# Saddle point and Hamiltonian structure in excitatory-inhibitory networks

**H. S. Seung, T. J. Richardson**
Bell Labs, Lucent Technologies
Murray Hill, NJ 07974
{seung|tjr}@bell-labs.com

**J. C. Lagarias**
AT&T Research
180 Park Ave. D-130
Florham Park, NJ 07932

**J. J. Hopfield**
Dept. of Molecular Biology
Princeton University
Princeton, NJ 08544

## Abstract

Because the dynamics of a neural network with symmetric interactions is similar to a gradient descent dynamics, convergence to a fixed point is the general behavior. In this paper, we analyze the global behavior of networks with distinct excitatory and inhibitory populations of neurons, under the assumption that the interactions between the populations are antisymmetric. Our analysis exploits the similarity of such a dynamics to a saddle point dynamics. This analogy gives some intuition as to why such a dynamics can either converge to a fixed point or a limit cycle, depending on parameters. We also show that the network dynamics can be written in a dissipative Hamiltonian form.

Dynamic neural networks with symmetric interactions provably converge to fixed points under quite general hypotheses [1, 2]. The convergence theory helped to establish the paradigm of dynamic neural computation with attractors[3, 4]. Little is known about the dynamics of networks whose connections are not symmetric. Expanding the convergence theory to include asymmetrical networks is nevertheless well motivated. First, the networks in real brains are asymmetric. Second, oscillations and complex nonperiodic behavior are observed in real brains, and these types of dynamical behavior cannot be realized in a symmetric network. Third, asymmetric networks admit dynamic behavior which offer a wider range of possibilities for neural computation.

In this paper, we consider a special class of asymmetric networks that consist of excitatory and inhibitory populations of neurons. In particular, we consider networks with two distinguished populations of neurons which interact via antisymmetrical connections, while within each population interactions are symmetric. Thus the class of networks considered here includes perhaps the most basic asymmetry of interaction in the brain, the asymmetry of excitation and inhibition. The networks we consider offer a rich repertoire of dynamical behaviors including oscillations[8] and traveling waves[5].

The mathematical content of the paper is the following. We define a general class of excitatory-inhibitory networks and introduce a Lyapunov function that establishes sufficient conditions for the global stability of fixed points. It is straightforward to verify the validity of the construction by a simple calculation. We apply the construction to a simple example network consisting of an excitatory population of neurons with recurrent inhibition from a single neuron[6, 7]. Because of space constraints we will concentrate on the formal aspects of our results. Moreover, we will make simplifying assumptions which will be relaxed in the full version of the paper.

What makes the Lyapunov function particularly interesting, over and above the convergence results it yields, is that it can be interpreted within a broad framework which borrows concepts from classical mechanics. We present two such complementary interpretations.

The first interpretation exploits the similarity of the excitatory-inhibitory network to saddle point dynamics. According to this viewpoint, the dynamical equations for the excitatory and inhibitory neurons are similar to gradient descent and ascent on a saddle function. The fixed points of the dynamics are saddle points of this function. The second method interprets the dynamics within a Hamiltonian framework and the Lyapunov function then arises as the energy of the dissipative Hamiltonian system.

We can apply the general constructions to yield sufficient conditions for convergence to fixed points excitatory-inhibitory networks. Equivalently, we obtain necessary conditions for oscillatory behavior. The more difficult task of finding sufficient conditions for oscillatory or chaotic behavior remains largely open. However, we hope that the analogies with saddle point and Hamiltonian dynamics will help in this task which will the subject of further work.

## 1 Excitatory-inhibitory network

Figure 1 depicts a network of two populations of neurons. The state of the network is described by two vectors $x \in R^m$ and $y \in R^n$ which represent the activities of the excitatory and inhibitory neurons respectively. The connections between the two populations are antisymmetric, while the connections within each population are symmetric. The dynamics of the network is given by

$$\tau_x \dot{x} + x = f(u - Ay + Bx) \,, \tag{1}$$

$$\tau_y \dot{y} + y = g(v + A^T x - Cy) \,. \tag{2}$$

Here we use $f$ to denote a vector of scalar functions which are applied component wise to the arguments, i.e., $f(x) = (f_1(x_1), \ldots, f_m(x_m))$. The scalar functions $f_i$ are generally assumed to monotonic non-decreasing. In order to simplify the notation we will assume that all of $f_i$ are the same, and we use $f$ to denote both the vector and scalar versions. Similarly, $g$ represents a vector of non-increasing scalar functions which, for convenience, we will assume are identical. To clarify the interpretation of the "excitatory and inhibitory" populations we should impose the additional constraint that the entries of matrices $A$, $B$, and $C$ are nonnegative, in which case the meaning becomes clear. However, although this assumption makes sense in a neurobiological context the mathematics in no way depends on it. The antisymmetry of interaction between $x$ and $y$ is manifest in the equations. The symmetry of interaction within each population is imposed by the constraints $B = B^T$ and $C = C^T$. The constant vectors $u$ and $v$ represent tonic input from external sources (or alternatively bias intrinsic to the neurons). The time constants $\tau_x$ and $\tau_y$ determine the speed of excitatory and inhibitory synapses, respectively. The ratio between these time constants greatly affects the global behavior of the network. In particular, when inhibition is much slower than excitation, $\tau_y \gg \tau_x$, oscillations may result.

The potential for oscillatory behavior in excitatory-inhibitory networks like (1) has long been known[8, 9]. The origin of oscillations can be understood from a simple two neuron model. Suppose that neuron 1 excites neuron 2, and receives inhibition back from neuron 2. Then the effect is that neuron 1 suppresses its own activity with an effective delay that depends on the time constant of inhibition. If this delay is long enough, oscillations result. However, these oscillations will die down to a fixed point, as the inhibition tends to dampen activity in the circuit. Only if neuron 1 also excites itself can the oscillations become sustained.

Therefore, whether oscillations are damped or sustained depends on the choice of parameters. In this paper we establish sufficient conditions for the global stability of fixed points in (1). The sufficient conditions indicate regimes in which it may be possible to prove other types of dynamical behavior, such as oscillations.

## 2 Lyapunov function

In this section we formally introduce our Lyapunov function and indicate conditions under which it can be used to prove global convergence to fixed points. A more detailed development will appear in a full paper.

As a preliminary, we establish some notational conventions. We define the antiderivatives $F$ and $\bar{F}$ componentwise,

$$F'(x) = f(x) \qquad \bar{F}'(x) = f^{-1}(x) \tag{3}$$

The antiderivatives $G$ and $\bar{G}$ are defined similarly.

For the purposes of this section it is convenient to assume that $f$ and $g$ are smooth. Furthermore we assume that all arguments are within the range of functions applied to them. In particular, $f$ may be defined on all of $R$ but may be bounded above and/or below so that $f^{-1}$ will be defined on some sub-interval of $R$. Note that the set $(x, y)$ lying in the range of $(f, g)$ is a positive invariant set under (1) and that its closure is a global attractor for the system.

The column vector of all ones is denoted by $\mathbf{1}$. Its dimensionality should be clear from context. It is useful to define the following functions

$$\begin{aligned} K_x(p, x) &= \frac{1}{\tau_x}\left(\mathbf{1}^T F(p) - x^T p + \mathbf{1}^T \bar{F}(x)\right) && (4) \\ K_y(q, y) &= \frac{1}{\tau_y}\left(\mathbf{1}^T G(q) - y^T q + \mathbf{1}^T \bar{G}(y)\right) . && (5) \end{aligned}$$

Note that both functions are convex in either of their arguments but they are not convex in both arguments in general. If $p = u - Ay + Bx$ then $K_x(p, x)$ is formally analogous to the kinetic energy associated with the $x$ variables. The function $K_x(p, x)$ attains its minimal value on the manifold $f(p) = x$. Similar statements apply to $K_y$.

The functions $K_x$ and $K_y$, along with the saddle function

$$S := -u^T x - \frac{1}{2}x^T Bx + v^T y - \frac{1}{2}y^T Cy + \mathbf{1}^T \bar{F}(x) + y^T A^T x - \mathbf{1}^T \bar{G}(y) \tag{6}$$

are the three pieces used to construct the Lyapunov function. The reason for the name "saddle function" will be explained later. A straightforward computation shows that the time derivative of

$$L := K_x(u - Ay + Bx, x) + K_y(v + A^T x - Cy, y) + rS \tag{7}$$

is given by

$$\begin{aligned}
\dot{L} &= \dot{x}^T B\dot{x} - \dot{y}^T C\dot{y} && (8)\\
&\quad -(\tau_x^{-1} + r)\dot{x}^T[u - Ay + Bx - f^{-1}(x)] && (9)\\
&\quad -(\tau_y^{-1} - r)\dot{y}^T[v + A^T x - Cy - g^{-1}(y)] && (10)
\end{aligned}$$

It will be seen that the functional $L$ can be of considerable utility in analyzing (1). Under various conditions on the parameters, the function $L$ serves as a Lyapunov function and it can therefore be used to prove global convergence of the network.

Note that $u - Ay + Bx = f^{-1}(\tau_x\dot{x} + x)$. Since $f^{-1}$ is monotonically increasing, $f^{-1}(a) - f^{-1}(b)$ has the same sign as $a - b$. Similar statements apply to $g$. It follows that if $-\tau_x^{-1} \leq r \leq \tau_y^{-1}$, then last two terms of (8) are nonpositive.

Without further constraints, the first two terms of $\dot{L}$ may be of arbitrary sign. In many cases the value of $r$ can be adjusted so that the first two terms of (8) are dominated by the second two.

For example, if $r$ can be chosen so that

$$\begin{aligned}
(a-b)^T B(a-b) &\leq (\tau_x^{-1} + r)(a-b)^T[f^{-1}(a) - f^{-1}(b)] && (11)\\
(a-b)^T C(a-b) &\geq (r - \tau_y^{-1})(a-b)^T[g^{-1}(a) - g^{-1}(b)] && (12)
\end{aligned}$$

for all $a$ and $b$ then we can conclude $\dot{L} \leq 0$. Assuming $L$ is bounded below, it then serves as a Lyapunov function of the dynamics.

To see the construction in action we turn to a simple example.

## 3 An Example: Global Inhibition

As an illustration of the use of the Lyapunov function, we consider a network with a single inhibitory neuron connected to a population of excitatory neurons. Each excitatory neuron feeds back to itself with strength $\beta$.

$$\begin{aligned}
\dot{x}_i + x_i &= [u_i - y + \beta x_i]^+ \ , && (13)\\
\dot{y} + y &= \sum_i x_i \ . && (14)
\end{aligned}$$

Here we have introduced the notation $[x]^+ = \max(x, 0)$. This is a special case of (1). The network is intended to minimize, at least locally, the functional

$$E := -\sum_i u_i x_i + \frac{1}{2}\left(\sum_i x_i\right)^2 + \frac{1-\beta}{2}\sum_i x_i^2$$

under the constraint $x_i \geq 0$. With the help of the Lyapunov function we can prove convergence to minima when $\beta < 2$. For convenience we assume $u_1 > u_2 > ... > u_n$. Let $I_{\text{on}}$ denote the subset of indices such that $i \in I_{\text{on}}$ if and only if $u_i - y + \beta x_i > 0$. Note that $I_{\text{on}}$ is time varying.

When $\beta < 1$ the network converges to the unique global minimum of $E$ which is characterized as follows. There is (generically) NOT TRUE a unique $k \in 1, ..., n$ such that $u_k > \frac{1}{k+1-\beta} \sum_{i=1}^{k} u_i > u_{k+1}$. The equations

$$\begin{aligned} x_i &= [u_i - y + \beta x_i]^+ \ , \\ y &= \frac{1}{k+1-\beta} \sum_{i=1}^{k} u_i \end{aligned}$$

define a fixed point of (13). This fixed point is a global attractor and minimizer of $E$. Note that $I_{\text{on}} = \{1, ..., k\}$ at the fixed point.

When $1 < \beta < 2$ the local minima of $E$ are of the "winner take all" type. There is (generically) a unique $k \in 1, ..., n$ such that $u_k > (2 - \beta) u_1 > u_{k+1}$. With the help of the Lyapunov function we can prove that (13) generically converges to one of $k$ fixed points characterized by: $I_{\text{on}} = \{j\}$ where $j \leq k$; $x_j = \frac{u_j}{2-\beta}$; $x_i = 0, i \neq j$.

For this example it is convenient to choose $r = 1 - \beta$ in which case the Lyapunov function reduces to

$$L = \dot{y}^2 + (1 - \beta) y^2 + \sum_{i=1}^{n} (y - u_i)^2 - \sum_{i \notin I_{\text{on}}} (y - u_i - \beta x_i)^2 ,$$

and its derivative is given by

$$\dot{L} = -2(2 - \beta) \dot{y}^2 - \beta \sum_{i \notin I_{\text{on}}} (y - u_i - \beta x_i) x_i \ .$$

Thus, for $\beta < 2$ it follows that $(x, y)$ converges to the positive invariant portion of the set $\{\dot{y} = 0\} \cap \{x_i = 0, i \notin I_{\text{on}}\}$. ¿From this one can conclude the convergence results described above.

This example has illustrated how the Lyapunov function (7) can be used to prove global stability of fixed points in excitatory-inhibitory networks. However, the origin of the Lyapunov function remains mysterious. It turns out that there are two "methods" by which the Lyapunov function can be derived, which give a deeper understanding of the mathematics of excitatory-inhibitory networks.

In the case $\beta > 2$ the fixed points of (13) are unstable. In this regime oscillations may arise as can be easily demonstrated for the case $n = 1$. For $n > 1$ oscillatory or chaotic behavior may arise.

## 4 Saddle dynamics

To see why the saddle function (6) deserves its name, consider the following,

$$-\frac{\partial S}{\partial x} = u - Ay + Bx - f^{-1}(x) \tag{15}$$

$$\frac{\partial S}{\partial y} = v + A^T x - Cy - g^{-1}(y) \tag{16}$$

The right hand sides of these equations have the same sign as $\dot{x}$ and $\dot{y}$ respectively by the monotonicity of $f$ and $g$. If $f$ and $g$ were the identity function, exact equality would hold. Let us assume that this is the case. The equations of motion (1) are then saddle point dynamics: They are gradient descent in $x$ and gradient ascent in $y$. Under certain circumstances, such a dynamics converges to a saddle point

$$\min_x \max_y S(x, y) \tag{17}$$

Sufficient conditions for convergence can be found by Lyapunov methods. The time derivative of the kinetic energy $T = \frac{1}{2}\tau_x|\dot{x}|^2 + \frac{1}{2}\tau_y|\dot{y}|^2$ is given by $-\dot{x}^T(\partial^2 S/\partial x^2)\dot{x} + \dot{y}^T(\partial^2 S/\partial y^2)\dot{y}$. If $\partial^2 S/\partial x^2$ is positive definite and $\partial^2 S/\partial y^2$ is negative definite, then $T$ is nonincreasing and is a Lyapunov function, since it is always bounded below.

If the kinetic energy is not a Lyapunov function, it can sometimes be fixed by adding a multiple of the saddle function. The time derivative of the saddle function is $\dot{S} = -\tau_x\dot{x}^2 + \tau_y\dot{y}^2$. By taking the linear combination $L = T + rS$, the terms that depend on $\dot{x}$ and $\dot{y}$ can be traded off against each other, so as to produce a nonincreasing $L$. When generalized to the case of nonlinear $f$ and $g$, this Lyapunov function turns into (7).

The resemblance of an excitatory-inhibitory network to a saddle point dynamics should be contrasted with the resemblance of a symmetric network to a gradient descent dynamics. Since gradient descent is a sure way of finding a local minimum of a potential function, fixed points of symmetric networks are almost always globally stable. A saddle point dynamics is clearly an unreliable way of finding a saddle point. In particular, it is easy to construct situations where a saddle point dynamics leads to a limit cycle about a saddle point. Furthermore, even when the dynamics does converge to the saddle point, damped oscillations are a clear possibility. Thus this viewpoint gives some insight into why excitatory-inhibitory networks often show oscillatory behavior.

## 5 Hamiltonian dynamics

The Lyapunov function (7) can also be derived from a dissipative Hamiltonian dynamics. To do this, we double the dimension of the state space $(x, y)$ by adding canonically conjugate momenta $(p_x, p_y)$, and consider the phase space dynamics

$$\begin{aligned}
\tau_x\dot{x} + x &= f(p_x) && (18)\\
\tau_y\dot{y} + y &= g(p_y) && (19)\\
\left(r + \frac{d}{dt}\right)(u - Ay + Bx - p_x) &= 0 && (20)\\
\left(r + \frac{d}{dt}\right)(v + A^T x - Cy - p_y) &= 0 && (21)
\end{aligned}$$

Clearly the linear space defined by $p_x = u - Ay + Bx$ and $p_y = v + A^T x - Cy$ is an attractive invariant manifold of this dynamics, provided that $r > 0$. On this invariant manifold, the phase space dynamics is equivalent to the state space dynamics (1).

The advantage of the phase space formulation is that the equations of motion can be written in canonical form with velocity-dependent terms,

$$\begin{aligned}
\dot{x} &= \frac{\partial H}{\partial p_x} && (22)\\
\dot{y} &= \frac{\partial H}{\partial p_y} && (23)
\end{aligned}$$

$$\dot{p_x} = -\frac{\partial H}{\partial x} + B\dot{x} - A\dot{y} - (\tau_x^{-1} + r)[p_x - f^{-1}(x)] \quad (24)$$

$$\dot{p_y} = -\frac{\partial H}{\partial y} + A^T\dot{x} - C\dot{y} - (\tau_y^{-1} - r)[p_y - g^{-1}(y)] \quad (25)$$

$$+2r(v + A^T x - Cy - p_y) \quad (26)$$

where the Hamiltonian is defined by

$$H(p_x, x, p_y, y) = K_x(p_x, x) + K_y(p_y, y) + rS(x, y) \quad (27)$$

On the invariant manifold, the Hamiltonian is identical to the Lyapunov function (7) defined previously. The rate of energy dissipation is given by

$$\dot{H} = \dot{x}^T B\dot{x} - (\tau_x^{-1} + r)\dot{x}^T[p_x - f^{-1}(x)] \quad (28)$$

$$-\dot{y}^T C\dot{y} - (\tau_y^{-1} - r)\dot{y}^T[p_y - g^{-1}(y)] \quad (29)$$

$$+2r\dot{y}^T(v + A^T x - Cy - p_y) \quad (30)$$

The last term vanishes on the invariant manifold, leaving a result identical to (8). Of course, it is also possible that the velocity-dependent terms may pump energy into the system, rather than dissipate it in which case oscillations or chaotic behavior may arise.

## 6 Conclusion

The dynamics of an excitatory-inhibitory network is closely related to a saddle point dynamics. Unlike a gradient dynamics, a saddle point dynamics can converge to a fixed point or to a limit cycle. This analogy gives some insight into the origins of oscillatory behavior in excitatory-inhibitory networks. Furthermore, it aids in the construction of Lyapunov stability arguments. The dynamics of an excitatory-inhibitory network also has a Hamiltonian structure.

## References


[1] M. A. Cohen and S. Grossberg. Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE*, 13:815–826, 1983.

[2] J. J. Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proc. Natl. Acad. Sci. USA*, 81:3088–3092, 1984.

[3] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proc. Nat. Acad. Sci. USA*, 79:2554–2558, 1982.

[4] S. Grossberg. Nonlinear neural networks: principles, mechanisms, and architectures. *Neural Networks*, 1:17–61, 1988.

[5] S. Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biol. Cybern.*, 27:77–87, 1977.

[6] S. Amari and M. A. Arbib. *Competition and cooperation in neural nets*, pages 119–165. Academic Press, New York, 1977.

[7] A. L. Yuille and N. M. Grzywacz. A winner-take-all mechanism based on presynaptic inhibition feedback. *Neural Comput.*, 1:334–347, 1989.

[8] H. R. Wilson and J. D. Cowan. A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik*, 13:55–80, 1973.

[9] B. Ermentrout. Complex dynamics in winner-take-all neural nets with slow inhibition. *Neural Networks*, 5:415–431, 1992.