a|)\times(1/|d|,1/4]$ under ϕ define the region $\bigcup_{i\in\mathcal{I}}R_i^S(|a|,|d|)$ where only saddle fixed points of (2) can lie. There are no repellors if $|\tilde{a}|,|\tilde{d}|\le 4$, that is, if $|d|(|a|-4)\le bc$ and $|a|(|d|-4)\le bc$, which is equivalent to $max\{|a|(|d|-4),|d|(|a|-4)\}\le bc$.

- in the case $ad=bc$, we have $\sqrt{\mathcal{D}(G_1,G_2)}=|aG_1+dG_2|$ and so $\lambda_1=0$. Hence, there are no repelling points.

- if $ad<bc$, in all fixed points (x,y) with

$$\phi(x,y)\in\left(0,\frac{1}{4}\right]\times\left(0,min\left\{\frac{1}{\tilde{d}},\frac{1}{4}\right\}\right]\cup\left(0,min\left\{\frac{1}{\tilde{a}},\frac{1}{4}\right\}\right]\times\left(0,\frac{1}{4}\right],$$

λ_1[13] is less than 1. If $2ad \geq bc$, this is true for all (x,y) such that $\phi(x,y) \in (0,1/4] \times (0,1/|d|) \cup (0,1/|a|) \times (0,1/4]$ and the preimages of $(G_1,G_2) \in (1/|a|,1/4] \times (0,1/|d|) \cup (0,1/|a|) \times (1/|d|,1/4]$ under ϕ define the region $\bigcup_{i\in\mathcal{I}} R_i^S(|a|,|d|)$ where only saddle fixed points of (2) can lie. There are no repellors if $\tilde{a},\tilde{d} \leq 4$, that is, if $min\{|a|(|d|+4), |d|(|a|+4)\} \geq bc$.

In general, we have shown that if

- $ad < bc$ and $ad + 4min\{|a|,|d|\} \geq bc$, or
- $ad = bc$, or
- $ad > bc$ and $ad - 4min\{|a|,|d|\} \leq bc$,

then there are no repellors. ∎

4 Quantitative analysis

In this section we are concerned with the actual position of fixed points of (2). We study, how the coefficients a, b, t_1, c, d and t_2 effect the position and the number of the fixed points. It is illustrative first to concentrate on a single neuron from a pair of neurons.

Denote the values of the weights associated with the self–loop of the selected neuron and with the interconnection link from the other neuron to the selected neuron by s and r respectively. The constant input to the selected neuron is denoted by t. If the activations of the selected neuron and the other neuron are u and v respectively, then the activation of the selected neuron at the next time step is[14] $g(su+rv+t)$. If the activation of the selected neuron is not to change, (u,v) should lie on the curve $f_{s,r,t}$:

$$v = f_{s,r,t}(u) = \frac{1}{r}\left(-t - su + ln\frac{u}{1-u}\right). \tag{20}$$

$ln(u/(1-u))\colon (0,1) \to \Re$, is a monotonically increasing function with

$$\lim_{u\to 0^+} ln\frac{u}{1-u} = -\infty \quad \text{and} \quad \lim_{u\to 1^-} ln\frac{u}{1-u} = \infty.$$

The linear function $-su+t$ cannot influence these assymptotical properties, it can, however, locally influence the "shape" of $f_{s,r,t}$. In particular while the effect of the constant term $-t$ is just a vertical shift of the whole function, $-su$ (if decreasing, i.e. if $s>0$, and "sufficiently large") has the power to overcome for a while the increasing tendencies of $ln(u/(1-u))$. More precisely, if $s>4$ then the term $-su$ causes the function $-su-t+ln(u/(1-u))$ to "bend" so that on

$$\left[\frac{1}{2} - \Delta(s), \frac{1}{2} + \Delta(s)\right]$$

it is decreasing, while it still increases on

$$\left(0, \frac{1}{2} - \Delta(s)\right) \cup \left(\frac{1}{2} + \Delta(s), 1\right).$$

$-su-t+ln(u/(1-u))$ is always concave and convex on $(0,1/2)$ and $(1/2,1)$ respectively. Finally, the coefficient r scales the whole function and flips it around the u–axis, if $r<0$. A graph of $f_{s,r,t}(u)$ is presented in figure 6.

[13] λ_1 is positive

[14] recall, that g the sigmoid function $g(\ell) = 1/(1+e^{-\ell})$

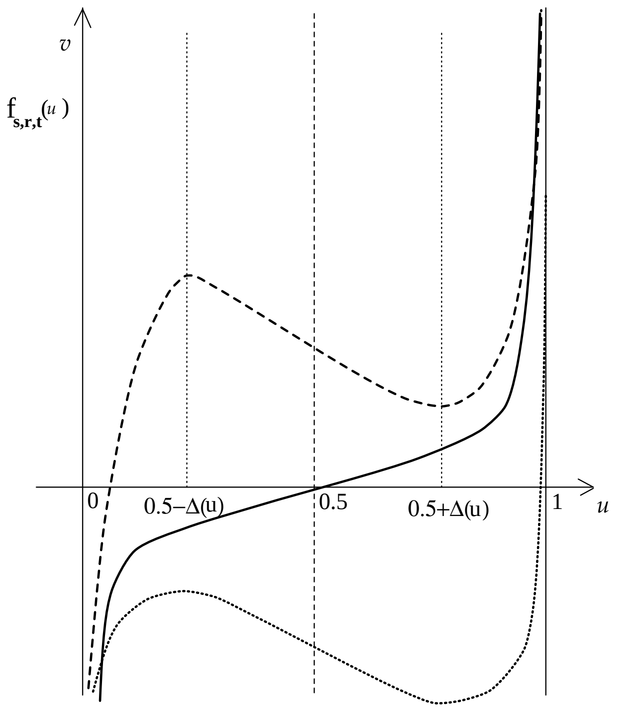

Figure 6: Graph of $f_{s,r,t}(u)$. Solid and dotted lines represent the cases $t, s = 0, r > 0$ and $t = 0, s > 4, r > 0$. Dashed line shows the graph when $t < 0, s > 4$ and $r > 0$. Negative external input t shifts the bended part into $v > 0$.

We characterize the neurons according to the sign of weights of the links stemming out of them. A neuron is said to be *greedy* if it self–excites itself, but inhibits the other neuron (the weight of the link to the other neuron is negative). A neuron is said to be *altruistic* if the opposite is true, i.e. if it self–inhibits itself, but excites the other neuron. An *enthusiastic* neuron excites both itself and the other neuron, while a *depressed* neuron inhibits everything including itself.

There are $\binom{4}{2} + 4 = 10$ possible cases of the coexistence of the two neurons. A fixed point represents a "compromise" achieved by both neurons in that the state of the system, once in a fixed point, does not change. Of course, just as with fixed points, the compromise can be characterized by various forms of stability. Based on the results from the previous section, in some cases we are able to predict the stability type of the fixed points of (2) according to their position in the neurons' activation space.

Each fixed point of (2) lies on the intersection of two curves $y = f_{a,b,t_1}(x)$, $x = f_{d,c,t_2}(y)$. We present some illustrative examples of the analysis of the position and the number of fixed points of (2) based on the characterization of neurons proposed above. Other cases would be analyzed in a similar manner. The external inputs are treated as artificial means to externally control the state of the system and the discussion of each case starts with an assumption that $t_1, t_2 = 0$. Signs of the coefficients a, b, c, d are marked by $+$ (if positive) and $-$ (if negative).

both neurons are enthusiastic: $(a, b, c, d) = (+, +, +, +)$ (figure 7)
$\circ$ $(t_1, t_2 = 0)$ Since $a, d > 0$, both functions f_{a,b,t_1} and f_{d,c,t_2} can "bend", but they bend before running into positive values (they bend outside $(0, 1)^2$). Since f_{a,b,t_1} and f_{d,c,t_2} pass only through $(1/2, 1) \times (0, 1)$ and $(0, 1) \times (1/2, 1)$ respectively, a fixed point only occurs in $(1/2, 1)^2$, the region

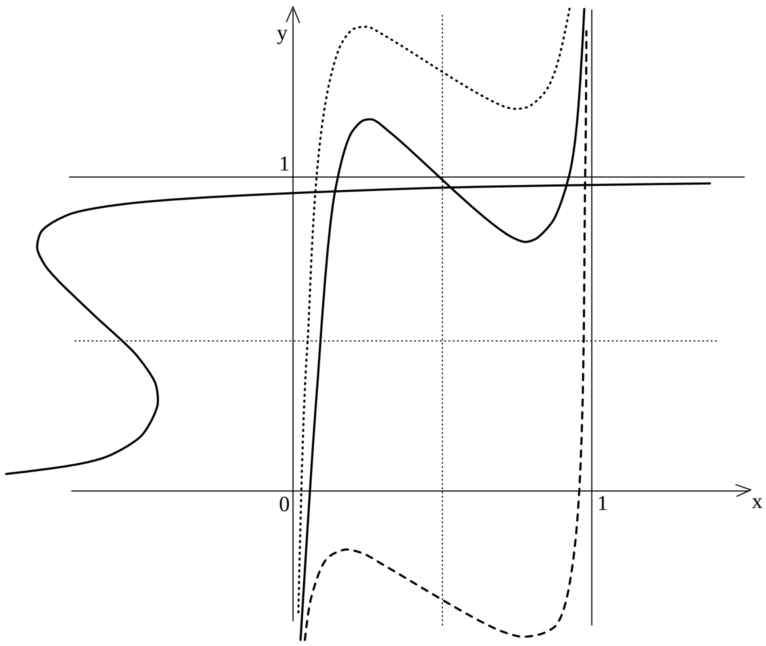

Figure 7: f_{a,b,t_1} and f_{d,c,t_2} when both neurons are enthusiastic.

of high activity of both neurons (f_{a,b,t_1} shown as a dashed line). There is no way to create a fixed point in a region, say, $(0,1/2)\times(1/2,1)$ corresponding to a state where the second neuron dominates over the first neuron.
• $(t_1, t_2 \neq 0)$ The only way to achieve the situation described above would be to use a large negative external input t_1 to the first neuron that would move the graph of f_{a,b,t_1} up, so that it intersects f_{d,c,t_2} in $(0,1/2)\times(1/2,1)$. If we artificially inhibited the first neuron by the external input too much, there may no longer be a fixed point in $(1/2,1)^2$ (f_{a,b,t_1} shown as a dotted line). However, if the self-excitation loop of the first neuron is strong enough, the bended shape of f_{a,b,t_1} can retain a fixed point in $(1/2,1)^2$ in spite of the external inhibition of the first neuron (f_{a,b,t_1} shown as a solid line).

both neurons are depressed: $(a,b,c,d) = (-,-,-,-)$
∘ $(t_1, t_2 = 0)$ Since $a, d < 0$, neither of the functions f_{a,b,t_1} and f_{d,c,t_2} can "bend". f_{a,b,t_1} and f_{d,c,t_2} pass through $(0,1/2)\times(0,1)$ and $(0,1)\times(0,1/2)$ respectively. A fixed point occurs in $(0,1/2)^2$, the region of low activity of both neurons.
• $(t_1, t_2 \neq 0)$ Positive external input to a neuron can, however, shift the fixed point towards high activity of that neuron.

an enthusiastic and a greedy neurons: $(a,b,c,d) = (+,-,+,+)$ (figure 8)
∘ $(t_1, t_2 = 0)$ f_{d,c,t_2} passes only through $(0,1)\times(1/2,1)$. The first (enthusiastic) neuron pays for being generous (it excites the second, greedy neuron) and there is no possibility of creating a fixed point in $(1/2,1)\times(0,1/2)$, the region of dominance of the enthusiastic neuron. Besides the possibility that there is a fixed point in $(0,1)\times(1/2,1)$ which may be close to either of the vertices[15] $(0,1)$ or $(1,1)$, there is a chance of having fixed points near both vertices (in fact there will be another fixed point "between" them) at one time. This can be achieved by a "cooperation"

[15] depending on how strong is the self–loop of the first neuron; f_{a,b,t_1} is shown as a dashed and a dotted lines

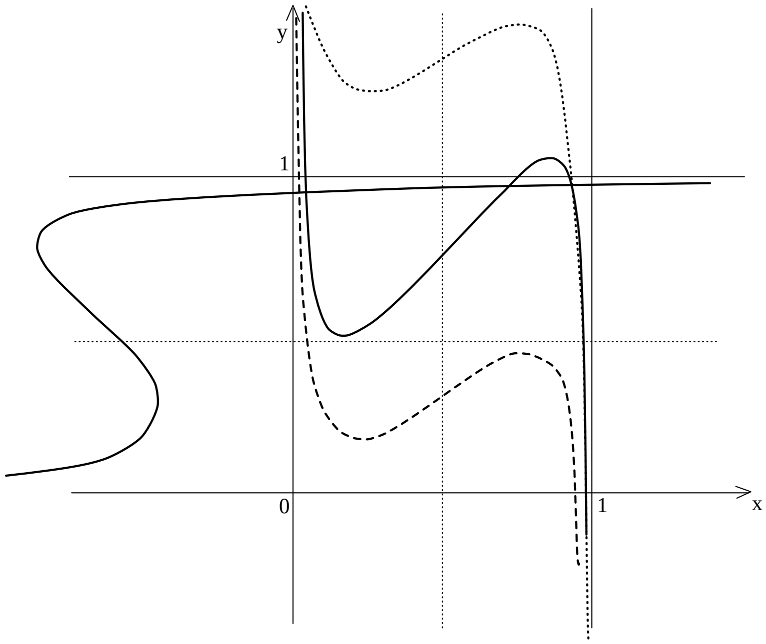

Figure 8: f_{a,b,t_1} and f_{d,c,t_2} in the case of an enthusiastic and a greedy neurons.

between the two neurons, in that the self–loop of the first neuron and the inhibition link from the second neuron have to have the "right" weights (they are neither too weak, nor too strong), so that the bended function f_{a,b,t_1} intersects f_{d,c,t_2} near both of the vertices (f_{a,b,t_1} shown as a solid line).

• ($t_1, t_2 \neq 0$) An interesting situation arises when the greedy neuron is externally inhibited, but has a strong self–loop, so that the bended part of f_{d,c,t_2} gets into $(0,1)^2$. Nine fixed points can be created. In general, a necessary condition on weights so that nine fixed points can exist is that the weights a, d on the self–loops are positive. This enables both functions f_{a,b,t_1} and f_{d,c,t_2} to "bend", and by moving the bended parts into $(0,1)^2$, create a complex intersection pattern.

a greedy and an altruistic neurons: $(a, b, c, d) = (+, +, -, -)$

∘ ($t_1, t_2 = 0$) Only a single fixed point in $(1/2, 1) \times (0, 1/2)$ can exist. Everything is in control of the greedy neuron.

• ($t_1, t_2 \neq 0$) By externally inhibiting the greedy neuron (moving the bended part of f_{a,b,t_1} upwards into $(0,1)^2$) more fixed points can be created. A strong external excitation of the altruistic neuron moves fixed points into $(0,1) \times (1/2, 1)$. There is even a possibility of creating a single fixed point in the region of dominance of the altruistic neuron ($(0,1/2) \times (1/2,1)$), if the greedy and altruistic neurons are strongly externally inhibited and excited respectively, but then the system is totally controlled by the external forces.

both neurons are greedy: $(a, b, c, d) = (+, -, -, +)$

∘ ($t_1, t_2 = 0$) Generally, if there are no external inputs, the case of two greedy neurons is the only case when there can be fixed points in the regions $(0,1/2) \times (1/2,1)$, $(1/2,1) \times (0,1/2)$ and $(1/2,1) \times (1/2,1)$ at one time. Even though the neurons inhibit each other, they can increase their self-excitation and through bended functions f_{a,b,t_1} and f_{d,c,t_2} introduce fixed points near the vertices $(1,0)$ and $(0,1)$ representing "winning" states of the first and second neuron respectively.

If they, moreover, "decide to cooperate" by not self-exciting themselves and inhibiting each other too much, a third fixed point near the high-activation vertex $(1,1)$ can be created. Hence, to create the most complex intersection pattern of f_{a,b,t_1} and f_{d,c,t_2} without external inputs, the two neurons should be "reasonably" greedy.

5 Creation of a new attractive fixed point through saddle node bifurcation

In this section we bring together the results from the last two sections. Normally, to detect stability types of the fixed points of (2), we would compute the position of the fixed points (which cannot, in general, be done analytically) and then linearize the system (2) in those fixed points, or directly use results of the section 3, where we have structured the network state space $(0,1)^2$ into areas where fixed points of particular stability types can lie. Fortunately, in some cases, these areas correspond to monotonicity intervals of the functions f_{a,b,t_1} and f_{d,c,t_2} defining the fixed points. The reasoning about the stability type of the fixed points can be based on the knowledge of where the functions intersect.

In this respect, the results of the section 3 will be useful when the neurons are enthusiastic or greedy, with a strong tendency to self excite themselves so that the functions f_{a,b,t_1} and f_{d,c,t_2} "bend", thus creating a possibility of complex intersection pattern in $(0,1)^2$.

For $a > 4$, denote the set

$$\left\{ (x, f_{a,b,t_1}(x)) |\ x \in \left(0, \frac{1}{2} - \Delta(a)\right) \right\}$$

of points lying on the "first outer branch" of $f_{a,b,t_1}(x)$ by $f^{\#0}_{a,b,t_1}$. Analogically, the set of points

$$\left\{ (x, f_{a,b,t_1}(x)) |\ x \in \left(0, \frac{1}{2} + \Delta(a)\right) \right\}$$

in the "second outer branch" of $f_{a,b,t_1}(x)$ is denoted by $f^{\#1}_{a,b,t_1}$. Finally, let f^{*}_{a,b,t_1} denote the set of points

$$\left\{ (x, f_{a,b,t_1}(x)) |\ x \in \left(\frac{1}{2} - \Delta(a), \frac{1}{2} + \Delta(a)\right) \right\}$$

on the "middle branch" of $f_{a,b,t_1}(x)$. Similarly, for $d > 4$, $f^{\#0}_{d,c,t_2}$, $f^{\#1}_{d,c,t_2}$ and f^{*}_{d,c,t_2} are used to denote the sets

$$\left\{ (f_{d,c,t_2}(y), y) |\ y \in \left(0, \frac{1}{2} - \Delta(d)\right) \right\},$$

$$\left\{ (f_{d,c,t_2}(y), y) |\ y \in \left(\frac{1}{2} + \Delta(d), 1\right) \right\},$$

and

$$\left\{ (f_{d,c,t_2}(y), y) |\ y \in \left(\frac{1}{2} - \Delta(d), \frac{1}{2} + \Delta(d)\right) \right\}$$

respectively.

Using the Theorem 4 we state the following corollary:

Corollary 4: *Assume that each of the neurons is either enthusiastic, or greedy and* $ad \geq bc/2$*. Then, attractive fixed points of (2) can lie only on the intersection of the outer branches of* f_{a,b,t_1}

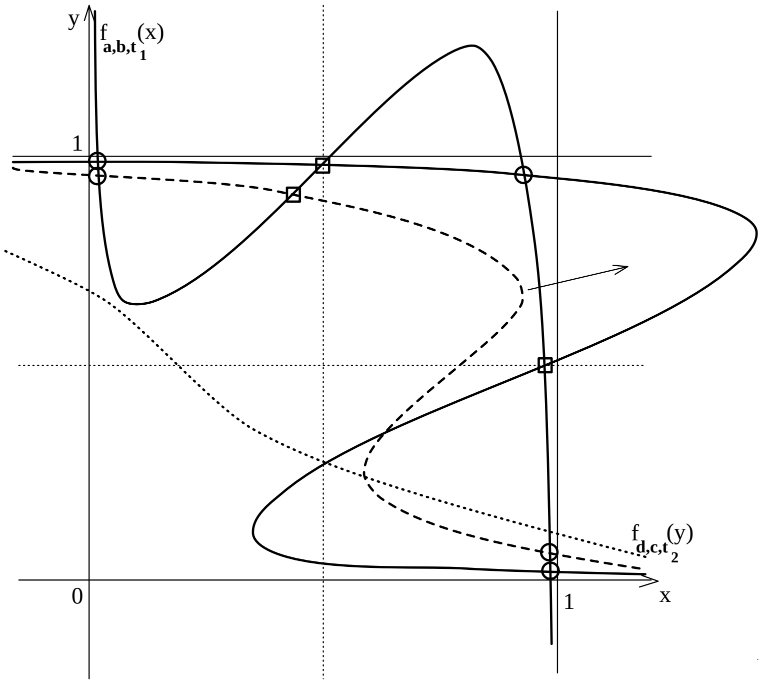

Figure 9: Geometrical illustration of saddle-node bifurcation in a recurrent neural network with two state neurons. Saddle and attractive points are marked with squares and circles respectively. $a, d > 0,\ b, c < 0$.

and f_{d,c,t_2}. Whenever the the middle branch of f_{a,b,t_1} intersects with an outer branch of f_{d,c,t_2} (or vice-versa), it corresponds to a saddle point of (2). In particular, all attractive fixed points of (2) are from

$$\bigcup_{i,j=0,1} f^{\#i}_{a,b,t_1} \cap f^{\#j}_{d,c,t_2}.$$

Every point from

$$f^{*}_{a,b,t_1} \cap \bigcup_{i=0,1} f^{\#i}_{d,c,t_2},$$

or

$$f^{*}_{d,c,t_2} \cap \bigcup_{i=0,1} f^{\#i}_{a,b,t_1}$$

is a saddle point of (2).

When both neurons self–excite themselves, Corollary 4 suggests that the usual scenario of creation of a new attractive fixed point is that typical of the saddle-node bifurcation in which a pair attractive + saddle fixed point is created. Attractive fixed points disappear in a reverse manner: an attractive point coalesces with with a saddle and they are annihilated. This is illustrated in figure 9. $f_{d,c,t_2}(y)$ shown as dashed curve intersects $f_{a,b,t_1}(x)$ in three points. By increasing d, f_{d,c,t_2} bends further (solid curve) and intersects with f_{a,b,t_1} in five points[16]. Saddle and attractive points are marked with squares and circles respectively. Note that as d increases attractive fixed points move closer to vertices $\{0, 1\}^2$.

[16] At the same time, $|c|$ has to be also appropriately increased so as to compensate for the increase in d so that the "bended" part of $f_{d,c}$ does not move radically to higher values of x.

This tendency, in the context of networks with exclusively self-exciting (or exclusively self-inhibiting) recurrent neurons, is discussed in [13]. Our result stated in Corollary 1, assumes two–neuron recurrent network. It only requires that the neurons have the same mutual interaction pattern ($bc > 0$) and gives a lower bound on the rate of convergence of the attractive fixed points of (2) towards some of the vertices $\{0,1\}^2$, as the absolute values of weights on the self–loops grow.

Corollary 1.1: *Assume* $bc > 0, |a| > 4, |d| > 4$. *Then all attractive fixed points of (2) lie in the* ε*-neighborhood of vertices of unit square, where*

$$\varepsilon = \sqrt{\left(\frac{1}{2} - \Delta(|a|)\right)^2 + \left(\frac{1}{2} - \Delta(|d|)\right)^2}.$$

6 Conclusion

The regions corresponding to stability types of fixed points of a two-neuron recurrent neural network were described based on the weight matrix of the network. The position of fixed points was investigated in the context of intersections of functions defining their x- and y-coordinates. It was shown that there is a correspondence between the stability regions for fixed points and monotonicity intervals of functions defining their position. When both neurons self-excite themselves and have the same mutual-interaction pattern, a new attractive fixed point is created through saddle node bifurcation. Assuming the same mutual interaction pattern between neurons, we give a lower bound on the rate of convergence of the attractive fixed points towards the saturated activation values, as the absolute values of weights on the self–loops grow.

Our ultimate goal is to extend the issues studied in this paper to a general case of n-neuron recurrent neural network. It is to be seen whether the reasoning in the space of derivatives of the sigmoid transfer function with respect to the weighted sum of neuron inputs, can be simplified to a more straightforward analysis of fixed point stability regions (as opposed to the case-analysis used in the proofs of this paper).

As explained in the introduction, training process during which recurrent neural networks learn to act as finite state machines can be interpreted from the point of view of bifurcation analysis [18]. Often, loops in state transition diagram of the finite state machine being learned are represented as attractive fixed points of the network. Understaning the fixed point potential of recurrent neural networks (number, stability, bifurcations of fixed points) can bring some light into the problem of neural complexity of finite state machines which (to our knowledge) has not been satisfactorily solved so far (see [1], [15] and [20]). Neural complexity of a finite state machine can be characterized as the minimal number of neurons needed so that the network can mimic the finite state machine.

References

[1] N. Alon, A.K. Dewdney, and T.J. Ott. Efficient simulation of finite automata by neural nets. *Journal of the Association of Computing Machinery*, 38(2):495–514, 1991.

[2] H. Anton. *Calculus with analytic geometry.* John Wiley and Sons, New York, NY, 1980.

[3] R.D. Beer. On the dynamics of small continuous–time recurrent networks. Technical Report CES–94–18, Case Western Reserve University, Cleveland, OH, 1994.

[4] E.K. Blum and X. Wang. Stability of fixed points and periodic orbits and bifurcations in analog neural networks. *Neural Networks*, (5):577–587, 1992.

[5] M.P. Casey. *Computation in Discrete-Time Dynamical Systems*. PhD thesis, University of California, San Diego, Department of Mathematics, March 1995.

[6] M.P. Casey. Relaxing the symmetric weight condition for convergent dynamics in discrete-time recurrent networks. Technical Report INC-9504, Institute for Neural Computation, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0112, 1995.

[7] A. Cleeremans, D. Servan-Schreiber, and J.L. McClelland. Finite state automata and simple recurrent networks. *Neural Computation*, 1(3):372–381, 1989.

[8] F. Cummins. Representation of temporal patterns in recurrent networks. *Submitted to the 15th Annual Conference of the Cognitive Science Society*, 1993.

[9] S. Das and M.C. Mozer. A unified gradient–descent/clustering architecture for finite state machine induction. In J.D. Cowen, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems 6*, pages 19–26. Morgan Kaufmann, 1994.

[10] K. Doya. Bifurcations in the learning of recurrent neural networks. In *Proc. of 1992 IEEE Int. Symposium on Circuits and Systems*, pages 2777–2780, 1992.

[11] J.L. Elman. Finding structure in time. *Cognitive Science*, 14:179–211, 1990.

[12] C.L. Giles, C.B. Miller, D. Chen, H.H. Chen, G.Z. Sun, and Y.C. Lee. Learning and extracting finite state automata with second–order recurrent neural networks. *Neural Computation*, 4(3):393–405, 1992.

[13] M.W. Hirsch. Saturation at high gain in discrete time recurrent networks. *Neural Networks*, 7(3):449–453, 1994.

[14] J.J. Hopfield. Neurons with a graded response have collective computational properties like those of two–state neurons. *Proceedings of the National Academy of Science USA*, 81:3088–3092, May 1984.

[15] B.G. Horne and D.R. Hush. Bounds on the complexity of recurrent neural network implementations of finite state machines. In J.D. Cowen, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems 6*, pages 359–366. Morgan Kaufmann, 1994. Also submitted to Neural Networks.

[16] L. Jin, P.N. Nikiforuk, and M.M. Gupta. Absolute stability conditions for discrete-time recurrent neural networks. *IEEE Transactions on Neural Networks*, (6):954–963, 1994.

[17] P. Manolios and R. Fanelli. First order recurrent neural networks and deterministic finite state automata. *Neural Computation*, 6(6):1155–1173, 1994.

[18] P. Tiňo, B.G. Horne, C.L. Giles, and P.C. Collingwood. Finite state machines and recurrent neural networks – automata and dynamical systems approaches. Technical Report UMIACS-TR-95-1, Institute for Advance Computer Studies, University of Maryland, College Park, MD 20742, 1995.

[19] P. Tiňo and J. Sajda. Learning and extracting initial mealy machines with a modular neural network model. *Neural Computation*, 7(4), 1995.

[20] H.T. Siegelmann, E.D. Sontag, and C.L. Giles. The complexity of language recognition by neural networks. In J. van Leeuwen, editor, *Algorithms, Software, Architecture (Proceedings of IFIP* 12^{th} *World Computer Congress)*, pages 329–335, Amsterdam, 1992. North–Holland.

[21] R.L. Watrous and G.M. Kuhn. Induction of finite–state languages using second–order recurrent networks. *Neural Computation*, 4(3):406–414, 1992.

[22] Z. Zeng, R.M. Goodman, and P. Smyth. Learning finite state machines with self–clustering recurrent networks. *Neural Computation*, 5(6):976–990, 1993.