*LETTER* Communicated by Garrett Stanley

# How Optimal Stimuli for Sensory Neurons Are Constrained by Network Architecture

**Christopher DiMattina**
*chris_dimattina@yahoo.com*
*Department of Neuroscience, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, U.S.A.*

**Kechen Zhang**
*kzhang4@jhmi.edu*
*Department of Biomedical Engineering, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, U.S.A.*

**Identifying the optimal stimuli for a sensory neuron is often a difficult process involving trial and error. By analyzing the relationship between stimuli and responses in feedforward and stable recurrent neural network models, we find that the stimulus yielding the maximum firing rate response always lies on the topological boundary of the collection of all allowable stimuli, provided that individual neurons have increasing input-output relations or gain functions and that the synaptic connections are convergent between layers with nondegenerate weight matrices. This result suggests that in neurophysiological experiments under these conditions, only stimuli on the boundary need to be tested in order to maximize the response, thereby potentially reducing the number of trials needed for finding the most effective stimuli. Even when the gain functions allow firing rate cutoff or saturation, a peak still cannot exist in the stimulus-response relation in the sense that moving away from the optimum stimulus always reduces the response. We further demonstrate that the condition for nondegenerate synaptic connections also implies that proper stimuli can independently perturb the activities of all neurons in the same layer. One example of this type of manipulation is changing the activity of a single neuron in a given processing layer while keeping that of all others constant. Such stimulus perturbations might help experimentally isolate the interactions of selected neurons within a network.**

## 1 Introduction

In sensory systems, the firing rate of a neuron often varies in a systematic manner with the parameters of the presented stimulus (Adrian, 1928). Characterizing the changes in mean neural firing rate with variations in stimulus

 

parameters remains the most basic step to explore the function of sensory systems, although temporal response properties and correlated responses among neurons may carry additional information about the sensory stimuli (see references in Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1997). It is sometimes of interest to identify the most effective or optimal stimulus for a neuron, that is the stimulus among all possible sensory inputs that elicits the maximum firing rate response. The intuitive idea about the optimal stimulus requires this stimulus to be a strict maximum or a peak in the sense that any excursion away from this stimulus in stimulus parameter space results in a decrease in firing rate. A strict maximum may not always exist. For example, when the stimulus-response relation of an auditory neuron is described by a function of quadratic form (Yu & Young, 2000), a strict maximum is impossible if the quadratic form is a saddle (Zhang, Anderson, & Young, 2004). From a theoretical point of view, whether a peak exists in the stimulus-response relation of a sensory neuron should be constrained ultimately by the architecture of the underlying neural network.

As illustrated in Figure 1, the location of the optimal stimulus for a neuron depends on the details of the synaptic connections. The input-output relation of the neurons here is the logistic gain function $g(x) = 1/(1 + e^{-x})$ (see Figure 1A). A single stimulus input $x$ in the range $[a, b]$ projects to two neurons, whose outputs are then weighted and summed to yield a single output $r$. Depending on the connection weights, the maximum response may occur either on the boundary (see Figure 1B) or in the interior (see Figure 1C) of the input space $[a, b]$. Imagine that one is trying to find the optimal stimulus for the output neuron. If one knows beforehand that the optimal stimulus lies only on the boundary as in Figure 1B, then one need only sample the two boundary points $a$ and $b$ of the stimulus interval. In contrast, when the response peaks somewhere in the middle, one may have to test many stimuli to locate the peak. In this letter, we generalize these notions about the nature of the optimal stimulus to arbitrary feedforward and recurrent networks.

A second and seemingly unrelated question is the extent to which one can control the activity pattern of a group of neurons within a neural network using sensory inputs only. It turns out that network controllability is related to whether the optimal stimulus occurs only at the boundary of the stimulus set, and the conditions depend on the anatomical patterns of convergence and divergence between the layers of the network.

## 2 Allowable Stimuli Should Form a Compact Set

In biological sensory systems, the input is provided by the primary sensory receptors, like retinal photoreceptors, cochlear hair cells, or mechanoreceptors in the skin. The activity of this initial sensory transduction layer forms a vector in a finite dimensional space that represents the stimuli impinging

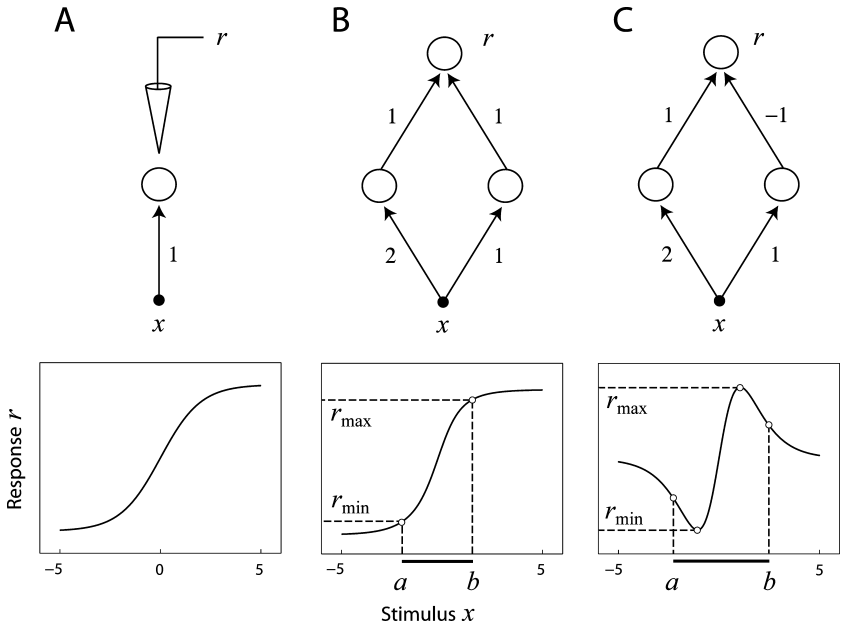


Figure 1: The stimulus that elicits the strongest response may occur at either the interior of the stimulus set or its boundary, depending on the network architecture. The stimulus set (the collection of all allowable stimuli) here is the interval $[a, b]$. **(A)** The standard logistic gain function is used for all neurons in this figure. The cone indicates a recording electrode. **(B)** The response $r$ of the neuron at the top attains its maximum at the boundary $b$ of the stimulus set (horizontal bar). **(C)** The neuron at the top attains its maximum response at the interior of the stimulus set.

on it. Since the stimuli used in sensory experiments are produced by devices like computer monitors, tactile probes, and sound generators, which have physical limits to the stimuli that can be produced, the space of all possible stimuli that can be produced is bounded by these constraints.

For visual stimuli, for instance, a computer monitor may be able to vary pixel intensity only between a minimum $I_{\min}$ and a maximum $I_{\max}$. Suppose there are $N$ pixels in the display; then the luminance of each pixel is bounded by

$$I_{\min} \leq I_i \leq I_{\max}, \quad i = 1, 2, \ldots, N. \tag{2.1}$$

Therefore, all images producible by the monitor lie within an $N$-dimensional hypercube given by $[I_{\min}, I_{\max}] \times \cdots \times [I_{\min}, I_{\max}]$ with $N$ factors for the

intensities of all the pixels. The boundary of this hypercube consists of stimuli with at least one pixel at either minimum or maximum luminance.

Another example is a sound-generating system constrained by the total output power. Suppose a sound stimulus is synthesized by summing sinusoidal waves of distinct frequencies:

$$s(t) = \sum_{i=1}^{N} a_i \cos(\omega_i t + \phi_i), \tag{2.2}$$

with frequencies $\omega_i$, amplitudes $a_i$, and phases $\phi_i$. The maximum total output power of the sound-generating system imposes an upper bound,

$$\sum_{i=1}^{N} a_i^2 \leq E, \tag{2.3}$$

with $E$ proportional to the maximum power. All sounds that can be generated here constitute a solid ball of radius $\sqrt{E}$ in the $N$-dimensional space of the amplitudes $(a_1, \ldots, a_N)$. A stimulus with the maximum power lies on the spherical surface, the boundary of the solid ball. In addition, the circular phase variables $\phi_1, \ldots, \phi_N$ form an $N$-dimensional torus. In the special case of stationary sound stimuli characterized by the spectral amplitudes but not the transients, the phase variables can be ignored (Yu & Young, 2000).

We assume that a sensory stimulus can be described by a finite set of parameters or a point in a finite dimensional Euclidean space. This assumption seems reasonable for all controlled experiments because in any conceivable experiment in a lab, one can independently manipulate only a finite number of stimulus parameters. For natural stimuli such as visual scenes and sounds from the environment, one can still describe any given set of stimuli to arbitrary precision by choosing a finite but large enough set of parameters. We assume further that the collection of all permissible or allowable stimuli in an experiment or in a natural environment should form a compact set, which will be denoted by $X$. *Compact* means bounded and closed in topology (Munkres, 1999). *Bounded* means that $X$ can be put inside a ball of finite radius. *Closed* means that $X$ includes all its boundary points. *Boundary point* is defined by the property that every ball, no matter how small, centered at a boundary point of $X$ should contain points both inside $X$ and outside $X$. By contrast, an *interior point* of $X$ is the center of any ball contained completely in $X$. The hypercube and the ball considered above are both compact as long as the boundary surfaces are included. We make the distinction between boundary point and interior point for mathematical convenience. As the distance between an interior point and a boundary point diminishes, their distinction eventually becomes meaningless for practical experiments.

A compact stimulus set is always a union of two disjoint sets: an interior that contains all the interior points and a boundary that contains all the boundary points (Munkres, 1999). The interior is possibly an empty set. The compactness condition is useful because a continuous function always attains its maximum and minimum on a compact set, and the image is also compact (Rudin, 1976). The input-output relation of a neuron is always continuous for the network models considered in this letter. Hence, given a compact set of allowable stimuli, a neuron always has a stimulus that elicits the strongest response as long as the input-output relation of the neuron is continuous. Moreover, the set of all responses of the neuron must also be compact. The compactness property also holds for the response patterns of a group of neurons. For brevity, the interior and the boundary of the allowable stimulus set are called *interior stimuli* and *boundary stimuli*, respectively. Similarly, *interior responses* and *boundary responses* are the interior and the boundary of the set of responses to all allowable stimuli, respectively.

As a special case, it is possible to have a compact stimulus set without an interior. Such a stimulus set consists entirely of boundary stimuli. For instance, let the sound signals in equation 2.2 have a constant energy; then inequality 2.3 becomes the equality $\sum_{i=1}^{N} a_i^2 = E$. This stimulus set corresponds to an $(N-1)$-dimensional sphere embedded in the $N$-dimensional Euclidean space of the amplitudes $(a_1, \ldots, a_N)$. The sphere has no interior and consists entirely of boundary points with respect to the topology of the Euclidean space of $(a_1, \ldots, a_N)$. In the theorems in this letter, we assume that our stimulus sets are compact with both interior and boundary.

The assumption of compactness requires that stimuli be suitably parameterized. For example, the luminance $I_i$ of pixel $i$ in equation 2.1 may be expressed by a new parameter on a logarithmic scale, $L_i = \log((I_{\max} - I_{\min})/(I_i - I_{\min}))$, which grows monotonically without bound as $I_i$ approaches $I_{\min}$. The hypercube considered above is no longer compact in terms of $L_i$.

How does one choose a suitable stimulus parameterization from many mathematically equivalent expressions? Since our focus here is on the biological neural network in the sensory system, the most important consideration is that the stimulus parameters should correspond to the input to the first layer of the biological neural network. For instance, the luminance of a pixel on a computer screen drives the activity of retinal photoreceptors, or the amplitude of a frequency component in a sound drives the activity of hair cells. The ionic current into a neuron is bounded naturally, and this is the real justification on assuming a bounded stimulus set. The main results in this letter do not rely on the exact details of stimulus parameters. General topological assumptions on compactness and the distinction between interior and boundary are sufficient for our purposes.

## 3 Network Architecture

**3.1 Gain Functions Are Increasing and Differentiable.** The input-output relation of each individual neuron is specified by a gain function that describes the dependence of the mean firing rate on the input. We assume that gain functions are increasing and differentiable with continuous derivatives. A gain function $g(u)$ is called *increasing* if $g(a) \leq g(b)$ for all $a < b$ and *strictly increasing* if $g(a) < g(b)$ for all $a < b$. A sufficient condition for being increasing or strictly increasing is $g'(u) \geq 0$ or $g'(u) > 0$ for all $u$, respectively. We further assume that zero derivative $g'(u) = 0$ may occur for a continuous interval of input $u$ but not at an isolated inflection point. This assumption allows a gain function to hold constant over some input interval, such as a zero firing rate for subthreshold inputs. These assumptions are biologically reasonable, and they cover most of the gain functions commonly used in neural modeling, some of which are illustrated in Figure 2.

Continuous differentiability helps simplify the mathematical treatment in the rest of the letter. Discontinuities in the derivative, like those caused by thresholds, can be readily smoothed out by approximation methods like the simple "surgery" shown in Figure 2C. The range of stimuli affected by the surgery is $[u_1, u_2]$, which can be made arbitrarily small so that the approximation becomes practically indistinguishable from the original model. Given any $u_1 < u_2$ and $g(u_1) < g(u_2)$, we replace the original gain function $g(u)$ in the interval $u \in [u_1, u_2]$ by constructing a continuously differentiable and strictly increasing function:

$$f(u) = g(u_1) + \int_{u_1}^{u} \phi(z)\, dz, \tag{3.1}$$

whose derivative is $f'(u) = \phi(u)$. Thus, $f(u)$ is strictly increasing if we choose $\phi(u) > 0$ for all $u \in [u_1, u_2]$. To guarantee the continuity of the new gain function and its derivative at the boundary points $u_1$ and $u_2$, we need the following constraints: $\phi(u_1) = g'(u_1)$, $\phi(u_2) = g'(u_2)$, and

$$\int_{u_1}^{u_2} \phi(z)\, dz = g(u_2) - g(u_1). \tag{3.2}$$

It is easy to find a continuous and positive $\phi(u)$ that satisfies all these constraints. Higher-order derivatives can be made continuous as well, but we need only the first derivative in this letter.

**3.2 Feedforward Network.** The general feedforward network considered in this letter is equivalent to a multilayer perceptron (Rumelhart, Hinton, & McClelland, 1986) which includes the original perceptron as

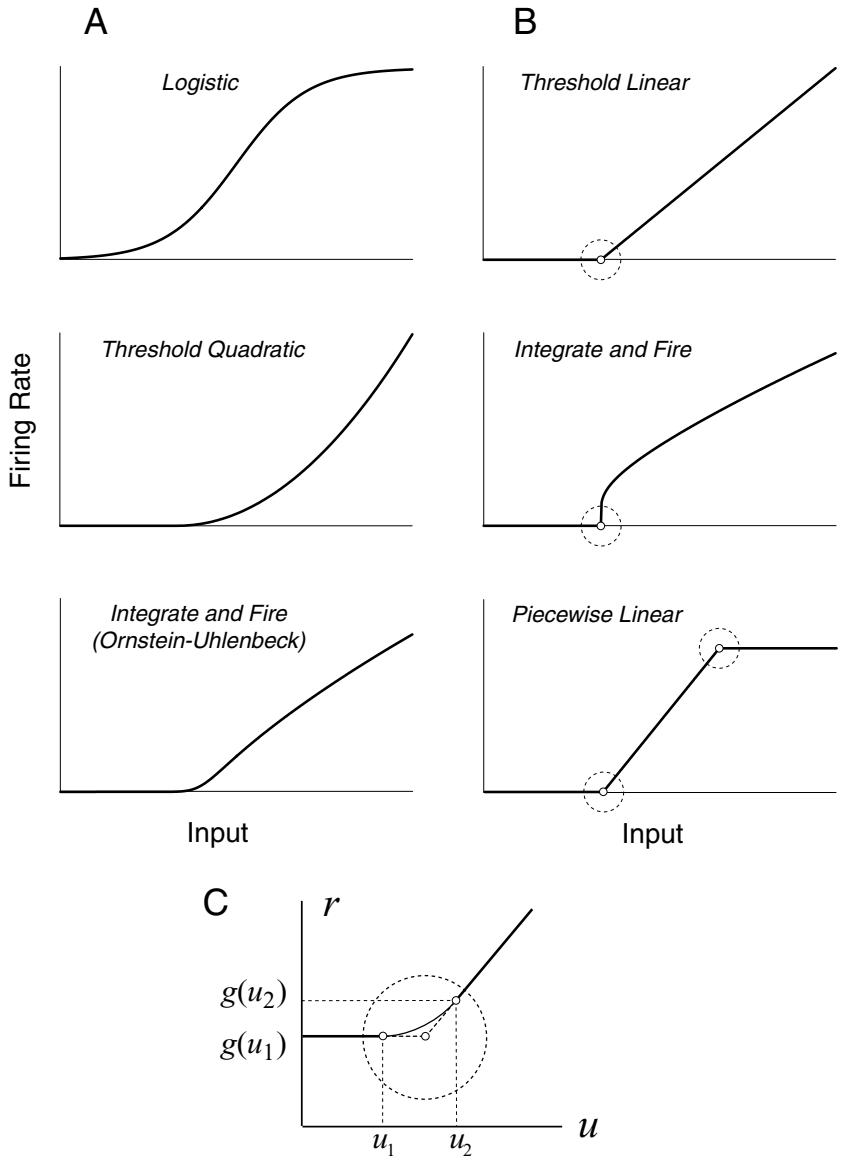


Figure 2: Examples of commonly used neuronal gain functions. **(A)** These gain functions are increasing and continuously differentiable. **(B)** These gain functions are increasing but not differentiable at the circled points. **(C)** Removal of a singularity by replacing the original function (dashed line segments) within the interval $[u_1, u_2]$ by a smooth, strictly increasing function (thin solid curve), with matching slopes at the end points $u_1$ and $u_2$.

a special case (Rosenblatt, 1962; Minsky & Papert, 1988). The activity $r_{i,m}$ of neuron $i$ in layer $m$ ($2 \leq m \leq M$) is given by

$$r_{i,m} = g_{i,m}\left(\sum_{j=1}^{K_{m-1}} w_{ij}^{(m)} r_{j,m-1}\right) = g_{i,m}\left(\mathbf{w}_i^{(m)} \cdot \mathbf{r}_{m-1}\right), \tag{3.3}$$

where $g_{i,m}$ is the gain function of layer $m$ neuron $i$, which receives synaptic connection with weight $w_{ij}^{(m)}$ from layer $m-1$ neuron $j$ with activity $r_{j,m-1}$, and $K_{m-1}$ is the number of neurons in layer $m-1$ (see Figure 3A). The gain functions of different neurons are allowed to be distinct. Weight $w_{ij}^{(m)}$ is positive if neuron $j$ is excitatory and negative if it is inhibitory, as usual. The last step in equation 3.3 replaces the summation in the first step by the inner product between the weight vector $\mathbf{w}_i^{(m)} = (w_{i1}^{(m)}, w_{i2}^{(m)}, \ldots, w_{iK_{m-1}}^{(m)})$ and the activity vector $\mathbf{r}_{m-1} = (r_{1,m-1}, r_{2,m-1}, \ldots, r_{K_{m-1},m-1})^{\mathrm{T}}$, with T indicating transpose. For convenience, we rewrite equation 3.3 in an equivalent vector form:

$$\mathbf{r}_m = \mathbf{g}_m\left(\mathbf{W}^{(m)}\mathbf{r}_{m-1}\right) = \begin{bmatrix} g_{1,m}\left(\mathbf{w}_1^{(m)} \cdot \mathbf{r}_{m-1}\right) \\ g_{2,m}\left(\mathbf{w}_2^{(m)} \cdot \mathbf{r}_{m-1}\right) \\ \vdots \\ g_{K_m,m}\left(\mathbf{w}_{K_m}^{(m)} \cdot \mathbf{r}_{m-1}\right) \end{bmatrix}, \tag{3.4}$$

where $\mathbf{g}_m = (g_{1,m}, g_{2,m}, \ldots, g_{K_m,m})$ gives the gain functions for all the neurons in layer $m$ and $\mathbf{W}^{(m)}$ is the $K_m \times K_{m-1}$ weight matrix whose $i$th row is $\mathbf{w}_i^{(m)}$. The recursive relation 3.4 is valid starting from the second layer ($m \geq 2$), while the first layer ($m = 1$) is the stimulus input vector given by

$$\mathbf{r}_1 = \mathbf{x} = \left(x_1, x_2, \ldots, x_{K_1}\right)^{\mathrm{T}}. \tag{3.5}$$

The feedforward network architecture illustrated in Figure 3A has neurons ordered into hierarchical layers, and connections are allowed between neurons only in consecutive layers. Any network without a closed loop of connections is functionally equivalent to a feedforward network of this form because any directed acyclic graph can be made into levels or layers with the first layer comprising input vertices and the final layer comprising only output vertices (Harary, Norman, & Cartwright, 1965). We adopt the following rearrangement: the layer assigned to a given neuron is the length of the longest path, where length is defined as the total number of synapses passed from the input layer. Additional ghost units that do not change their inputs (with identity gain function $g(u) = u$ and output weight $= 1$) can then be added to eliminate projections that skip layers.

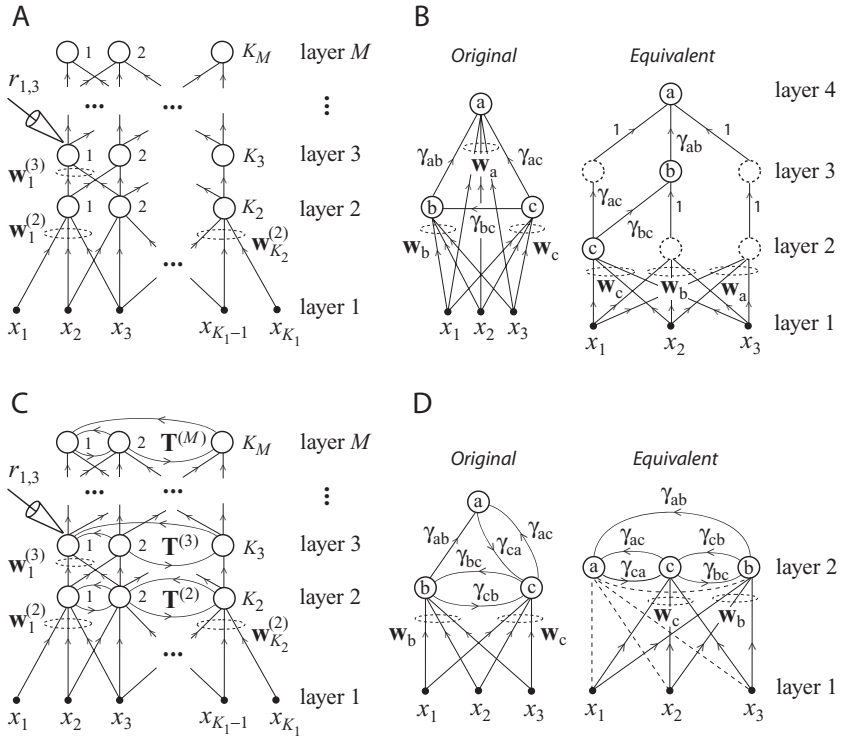


Figure 3: Neural network architectures considered in this letter. Dashed ellipses specify the weight vectors, and the cones indicate recording electrodes. **(A)** A generic layered feedforward network here is a multilayer perceptron. **(B)** Any network without loops (left) can be transformed into an equivalent layered feedforward network (right). Dashed circles indicate "ghost units" or input pass-through. This example comes from a proposed model of functional circuitry in the dorsal cochlear nucleus in the cat auditory system (Young, 1998). **(C)** The general form of a layered recurrent network allows arbitrary lateral connections among neurons in the same layer, but only feedforward connections between successive layers. **(D)** An arbitrary recurrent network (left) is equivalent to a layered recurrent network with two layers (right). Dashed lines indicate zero connections.

This trick can also ensure that output neurons are located only in the final layer. Figure 3B shows an example of rearranging a network without loops into an equivalent layered network.

**3.3 Recurrent Network.** The general structure of a layered recurrent network considered in this letter is shown in Figure 3C. Compared with

the feedforward network in Figure 3A, the feedforward connections are organized in the same way, but now arbitrary recurrent connections are added to neurons within the same layer. Although no feedback across layers is allowed in this layered network, any network with feedback can always be rearranged into a layered network consisting of only two layers: a recurrent layer (layer 2) with arbitrary lateral connections and an input layer (layer 1), which can potentially connect with any neuron in layer 2. Figure 3D shows an example of rearranging a network with feedback into an equivalent layered network.

We use standard continuous firing rate dynamics (Wilson & Cowan, 1972; Amari, 1972; Hopfield, 1984). The firing rates of neurons in layer $m$ ($2 \leq m \leq M$) are described by a vector

$$\mathbf{r}_m = \mathbf{g}_m(\mathbf{v}_m) = \begin{bmatrix} g_{1,m}(v_{1,m}) \\ g_{2,m}(v_{2,m}) \\ \vdots \\ g_{K_m,m}(v_{K_m,m}) \end{bmatrix}, \tag{3.6}$$

where the net input $\mathbf{v}_m$ is interpreted as the membrane potentials or the synaptic currents, and $\mathbf{g}_m$ is the vector form of the gain functions. The vector-matrix form of the dynamical equation reads

$$\frac{d\mathbf{v}_m}{dt} = -\mathbf{D}^{(m)}\mathbf{v}_m + \mathbf{T}^{(m)}\mathbf{g}_m(\mathbf{v}_m) + \mathbf{W}^{(m)}\mathbf{r}_{m-1}, \tag{3.7}$$

where $\mathbf{D}^{(m)}$ is a diagonal matrix with positive entries, $\mathbf{T}^{(m)}$ is a $K_m \times K_m$ matrix for the strength of recurrent connections among neurons in layer $m$, and $\mathbf{W}^{(m)}\mathbf{r}_{m-1}$ is the input from layer $m-1$ for $2 \leq m \leq M$. Similar to equation 3.5 for the feedforward network, here the firing rates for the first layer are the stimulus input: $\mathbf{r}_1 = \mathbf{x}$.

## 4 Feedforward Network Results

**4.1 Basic Results.** To examine how the activity of any layer $m$ ($2 \leq m \leq M$) depends on the activity of the previous layer $m-1$ in a feedforward network, consider the differential formula derived from equation 3.4:

$$d\mathbf{r}_m = \mathbf{J}^{(m)} d\mathbf{r}_{m-1}, \tag{4.1}$$

where the Jacobian matrix is

$$\mathbf{J}^{(m)} = \mathbf{G}^{(m)}\mathbf{W}^{(m)}, \tag{4.2}$$

where

$$\mathbf{G}^{(m)} = \begin{bmatrix} g'_{1,m}\left(\mathbf{w}_1^{(m)} \cdot \mathbf{r}_{m-1}\right) & 0 & \cdots & 0 \\ 0 & g'_{2,m}\left(\mathbf{w}_2^{(m)} \cdot \mathbf{r}_{m-1}\right) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & g'_{K_m,m}\left(\mathbf{w}_{K_m}^{(m)} \cdot \mathbf{r}_{m-1}\right) \end{bmatrix} \tag{4.3}$$

is a diagonal matrix of the derivatives of the gain functions. Applying equation 4.1. recursively all the way down the lower layers, we have

$$d\mathbf{r}_m = \mathbf{J}^{(m)}\mathbf{J}^{(m-1)} \ldots \mathbf{J}^{(2)} d\mathbf{x}, \tag{4.4}$$

where in the last step, $\mathbf{r}_1 = \mathbf{x}$ is used (see equation 3.5). As a simple application of the chain rule for derivatives, this result is reminiscent of the derivation of the backpropagation learning rule, although the latter involves derivatives with respect to the weights rather than the stimulus (Rumelhart et al., 1986). We can also write the derivatives equivalently as the Jacobian:

$$\frac{\partial \mathbf{r}_m}{\partial \mathbf{x}} = \mathbf{J}^{(m)}\mathbf{J}^{(m-1)} \ldots \mathbf{J}^{(2)}. \tag{4.5}$$

Equations 4.4 and 4.5 show how the activity $\mathbf{r}_m$ of layer $m$ is affected by infinitesimal changes of the input $\mathbf{x}$.

Recall that the *rank* of a matrix is the maximum number of linearly independent rows or columns. An $m \times n$ matrix $\mathbf{A}$ is called *full rank* or *nondegenerate* if

$$\operatorname{rank} \mathbf{A} = \min\{m, n\}, \tag{4.6}$$

which is the maximum rank possible. If the rank is less than the maximum possible, we call the matrix *degenerate*. Let $\mathbf{A}$ be an $m \times n$ matrix and $\mathbf{B}$ be an $n \times k$ matrix. If $\operatorname{rank} \mathbf{B} = n$, then

$$\operatorname{rank} \mathbf{AB} = \operatorname{rank} \mathbf{A}. \tag{4.7}$$

Similarly, if $\operatorname{rank} \mathbf{A} = n$, then $\operatorname{rank} \mathbf{AB} = \operatorname{rank} \mathbf{B}$, which can be applied to the Jacobian matrix in equation 4.2 to obtain

$$\operatorname{rank} \mathbf{J}^{(m)} = \operatorname{rank} \mathbf{W}^{(m)}, \tag{4.8}$$

assuming $g'_{i,m} \neq 0$ so that the diagonal matrix $\mathbf{G}^{(m)}$ in equation 4.3 is nondegenerate.

We present the first result as a theorem:

**Theorem 1.** *Suppose a feedforward neural network satisfies the following conditions:*

1. *Each layer contains no more neurons than the layer below it.*
2. *The weight matrices connecting successive layers have full rank.*
3. *All gain functions are continuously differentiable with positive derivatives.*

*Then given any compact stimulus set comprising an interior and a boundary, the response of any neuron in the network can attain a maximum or minimum only at the boundary of the stimulus set but never at its interior.*

Before proving the theorem, we express the three conditions more explicitly. Condition 1 means that

$$K_1 \geq K_2 \geq \cdots \geq K_M, \tag{4.9}$$

where $K_m$ is the number of neurons in layer $m$, assuming there is a total of $M$ layers (see Figure 3A). Condition 2 means that

$$\operatorname{rank} \mathbf{W}^{(m)} = \min\{K_m, K_{m-1}\} = K_m, \tag{4.10}$$

which follows from equations 4.6 and 4.9. Here $\mathbf{W}^{(m)}$ is a $K_m \times K_{m-1}$ matrix that specifies the synaptic weights from neurons in layer $m-1$ to neurons in layer $m$. Finally, condition 3 means that $g'_{i,m}(u) > 0$ for gain function $g_{i,m}$ of neuron $i$ in layer $m$.

**Proof.** Consider the response $r_{i,m}$ of a single neuron $i$ in layer $m$. A necessary condition for $r_{i,m}$ to attain a maximum (or minimum) in the interior of the stimulus set is that the gradient with respect to the stimulus $\mathbf{x}$ must vanish, because otherwise one can always change the stimulus along the gradient to further increase (or decrease) the response. Writing this necessary condition using equations 4.2, 4.3 and 4.5 and taking the $i$th row of $\mathbf{J}^{(m)}$, we have

$$\frac{\partial r_{i,m}}{\partial \mathbf{x}} = g'_{i,m}\big(\mathbf{w}_i^{(m)} \cdot \mathbf{r}_{m-1}\big)\mathbf{w}_i^{(m)}\mathbf{J}^{(m-1)} \ldots \mathbf{J}^{(2)} = \mathbf{0}, \tag{4.11}$$

which is equivalent to

$$\mathbf{w}_i^{(m)}\big(\mathbf{J}^{(m-1)} \ldots \mathbf{J}^{(2)}\big) = \mathbf{0} \tag{4.12}$$

because $g'_{i,m} > 0$. Since the product of the Jacobian matrices in the parentheses has full rank (see below), we must have the row vector $\mathbf{w}_i^{(m)} = \mathbf{0}$ as required by linear independence. This is a contradiction because it implies that neuron $i$ in layer $m$ receives no synaptic input at all, contradicting our assumption that the weight matrix $\mathbf{W}^{(m)}$ has full rank. The contradiction establishes that a maximum (or minimum) cannot be attained at any interior point. Since a continuous function always has a maximum and a minimum on a compact set, those can be attained only at the boundary.

The only thing left to show is that the product of Jacobian matrices in equation 4.12 indeed has full rank. Note that every $\mathbf{J}^{(n)}$ has full rank or $\operatorname{rank}\mathbf{J}^{(n)} = K_n$, which follows from equations 4.8 and 4.10. Applying equation 4.7 repeatedly, we obtain

$$\operatorname{rank}\left(\mathbf{J}^{(n)}\ldots\mathbf{J}^{(3)}\mathbf{J}^{(2)}\right) = \operatorname{rank}\left(\mathbf{J}^{(n)}\ldots\mathbf{J}^{(3)}\right) = \cdots = \operatorname{rank}\mathbf{J}^{(n)} = K_n \quad (4.13)$$

for $2 \leq n \leq M$. This verifies that the product in equation 4.12 has full rank.

This theorem holds true not only for every neuron in the top layer, but also for every neuron in the other layers throughout the entire network, including all the hidden units in the intermediate layers. Condition 1 for convergent connection hierarchy and condition 2 for nondegenerate weight matrices are sufficient for theorem 1 to hold true, but none of them is necessary. For example, if all connections in a network are excitatory (or inhibitory), then any increase of the input will lead to an increase (or decrease) of response for all neurons throughout the network. The network in Figure 1B is an example with all positive weights. In such networks, both the maximum and the minimum responses can be attained only by boundary stimuli regardless of whether conditions 1 and 2 are satisfied. In other words, the conclusions of the theorem may hold true for some networks that have divergent layers or degenerate weight matrices. But if either condition 1 or condition 2 is removed, one can find counterexamples that defeat the theorem. Therefore, in general, the two conditions cannot be relaxed, although they are unnecessary for some networks.

Condition 1 requires that each layer contain no more neurons than a lower layer. A network that does not satisfy this condition is shown in Figure 1C. The theorem does not apply to this network. In general, condition 1 is not as restrictive as it first may appear because when one applies the theorem to a neuron of interest, one needs to include only lower-layer neurons that actually affect its activity. The resultant effective network can satisfy condition 1 even when the whole network does not. We return to this issue in section 6.2.

**4.2 Stimulus Perturbation for Response Control.** The network considered in the preceding section has another important property: the activities

of all the neurons in any given layer of the network can be controlled simultaneously and independently by the stimulus. To see this, first consider a local linear approximation of equation 4.4

$$\Delta \mathbf{r}_m \approx \mathbf{J} \Delta \mathbf{x}, \tag{4.14}$$

where $\Delta \mathbf{r}_m$ is the change of the response caused by a small but finite change $\Delta \mathbf{x}$ of the stimulus, and

$$\mathbf{J} \equiv \mathbf{J}^{(m)} \mathbf{J}^{(m-1)} \dots \mathbf{J}^{(2)} \tag{4.15}$$

is the overall Jacobian matrix. Given any desired change $\Delta \mathbf{r}_m$ of the response pattern, how can it be attained by a proper stimulus change $\Delta \mathbf{x}$? One can solve for $\Delta \mathbf{x}$ from equation 4.14 provided that $\mathbf{J}$ has full rank, and the standard solution is

$$\Delta \mathbf{x} \approx \mathbf{J}^{\dagger} \Delta \mathbf{r}_m + (\mathbf{I} - \mathbf{J}^{\dagger} \mathbf{J}) \mathbf{z}, \tag{4.16}$$

where $\mathbf{J}^{\dagger}$ is Moore-Penrose pseudoinverse, $\mathbf{I}$ is an identity matrix, and vector $\mathbf{z}$ is arbitrary so that the solution in general is not unique (Ben-Israel & Greville, 2003). Here the Jacobian $\mathbf{J}$ needs to have full rank because otherwise any response vector $\Delta \mathbf{r}_m$ lying outside the range space of $\mathbf{J}$ can never be elicited by any stimuli.

To test whether the overall Jacobian $\mathbf{J}$ has full rank in an experiment, one may use $K_1$ stimuli to elicit $K_1$ responses and then write the approximate equation 4.14 in the matrix form $\mathbf{R} \approx \mathbf{J}\mathbf{X}$. Here $\mathbf{X}$ is a $K_1 \times K_1$ square matrix, with each column being a single stimulus vector $\Delta \mathbf{x}$, and different stimulus vectors are chosen to be linearly independent; $\mathbf{R}$ is a $K_m \times K_1$ response matrix, with each column being a response vector $\Delta \mathbf{r}_m$, which should be averaged over repeated trials to reduce noise. Then

$$\mathbf{J} \approx \mathbf{R}\mathbf{X}^{-1} \tag{4.17}$$

is an estimate of the Jacobian from the experimental data. Here the rank of the Jacobian $\mathbf{J}$ is always equal to the rank of the response matrix $\mathbf{R}$, or $\operatorname{rank} \mathbf{J} = \operatorname{rank} \mathbf{R}$.

The argument above shows how to find a stimulus to generate any desired change of response pattern by linear approximation. We prove below that an exact inverse function actually exists so that the desired response can be generated by a proper stimulus exactly, not just approximately.

**Theorem 2.** *Suppose a feedforward neural network satisfies the same three conditions in theorem 1 and the set of allowable stimuli is compact with an interior and a boundary. Then the following statements hold:*

*a. Given any interior stimulus and the corresponding pattern of responses in any layer of the network, any arbitrary but sufficiently small changes to the response pattern can be produced exactly by some stimulus.*
*b. All boundary response patterns can arise only from boundary stimuli. That is, every response pattern in the topological boundary of the set of all possible responses in any given layer of the network must come from some stimulus on the boundary of the stimulus set.*

**Proof.** Let the response vector $\mathbf{r}_m$ of neurons in any given layer $m$ be

$$\mathbf{r}_m = \mathbf{f}(\mathbf{x}), \tag{4.18}$$

where $\mathbf{x}$ is the stimulus vector and function $\mathbf{f}$ is continuously differentiable since it is built by linear combinations and compositions of continuously differentiable gain functions. Consider the same Jacobian matrix,

$$\frac{\partial \mathbf{r}_m}{\partial \mathbf{x}} = \mathbf{J}, \tag{4.19}$$

as in equation 4.5, using the notation of equation 4.15. We have rank $\mathbf{J} = K_m$ (full rank) according to the argument for equation 4.13. By assumption, the stimulus vector $\mathbf{x}$ contains $K_1$ variables, whereas the response vector $\mathbf{r}_m$ contains $K_m$ variables, which may be fewer because $K_1 \geq K_m$ according to equation 4.9. A unique inverse function of $\mathbf{f}$ does not exist when $K_1 > K_m$ because different stimuli may generate identical responses. However, it is always possible to find a unique inverse function in a subspace of dimension $K_m$ for the stimulus parameters according to the rank theorem, a variant of the implicit function theorem (Spivak, 1965; Rudin, 1976). A more explicit explanation is as follows. To specify a unique inverse function, first select any subset of $K_m$ stimulus variables from a total of $K_1$ to define a restricted stimulus vector $\tilde{\mathbf{x}}$ such that the new Jacobian $\partial \mathbf{r}_m / \partial \tilde{\mathbf{x}}$ is an invertible $K_m \times K_m$ matrix. Assuming from now on that all remaining $K_1 - K_m$ stimulus variables are always fixed, we can regard the response $\mathbf{r}_m$ as a function of the selected stimulus variables $\tilde{\mathbf{x}}$ only, namely,

$$\mathbf{r}_m = \tilde{\mathbf{f}}(\tilde{\mathbf{x}}). \tag{4.20}$$

The theorem follows from the fact that $\tilde{\mathbf{f}}$ is locally invertible, as detailed below.

*Proof of statement (a):* Given any interior stimulus $\mathbf{x}^0$ together with its response $\mathbf{r}_m^0 = \mathbf{f}(\mathbf{x}^0)$, consider the restricted stimulus vector $\tilde{\mathbf{x}}^0$ and the function $\tilde{\mathbf{f}}$ as in equation 4.20. Since the Jacobian $[\partial \mathbf{r}_m / \partial \tilde{\mathbf{x}}]_{\tilde{\mathbf{x}}=\tilde{\mathbf{x}}^0}$ is invertible as described above, we can apply the inverse function theorem (Rudin, 1976) to find an open subset $U$ in the restricted stimulus space and an open subset

$V = \tilde{\mathbf{f}}(U)$ in the response space such that $\tilde{\mathbf{x}}^0 \in U$ and $\mathbf{r}_m^0 \in V$, and the mapping between $U$ and $V$ is one-to-one and the inverse $\tilde{\mathbf{f}}^{-1}$ is continuously differentiable. In other words, pick any desired response pattern $\mathbf{r}_m \in V$, and the inverse function can always find a stimulus,

$$\tilde{\mathbf{x}} = \tilde{\mathbf{f}}^{-1}(\mathbf{r}_m) \in U, \tag{4.21}$$

which should elicit the desired response exactly. This proves statement (a).

*Proof of statement (b):* The inverse function theorem implies that the response $\mathbf{r}_m^0 = \mathbf{f}(\mathbf{x}^0)$ elicited by the interior stimulus $\mathbf{x}^0$ is contained by the open set $V$, or $\mathbf{r}_m^0 \in V$. Since every point in an open set is an interior point, the response $\mathbf{r}_m^0$ is an interior point. This shows that an interior stimulus can elicit only an interior response. In other words, a boundary response cannot arise from an interior stimulus, and the only possibility left is its arising from a boundary stimulus.

Statement (a) shows that one can find proper stimulus to perturb, simultaneously and arbitrarily, the activities of all neurons in any single layer of the network. That is, pick any layer of the network; then all the neurons in that layer are simultaneously and independently controllable. However, in general, one may not be able to control neurons from two different layers at the same time, although each layer is completely controllable by itself. Another caveat is that the size of the desired change of responses, although finite, should be sufficiently small. The proof involving the inverse function theorem is an existence proof that does not specify the function explicitly. To estimate how small is "sufficiently small," one has to impose additional constraints on the network, such as on the maximum slope of the gain functions and the maximum synaptic weights.

As an example of simultaneous control, theorem 2 implies that it is possible to alter the response of any given single neuron up or down while keeping the responses of all other neurons in the same layer constant. Perturbing the activity of a single neuron is just a special case of the arbitrary changes to the activity pattern as described in statement (a) of theorem 2. Finally, theorem 2 may be extended to allow the Jacobian in equation 4.19 to be degenerate rather than full rank, in which case the number of neurons in a given layer that can be controlled simultaneously is equal to the rank of the Jacobian (see section 6.1).

**4.3 Topological Interior and Boundary of Stimuli and Responses.** Statement (b) of theorem 2 needs further explanation. As illustrated by Figure 4, the maximum and minimum responses of a neuron are special points belonging to the boundary of the set of all possible responses, and therefore they must arise from boundary stimuli by statement (b). This argument may serve as another proof of theorem 1. Statement (b) of theorem 2

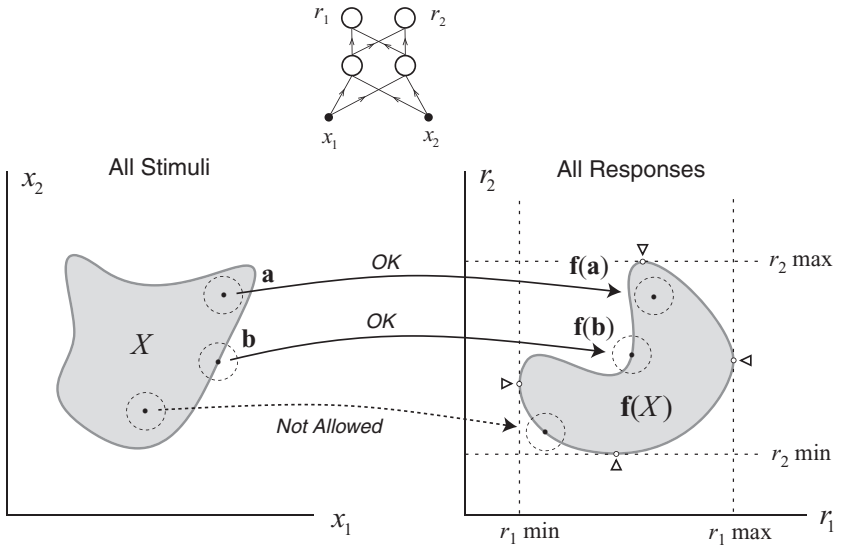


Figure 4: Stimulus-response topology according to theorem 2. Although a network with only two inputs and two outputs is illustrated here for clarity, the general conclusions are valid for arbitrary dimensions. The arrow $\mathbf{a} \rightarrow \mathbf{f}(\mathbf{a})$ shows that an interior stimulus (i.e., in the interior of the set of all stimuli) always elicits an interior response (i.e., in the interior of the set of all responses). The arrow $\mathbf{b} \rightarrow \mathbf{f}(\mathbf{b})$ and the dashed arrow below show that a boundary response (i.e., on the boundary of the set of all responses) can arise only from a boundary stimulus (i.e., on the boundary of the set of all stimuli), never from an interior stimulus. The maximum and minimum responses of each neuron (open arrow heads) are special cases of boundary responses, and as such, they must arise from some boundary stimuli.

is more general than theorem 1: while theorem 1 ensures only that the maximum and minimum responses of each individual neuron must originate from the stimulus boundary, theorem 2 asserts that in fact, the entire boundary of the pattern of responses in any given layer must arise also from the stimulus boundary.

Under the conditions assumed in the theorems above, an interior stimulus always elicits an interior response but never a boundary response. By contrast, a boundary stimulus may elicit either a boundary response or an interior response, depending on the details of the network. These general results are summarized in Table 1.

To see all these possibilities intuitively, consider a translucent balloon casting its shadow on the ground. A point inside the balloon is always cast to the interior of the shadow, whereas a point on the surface of the balloon

Table 1: Stimulus-Response Topology

| | *Interior Response* | *Boundary Response* |
|---|---|---|
| *Interior stimulus* | Always | Never |
| *Boundary stimulus* | Permissible | Permissible |

may end up either in the interior of the shadow or on its boundary. One can readily construct a feedforward network that is analogous to this example by taking the balloon as the stimulus set and the shadow as the response set.

**4.4 Gain Functions with Flat Segments.** The results in the preceding sections assume positive derivatives of the gain functions. Biological gain functions may have flat segments such as when input is below threshold. The vanishing derivative in the flat segment can induce degeneracy in the Jacobian matrix, making the treatment in the preceding section inadequate. The following theorem shows that with flat segments in gain functions, the maximum (or minimum) response is attained still only by boundary stimuli, but now some interior stimuli may elicit an identical response. Nonetheless, a maximum (or minimum) response at the interior is not a true peak (or valley) in the stimulus-response relation but a flat plateau along some dimension.

**Theorem 3.** *Suppose a feedforward neural network satisfies conditions 1 and 2 in theorem 1, but the gain functions have nonnegative derivatives, allowing zero derivatives for continuous flat segments. Then for any stimulus set that is compact with an interior and a boundary, the following statements hold:*

a. *The response of any neuron in the network to any interior stimuli can never be greater (or less) than the maximum (or minimum) response to boundary stimuli.*
b. *For any neuron in the network, if an interior stimulus elicits a response that matches the maximum or minimum boundary responses, then it is possible to perturb the stimulus continuously without altering the response. In other words, a strict peak or valley in stimulus-response relation cannot exist.*

**Proof.** We treat the gain functions with flat segments as the limit of a series of strictly increasing functions so as to apply theorem 1. Define a series of strictly increasing gain functions,

$$g_{i,m,N}(u) = g_{i,m}(u) + u/N, \tag{4.22}$$

for each neuron $i$ in layer $m$, with $N = 1, 2, \ldots$. We have $g'_{i,m,N}(u) > 0$ because $g'_{i,m}(u) \geq 0$ for the original gain function, and $g_{i,m,N}(u) \to g_{i,m}(u)$ as $N \to \infty$.

Now consider a series of networks ($N = 1, 2, \ldots$) with everything identical to that of the original network except that the gain functions $g_{i,m}$ are replaced by $g_{i,m,N}$. Let $f(\mathbf{x})$ be the response to stimulus $\mathbf{x}$ by any given neuron in the original network, and let $f_N(\mathbf{x})$ be the response of the same neuron in the network series. We claim that

$$\lim_{N\to\infty} f_N(\mathbf{x}) = f(\mathbf{x}), \tag{4.23}$$

and the convergence is uniform for all stimuli $\mathbf{x} \in X$, with $X$ being the stimulus set. The uniform convergence property is needed to prove that the maximum response of the original network is the limit of the maximum responses of the network series since the maxima of the series may be attained at different locations on the stimulus boundary.

To show the uniform convergence of equation 4.23, replace $1/N$ in $f_N(\mathbf{x})$ by a continuous variable $\zeta \in [0, 1]$ and write $F(\mathbf{x}, \zeta) = f_N(\mathbf{x})$ for $\zeta > 0$. The case $\zeta = 0$ corresponds to the original network, and $F(\mathbf{x}, 0) = f(\mathbf{x})$. Now we have a continuous function $F(\mathbf{x}, \zeta)$ defined on the product space $X \times [0, 1]$, which is compact because $X$ is compact. The function $F(\mathbf{x}, \zeta)$ is continuous because it is made of linear combinations and recursive compositions of gains functions that are always continuous (see equations 3.3 and 4.22). Since any continuous function on a compact set is uniformly continuous (Heine-Cantor theorem), $F(\mathbf{x}, \zeta)$ is uniformly continuous. This means that for any given $\varepsilon > 0$, there exists a $\delta$ such that

$$\sqrt{\|\mathbf{x}_1 - \mathbf{x}_2\|^2 + |\zeta_1 - \zeta_2|^2} < \delta \tag{4.24}$$

implies

$$|F(\mathbf{x}_1, \zeta_1) - F(\mathbf{x}_2, \zeta_2)| < \varepsilon \tag{4.25}$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in X$, and $\zeta_1, \zeta_2 \in [0, 1]$, with $\|\|$ indicating Euclidean distance. In particular, setting $\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{x}$, $\zeta_2 = 0$, and $\zeta_1 = 1/N$, we have $F(\mathbf{x}_1, \zeta_1) = f_N(\mathbf{x})$ and $F(\mathbf{x}_2, \zeta_2) = f(\mathbf{x})$. Then the statement that equation 4.24 implies equation 4.25 becomes that $N > 1/\delta$ implies

$$|f_N(\mathbf{x}) - f(\mathbf{x})| < \varepsilon \tag{4.26}$$

for all $\mathbf{x} \in X$. This verifies that the convergence in equation 4.23 is uniform.

Since a continuous function always has a maximum on a compact set (Weierstrass theorem), we suppose that among all boundary stimuli $\partial X$, the stimulus $\mathbf{b} \in \partial X$ elicits the maximum response $f(\mathbf{b})$ in the original network, whereas the stimulus $\mathbf{b}_N \in \partial X$ elicits the maximum response $f_N(\mathbf{b}_N)$ in the network series. Here $\partial X \subset X$ denotes the boundary of $X$. Next we show

that equation 4.26 implies

$$|f_N(\mathbf{b}_N) - f(\mathbf{b})| < \varepsilon. \tag{4.27}$$

Split equation 4.26 into two inequalities:

$$f_N(\mathbf{x}) < f(\mathbf{x}) + \varepsilon, \tag{4.28}$$

$$f_N(\mathbf{x}) > f(\mathbf{x}) - \varepsilon. \tag{4.29}$$

Since $f(\mathbf{b})$ is a maximum boundary response, we have $f(\mathbf{x}) \leq f(\mathbf{b})$ for any $\mathbf{x} \in \partial X$. This inequality together with equation 4.28 implies

$$f_N(\mathbf{b}_N) < f(\mathbf{b}) + \varepsilon \tag{4.30}$$

by putting $\mathbf{x} = \mathbf{b}_N$. Since the maximum response $f_N(\mathbf{b}_N) \geq f_N(\mathbf{x})$ for any $\mathbf{x}$ by theorem 1, combining this inequality with equation 4.29 and then setting $\mathbf{x} = \mathbf{b}$, we obtain

$$f_N(\mathbf{b}_N) > f(\mathbf{b}) - \varepsilon. \tag{4.31}$$

Combining equations 4.30 and 4.31 yields equation 4.27, and the arbitrariness of $\varepsilon$ proves that

$$\lim_{N\to\infty} f_N(\mathbf{b}_N) = f(\mathbf{b}). \tag{4.32}$$

Now for any given interior stimulus $\mathbf{a}$, applying theorem 1 to the network series yields

$$f_N(\mathbf{a}) < f_N(\mathbf{b}_N). \tag{4.33}$$

Taking the limit $N \to \infty$ on both sides of equation 4.33 and then using limits 4.23 and 4.32, we finally obtain

$$f(\mathbf{a}) \leq f(\mathbf{b}). \tag{4.34}$$

This is equivalent to statement (a) for the maximum case. The proof for the minimum case is analogous.

*Proof of statement (b):* The neuron of interest must receive input (possibly indirectly) from at least one lower-level neuron whose input lies in the flat segment of its gain function, because otherwise theorem 1 would apply and contradict the assumption about the match. Let $m$ be lowest layer with a neuron in the flat segment. Then everything below layer $m$ satisfies the conditions for theorem 2, and we can find a proper stimulus to arbitrarily

perturb the inputs to layer $m$ using an inverse function. The only caveat here is that the original statement of theorem 2 is about the response rather than the input, but the conclusion can be readily extended to input by taking the gain functions in layer $m$ as the identity function $g_{i,m}(u) = u$. Now slightly increase or decrease the input to the selected neuron in layer $m$ while keeping it always within the flat segment, but without changing the inputs to any other neuron in the same layer. The neuron of interest, sitting at a layer higher than $m$, should maintain the same response because the outputs from layer $m$ never change while the stimulus is being altered.

## 5 Recurrent Network Results

We consider recurrent networks that exhibit global asymptotic stability, which implies that given any stimulus, the network always settles into the same response pattern or final equilibrium state, which is determined by the stimulus only regardless of the initial state of the network (Bhatia & Szegö, 1970). This assumption seems reasonable for sensory neurons in many neurophysiological studies, and it allows us to derive results parallel with those for the feedforward networks, but it ignores intrinsically oscillatory networks (Wilson & Cowan, 1972) such as those in the olfactory system (Laurent et al., 2001).

First, we show that when confined to equilibrium states, a globally stable recurrent network may resemble the feedforward network considered before. Whenever the network described by dynamical equation 3.7 reaches an equilibrium state, we have $d\mathbf{v}_m/dt = \mathbf{0}$, and the algebraic equation,

$$\mathbf{F}(\mathbf{u}_m) \equiv \mathbf{D}^{(m)}\mathbf{u}_m - \mathbf{T}^{(m)}\mathbf{g}_m(\mathbf{u}_m) - \mathbf{W}^{(m)}\mathbf{s}_{m-1} = \mathbf{0}, \tag{5.1}$$

for $2 \leq m \leq M$. Global stability implies that this equation has a unique solution of $\mathbf{u}_m$ for each given input $\mathbf{s}_{m-1}$. For clarity, instead of $\mathbf{v}_m$, we have used different variables $\mathbf{u}_m = (u_{1,m}, u_{2,m}, \ldots, u_{K_m,m})^{\mathrm{T}}$ to denote the equilibrium state, and

$$\mathbf{s}_{m-1} = \mathbf{g}_{m-1}(\mathbf{u}_{m-1}) \tag{5.2}$$

is the counterpart of equation 3.6 for equilibrium state in layer $m-1$ ($2 \leq m \leq M$). The first layer is the stimulus input:

$$\mathbf{s}_1 = \mathbf{x}. \tag{5.3}$$

To ensure that equation 5.2 makes sense for the first layer, we define $\mathbf{u}_1 = \mathbf{x}$ and let the gain functions $\mathbf{g}_1 = (g_{1,1}, g_{2,1}, \ldots, g_{K_1,1})$ be the identity functions, namely, $g_{i,1}(u) = u$. According to the implicit function theorem (Rudin, 1976; Krantz & Parks, 2002), equation 5.1 defines the equilibrium

state $\mathbf{u}_m$ locally as an implicit function of the activity $\mathbf{s}_{m-1}$ from the lower layer $m-1$, provided that the matrix

$$\mathbf{M}^{(m)}(\mathbf{u}_m) \equiv \frac{\partial \mathbf{F}(\mathbf{u}_m)}{\partial \mathbf{u}_m} = \mathbf{D}^{(m)} - \mathbf{T}^{(m)}\mathbf{G}^{(m)}(\mathbf{u}_m) \tag{5.4}$$

is invertible, where

$$\mathbf{G}^{(m)}(\mathbf{u}_m) = \mathrm{diag}\left(g'_{1,m}(u_{1,m}), g'_{2,m}(u_{2,m}), \ldots, g'_{K_m,m}(u_{K_m,m})\right) \tag{5.5}$$

is a diagonal matrix of the gain function derivatives. The matrix $\mathbf{M}^{(m)}(\mathbf{u}_m)$ is guaranteed to be invertible if the linearized system is nondegenerate. This is because the linearized version of equation 3.7 for layer $m$ around the equilibrium point $\mathbf{u}_m$ is actually

$$\frac{d\mathbf{z}_m}{dt} = -\mathbf{M}^{(m)}(\mathbf{u}_m)\mathbf{z}_m, \tag{5.6}$$

where $\mathbf{z}_m = \mathbf{v}_m - \mathbf{u}_m$, and the input term from layer $m-1$ disappears because it is assumed to be fixed. Global stability implies that here, the eigenvalue of $-\mathbf{M}^{(m)}(\mathbf{u}_m)$ cannot have positive real part, although it is still possible to have 0 as an eigenvalue and therefore a degenerate $\mathbf{M}^{(m)}(\mathbf{u}_m)$ (see appendix A).

By focusing on how the equilibrium state depends on the input, we can treat the recurrent network in a manner similar to the feedforward network as long as $\mathbf{M}^{(m)}(\mathbf{u}_m)$ is invertible. Write the differential form of equation 5.1 as

$$d\mathbf{u}_m = \mathbf{M}^{(m)}(\mathbf{u}_m)^{-1}\mathbf{W}^{(m)}d\mathbf{s}_{m-1}. \tag{5.7}$$

Combining this with $d\mathbf{s}_{m-1} = \mathbf{G}^{(m-1)}(\mathbf{u}_{m-1})d\mathbf{u}_{m-1}$, the differential form of equation 5.2, we obtain the recursive relation

$$d\mathbf{u}_m = \mathbf{J}^{(m)}d\mathbf{u}_{m-1} \tag{5.8}$$

where

$$\mathbf{J}^{(m)} = \mathbf{M}^{(m)}(\mathbf{u}_m)^{-1}\mathbf{W}^{(m)}\mathbf{G}^{(m-1)}(\mathbf{u}_{m-1}) \tag{5.9}$$

is analogous to the Jacobian in the treatment of the feedforward network. Applying equation 5.8 repeatedly all the way down to layer 1 with equation 5.3, we have

$$d\mathbf{u}_m = \mathbf{J}^{(m)}\mathbf{J}^{(m-1)}\ldots\mathbf{J}^{(2)}d\mathbf{x}, \tag{5.10}$$

which is the counterpart of equation 4.4 for the feedforward network, and

$$\frac{\partial \mathbf{u}_m}{\partial \mathbf{x}} = \mathbf{J}^{(m)}\mathbf{J}^{(m-1)}\ldots\mathbf{J}^{(2)} \tag{5.11}$$

is the counterpart of equation 4.5.

In summary, equations 5.10 and 5.11 mean that the equilibrium state $\mathbf{u}_m$ of the recurrent network plays the same formal role as the response $\mathbf{r}_m$ in the feedforward network. As before, assuming strictly increasing gain functions, now if the weight matrix $\mathbf{W}^{(m)}$ has full rank, then $\mathbf{J}^{(m)}$ also has full rank and rank $\mathbf{J}^{(m)} = \text{rank}\,\mathbf{W}^{(m)}$, which is the same as equation 4.8. Equation 4.13 also still holds:

$$\text{rank}\left(\mathbf{J}^{(m)}\ldots\mathbf{J}^{(3)}\mathbf{J}^{(2)}\right) = \text{rank}\left(\mathbf{J}^{(m)}\ldots\mathbf{J}^{(3)}\right) = \cdots = \text{rank}\,\mathbf{J}^{(m)} = K_m. \tag{5.12}$$

Now we can state an analogous set of results for the recurrent network. The proof is similar to that of the feedforward counterpart so that details will be omitted.

**Theorem 4.** *Suppose a hierarchical neural network has both feedforward connections and recurrent connections within each layer, and satisfies the following conditions:*

1. *Each layer contains no more neurons than the layer below it.*
2. *The weight matrices for feedforward connections between successive layers have full rank.*
3. *All gain functions are continuously differentiable with positive derivatives.*
4. *For any given stimulus, the network state always approaches the same response regardless of the initial state, and the linearized system is also always stable.*

*Then for any stimulus set that is compact with an interior and a boundary, the following statements hold:*

a. *The response of any neuron in the network can attain a maximum or minimum only at the boundary of the stimulus set, never at its interior.*
b. *Given any interior stimulus and the corresponding pattern of responses in any layer of the network, any arbitrary but sufficiently small changes to the response pattern can be generated exactly by some stimuli.*
c. *All boundary responses in any given layer can arise only from boundary stimuli.*

*Here "response" always refers to the equilibrium state of the network under a fixed stimulus.*

The statements of this theorem are identical to those in theorems 1 and 2 for feedforward networks. Conditions 1 to 3 are identical to those in theorem 1 except that the weight matrices in condition 2 are now explicitly identified as the feedforward connections. The recurrent connections are constrained implicitly by condition 4 on global stability. Since a globally stable nonlinear system may sometimes allow a degenerate linearized system (see appendix A), the stability of the linearized system is required explicitly here. The stability of the linearized system, 5.6, implies that the eigenvalues of matrix $-\mathbf{M}^{(m)}(\mathbf{u}_m)$ can have only negative real parts, which in turn implies the invertibility of this matrix.

The explicit requirement for a stable linearized system is no longer needed if a sufficient condition for global stability can automatically guarantee the stability of the linearized system. The following example is a generic two-layer recurrent network similar to that shown in Figure 2D. Since only the second layer ($m = 2$) needs to be considered, we simply drop the index $m$ in equation 5.4 and write

$$\mathbf{M}(\mathbf{u}) \equiv \mathbf{D} - \mathbf{T}\mathbf{G}(\mathbf{u}), \tag{5.13}$$

where $\mathbf{G}(\mathbf{u}) = \mathrm{diag}(g_1'(u_1), g_2'(u_2), \ldots, g_{K_2}'(u_{K_2}))$ is the simplified version of equation 5.5. One sufficient condition for global stability is that the symmetric matrix

$$\mathbf{S} \equiv \mathbf{D}\bar{\mathbf{G}}^{-1} - \frac{1}{2}(\mathbf{T} + \mathbf{T}^{\mathrm{T}}) \tag{5.14}$$

is positive definite, where $\bar{\mathbf{G}} = \mathrm{diag}(\bar{g}_1', \bar{g}_2', \ldots, \bar{g}_{K_2}')$ are the maximum slopes of the gain functions with $\bar{g}_i' \geq g_i'(u)$ for all $u$ (Forti & Tesi, 1995; Lu & Chen, 2003). To see why this condition guarantees the invertibility of $\mathbf{M}(\mathbf{u})$, define

$$\mathbf{S}(\mathbf{u}) \equiv \mathbf{D}\mathbf{G}(\mathbf{u})^{-1} - \frac{1}{2}(\mathbf{T} + \mathbf{T}^{\mathrm{T}}), \tag{5.15}$$

which can be rewritten as

$$\mathbf{S}(\mathbf{u}) = \frac{1}{2}\Big(\mathbf{M}(\mathbf{u})\mathbf{G}(\mathbf{u})^{-1} + \big(\mathbf{M}(\mathbf{u})\mathbf{G}(\mathbf{u})^{-1}\big)^{\mathrm{T}}\Big) \tag{5.16}$$

using equation 51.3. The positive definiteness of $\mathbf{S}$ implies the positive definiteness of $\mathbf{S}(\mathbf{u})$ because their difference $\mathbf{S}(\mathbf{u}) - \mathbf{S}$ is a positive diagonal matrix by equations 5.14 and 5.15. It follows that $\mathbf{M}(\mathbf{u})$ must be invertible, because otherwise $\mathbf{S}(\mathbf{u})$, as given by equation 5.16, would be degenerate rather than positive definite. In conclusion, if one replaces condition 4 by the assumption that $\mathbf{S}$ in equation 5.14 is positive definite, then both global

stability and local stability follow automatically, and the conclusions of theorem 4 stay the same.

Why are the results of the recurrent network so similar to those of the feedforward network? This is because a globally stable recurrent network can be approximated linearly around an equilibrium point as a feedforward network with the same architecture (same number of layers and same number of neurons in each layer but without the recurrent connections). The locally equivalent feedforward network has connection weights that can incorporate the effects of the recurrent connections of the recurrent network.

Despite their similarity, a globally stable recurrent network in general is not strictly equivalent to a feedforward network with the same architecture, in the sense that they cannot have an identical response to all possible stimuli. Their similarity holds only for a single local region of the stimulus space and cannot be extended globally. An explicit example is given in appendix B.

Theorem 4 for recurrent networks may be generalized to gain functions with cutoff or saturation in a manner similar to theorem 3. Since we have already considered the parallels between the feedforward network and globally stable recurrent network, further discussion will be omitted.

## 6 Extensions and Applications

**6.1 Controllability of Network with Degenerate Weights.** We have shown that the exact same conditions on neural network architecture that guarantee that for all neurons in any layer the optimal stimulus must lie on the boundary of a compact stimulus space also guarantee that with suitable inputs, one can locally control the activity of all of the neurons of a given layer. However, the theorems in the preceding sections always assume full rank or nondegenerate weight matrices for feedforward connections between consecutive layers. Now we relax this condition and consider the consequences of degenerate weight matrices for the controllability of neurons by perturbations of stimuli.

In a network allowing degenerate connections between layers, how many neurons in layer $m$ can be controlled simultaneously and independently by an external stimulus? Consider the rank of the final Jacobian $\mathbf{J}$, which is $\partial \mathbf{r}_m / \partial \mathbf{x}$ in equation 4.5 for feedforward networks or $\partial \mathbf{u}_m / \partial \mathbf{x}$ in equation 5.11 for recurrent networks. The rank of Jacobian $\mathbf{J}$ is equal to the maximum number of neurons in layer $m$ that can be controlled simultaneously and independently by the stimulus $\mathbf{x}$. The rank may be measured experimentally by first estimating $\mathbf{J}$ locally around some network state using equation 4.17. A perturbation of response pattern can be produced by a suitable stimulus only if it lies in the range space of the Jacobian.

How is the rank of the final Jacobian determined by the architecture of the network? Consider the same feedforward network as in theorem 1,

except that now we allow arbitrary numbers of neurons $K_1, K_2, \ldots, K_m$ in successive layers and allow degenerate weight matrices. Now equations 4.2, 4.5, and 4.8 remain valid, but inequality 4.9 and equation 4.10 no longer hold. Since the overall Jacobian is $\mathbf{J} = \mathbf{J}^{(m)}\mathbf{J}^{(m-1)} \ldots \mathbf{J}^{(2)}$ by equation 4.5, we have

$$\operatorname{rank} \mathbf{J} \leq \min \left\{ \operatorname{rank} \mathbf{J}^{(m)}, \ldots, \operatorname{rank} \mathbf{J}^{(2)} \right\} \tag{6.1}$$

because the rank of a product of matrices cannot exceed the rank of any matrix in the product. Using equation 4.8, we have $\operatorname{rank} \mathbf{J}^{(n)} = \operatorname{rank} \mathbf{W}^{(n)} \leq \min\{K_n, K_{n-1}\}$ since $\mathbf{W}^{(n)}$ has dimension $K_n \times K_{n-1}$. Substitution into equation 6.1 yields an upper bound,

$$\operatorname{rank} \mathbf{J} \leq \min\{K_1, K_2, \ldots, K_m\}. \tag{6.2}$$

Therefore, the number of simultaneously and independently controllable neurons in any given layer is bounded by the smallest number of neurons in any lower layer of the network.

The upper bound can be improved if we know how many neurons in each layer fall in the flat range of their gain functions (saturated or below threshold). These neurons reduce the rank of $\mathbf{J}^{(m)} = \mathbf{G}^{(m)}\mathbf{W}^{(m)}$ in equation 4.2, because $\mathbf{G}^{(m)}$ no longer has full rank due to zero diagonal terms. If layer $n$ has $S_n$ neurons in the flat range, then $\operatorname{rank} \mathbf{G}^{(n)} = K_n - S_n$, and $\operatorname{rank} \mathbf{J}^{(n)} \leq K_n - S_n$. This leads to a tighter upper bound:

$$\operatorname{rank} \mathbf{J} \leq \min\{K_1 - S_1, \ldots, K_m - S_m\}. \tag{6.3}$$

Although the upper bounds 6.2 and 6.3 are derived for feedforward networks, they also hold true for layered recurrent networks (see section 3.3). Due to the similarity between the two types of systems (see section 5), further discussion is omitted.

Suppose layer $m$ has more neurons than the rank $k = \operatorname{rank} \mathbf{J}$ of the Jacobian. Then which particular subsets of $k$ neurons in layer $m$ are simultaneously and independently controllable? Select $k$ neurons, and delete all entries related to the unselected neurons from $\mathbf{J}$. If the remaining matrix has full rank, then these $k$ neurons are controllable. The selection of the $k$ neurons may not be unique. We can also consider the range space of the Jacobian matrix and identify all the neurons with indices corresponding to the indices of the components of the range space. The number of the identified neurons should be no less than $k$, and any $k$ of these identified neurons can be controlled simultaneously and independently.

The layer with the smallest number of neurons forms an information processing bottleneck according to equations 6.2 and 6.3, and this consequence might imply that in sensory systems, the maximum number of neurons in a

higher stage of processing, which can be simultaneously and independently controlled, is limited by the number of neurons in the bottleneck stages. For instance, in the cat auditory system, the bottleneck actually occurs at the input layer since there are only 3000 inner hair cells (Ryugo, 1992). This would imply that at most, this number of cells at higher processing stages would be simultaneously and independently controllable using sound inputs only.

**6.2 Convergence and Divergence in Sensory Networks.** The theorems in this letter rely on the assumption of convergent network structure, with fewer neurons in higher layers than in lower layers. Thus, it might appear at first glance that these results would have very limited applicability to the mammalian central nervous system, because in mammalian sensory systems, there is in general a massive divergence of connections from the periphery to higher processing areas. For instance, in the cat auditory system, only 3000 inner hair cells are innervated by roughly 50,000 auditory nerve fibers, which in turn provide divergent input to an even greater number of neurons at the next stage of synaptic processing in the cochlear nucleus (Ryugo, 1992; Young, 1998). Similarly, in the primate visual system, the optic nerve represents an information bottleneck, as it contains only about $\sim$1 million fibers compared to the to the roughly 160 million neurons downstream in one hemisphere of the primary visual cortex (Spear, Kim, Ahmad, & Tom, 1996; O'Kusky & Colonnier, 1982; Schein & de Monastero, 1987).

However, the assumptions of the theorems are not as restrictive as they first may appear. This is because they apply only to the actual functional subnetwork which connects a neuron of interest to the sensory periphery. All neurons that do not contribute to the response of the neuron of interest can be ignored. This notion of a functional subnetwork is illustrated in Figure 5. In Figure 5A, neuron $\alpha$ in layer 3 is at the top of a convergent pyramid (shaded), in which each neuron receives connections from multiple neurons in the layer immediately below it, and so on recursively all the way down to the periphery. Since only those shaded neurons contribute to the response of neuron $\alpha$, all other neurons can be ignored. Such convergent functional subnetworks can be entirely compatible with a global divergence pattern with more neurons in higher layers. Conversely, even when there is global convergence pattern with fewer neurons in higher layers, the functional subnetwork may not be convergent and our theorems still may not apply (see Figure 5D). Similarly, this notion of a functional subnetwork can be readily generalized to the layered recurrent networks considered in previous sections by including all contributing neurons regardless of whether the connection is feedforward or recurrent. So to reiterate, the theorems apply only to the actual functional subnetwork, which connects the neurons of interest to the periphery.

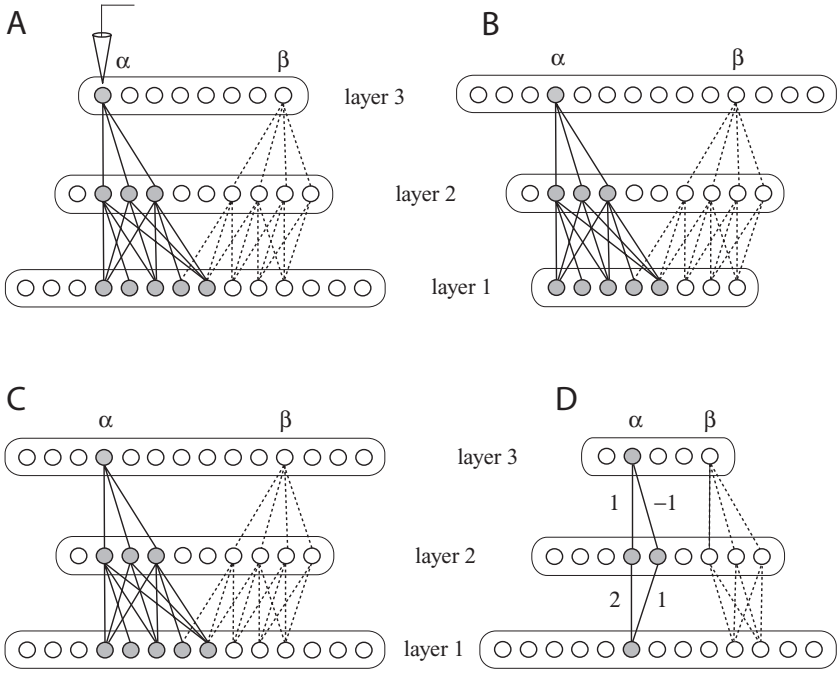


Figure 5: The response of a neuron ($\alpha$ or $\beta$) is determined by the functional subnetwork (shaded or dashed) that connects it to the periphery, regardless of the total numbers of neurons in different layers. The examples here show identical functional subnetworks for neurons $\alpha$ and $\beta$ when the total numbers of neurons from layer 1 to layer 3 are convergent (**A**), divergent (**B**), or have a bottleneck (**C**). The theorems in this letter apply to neurons $\alpha$ and $\beta$ in **A**, **B**, and **C**, but not in **D**, because in the last case the functional subnetworks are not convergent despite the fact that the total numbers of neurons in different layers are convergent. In fact, neuron $\alpha$ in **D** does have an optimal stimulus in the interior (the same example as in Figure 1C).

Consider the concrete example of the visual system. Although there is massive divergence as one goes from the optic nerve to the cortex, it is conceivable that the functional feedforward network connecting a simple cell in the primary visual cortex to the periphery is actually convergent, because a single simple cell receives convergent inputs from multiple cells in lateral geniculate nucleus (LGN) (Tanaka, 1983; Reid & Alonso, 1995). The number of LGN cells is approximately equal to the number of retinal ganglion cells (approximately 1 million in macaque monkeys; see Spear et al., 1996; Schein & de Monastero, 1987), which in turn is one or two orders of magnitude less than the number of retinal photoreceptors

(Packer, Hendrickson, & Curcio, 1989). Thus, it is quite possible that in a feedforward approximation, a single V1 cell may sit on top of a convergent functional subnetwork to which the theorems would apply (see Figure 5C).

**6.3 Nonmonotonic Responses.** The theoretical result that the optimal stimulus is expected to lie on the boundary of the stimulus set under very general conditions may seem to contradict many experimental results that show nonmonotonic responses. For example, a simple cell in primary visual cortex has a preferred orientation corresponding to the peak response in the orientation tuning curve (Hubel & Wiesel, 1962). Since nonmonotonic responses are not uncommon among neurons in various sensory modalities, one may wonder why the conditions of theorems are violated so often given their generality. Here we show that nonmonotonic responses do not necessarily violate the conditions of the theorems if the results are obtained by using stimuli restricted to a lower-dimensional subset of the full stimulus space. An optimal stimulus obtained this way may not be a genuine peak in the stimulus-response relationship because small perturbations of the stimulus in additional dimensions could further increase the response. The possibility of true violation of the conditions of the theorems is discussed at the end of this section.

The simple example in Figure 6 illustrates how restricting the stimulus space to a lower-dimensional subset can lead to nonmonotonic responses. All the neurons in this network have the standard logistic gain function $g(x) = 1/(1 + e^{-x})$ as shown in Figure 1A. This network satisfies all the conditions of theorem 1, and therefore in the two-dimensional stimulus space $(x_1, x_2)$, the optimal stimulus always lies on the boundary of any compact set of stimuli (see Figures 6C and 6D). However, the situation changes completely when the stimulus is restricted to a one-dimensional subspace. Suppose we allow only input $x_1$ to vary freely while keeping the other input fixed at $x_2 \equiv 0$. Then the network becomes identical to the one in Figure 1C, which has a tuning curve with a peak. Thus, in the one-dimensional stimulus space $x_1$, the optimal stimulus no longer lies on the boundary (see Figures 1C and 6B). Similarly, if we keep $x_1$ constant and allow $x_2$ to vary, then the response to stimulus $x_2$ also has a peak (Figure 6B). In fact, by fixing any one input at any given level, the response to the remaining input always has a peak (see Figure 6B). Nonmonotonic responses become possible here because when the stimulus is restricted to a one-dimensional subspace, the functional subnetwork has only one input but two neurons in the next layer, and thus is no longer convergent. This example suggests that when a putative optimal stimulus is found in experiment, it may be useful to perturb the stimulus by altering parameters in additional dimensions of the full stimulus space so as to test whether it is possible to further increase the peak response.

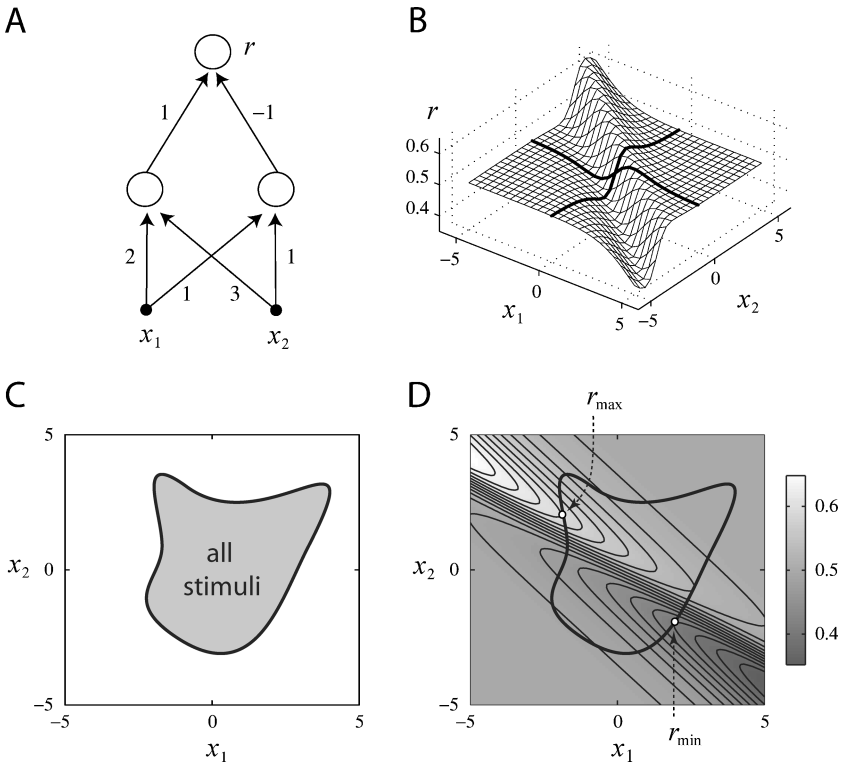


Figure 6: A neuron whose optimal stimulus lies on the boundary of a stimulus set may respond nonmonotonically if the stimuli are restricted to a lower-dimensional subspace. (**A**) This network with two inputs is reduced to the example in Figure 1C if input $x_2$ is fixed at $x_2 \equiv 0$. The numbers by the arrows indicate the synaptic weights. (**B**) The tuning curve for any one input (either $x_1$ or $x_2$) always has a peak when the other input is kept constant. The thick lines indicate the special cases for $x_1 \equiv 0$ or $x_2 \equiv 0$. (**C**) An arbitrary compact stimulus set in the two-dimensional stimulus space. (**D**) The same stimulus set as in **C** is shown together with the responses (shades of gray) and the iso-response contours. Both the optimal stimulus, which elicits the maximum response $r_{\max}$, and the stimulus that elicits the minimum response $r_{\min}$ must lie on the boundary of the stimulus set according to the theorems in this letter.

Now we return to the simple cell example and show that similar ideas can be applied. Consider a single-layer feedforward model of a simple cell, and let the response $r$ be given by

$$r = g\left(\sum_{i=1}^{N} w_i x_i\right), \tag{6.4}$$

where $g$ is a monotonically increasing gain function, the stimulus pattern $(x_1, \ldots, x_N)$ corresponds roughly to the activity in the retina or the lateral geniculate nucleus, and the weight pattern $(w_1, \ldots, w_N)$ is arranged properly to generate a preferred orientation (Hubel & Wiesel, 1962). This network has only one weight vector, and theorem 1 applies trivially, leading to the conclusion that the optimal stimulus should always lie on the boundary of the stimulus set. It is known that the responses of many V1 neurons increase monotonically with increasing stimulus contrast (Maffei & Fiorentini, 1973; Albrecht & Hamilton, 1982). For these neurons, the optimal stimulus is a bar of the best orientation presented at maximum contrast possible, and this stimulus lies on the boundary of the stimulus space (see section 2). In terms of the model in equation 6.4, the optimal stimulus should maximize the luminance for all input pixels $x_i$ with positive weights ($w_i > 0$) while minimizing the luminance for all pixels with negative weights ($w_i < 0$). Even for higher-level visual areas like V4 and IT, the firing rates of many neurons tend to increase with stimulus contrast (Cheng, Hasegawa, Saleem, & Tanaka, 1994; Oram, Xiao, Dritschel, & Payne, 2002), thus suggesting that the optimal stimuli in the pixel space for these potentially highly nonlinear neurons might lie also on the boundary composed of maximum contrast stimuli.

Nonmonotonic responses of the perceptron model in equation 6.4 are possible only when the original stimulus set is restricted to a lower-dimensional subset. When an oriented bar with fixed shape and luminance is used to measure the orientation tuning curve, the only freely varying parameter is the orientation angle. This stimulus subset is only one-dimensional. As another example, when the vector length of the stimulus pattern is fixed, namely, $\sum_{i=1}^{N} x_i^2 = \text{constant}$, the optimal stimulus pattern is proportional to the weight pattern, or $(x_1, \ldots, x_N) \propto (w_1, \ldots, w_N)$. In this case, the stimulus set is a sphere, which is of lower dimension than the original stimulus set and has no interior (see section 2). Therefore, these examples can be readily reconciled with the theorems in this letter.

In conclusion, the "optimal stimuli" found in many experiments that show nonmonotonic or peaked tuning curves for various stimulus parameters do not necessarily contradict the theorems in this letter, because peaking in one dimension of the stimulus space does not imply a strict maximum in the other stimulus dimensions. To test whether a putative optimal stimulus is a true peak, one needs to perturb the stimulus along various dimensions and examine whether any perturbation away from the optimum reliably reduces the response. On the other hand, true optimal stimuli in the sense of strict maxima in the stimulus-response relation may well exist in the sensory pathways and serve useful functions, but these neurons would have to violate the conditions of the theorems in this letter. For example, the weight matrix might become degenerate in order to achieve stimulus specificity by synaptic learning processes. Synaptic learning is implicated here because if one assigns the weights randomly by drawing from any continuous

probability distributions, the weight matrix would be unlikely (with probability 0) to be exactly degenerate. Experimental examination of the relationship between optimal stimuli and the underlying network architecture would ultimately require the development of new methods that can efficiently measure the connection weight matrices in sensory neuronal networks.

**6.4 Optimizing Quantities Other Than Firing Rate.** In the preceding sections, "response" of a neuron always refers to its firing rate. When the response is taken as some quantity measured from the spike train other than the firing rate, the conclusions in the preceding sections may or may not hold. In this section we provide concrete examples to illustrate both possibilities. For example, suppose that latency of the first spike is used to quantify the response. Let $r = f(x)$ be the mean firing rate response of a neuron to stimulus $x$. For Poisson model with mean rate $f(x)$, the latency has an exponential distribution, and its mean value is inversely proportional to the mean firing rate:

$$\tau(x) = 1/f(x). \tag{6.5}$$

Thus, the stimulus that maximizes the firing rate will minimize latency and vice versa. The same conclusions we have derived for the mean firing rate measure will hold with the latency measure in this hypothetical example.

For another example, the Fisher information measure may yield different conclusions from that of the mean firing rate. Given the firing rate function $f(x)$ as considered above, the Fisher information for spikes within a unit time interval is

$$I(x) = f'(x)^2 \big/ f(x), \tag{6.6}$$

assuming Poisson spike statistics (Seung & Sompolinsky, 1993). The stimulus that maximizes the Fisher information is a solution to the algebraic equation $I'(x) = 0$. In special cases such as when $I(x) = \Psi(f(x))$ where $\Psi$ is any strictly monotonic function, the Fisher information $I(x)$ and the firing rate $f(x)$ always have identical maximum and minimum. However, in general, they are different. For example, in Figure 1B, the firing rate is maximized at a boundary point of the stimulus set $[a, b]$, that is, $\hat{x} = \arg\max_x f(x) = b$, whereas the Fisher information reaches its peak at $\hat{x} = \arg\max_x I(x) = \ln\left(\frac{1}{12}\left(1 + \sqrt[3]{217 + 12\sqrt{327}} + \sqrt[3]{217 - 12\sqrt{327}}\right)\right) \approx -0.3212$, which is an interior point of the interval $[a, b]$. Therefore, the theorems in this letter may not hold when the response is quantified by Fisher information. The theorems also may not apply when the optimization involves an ensemble

of stimuli rather than an individual stimulus, such as the mutual information between stimuli and responses (Machens, 2002).

## 7 Discussion

We have analyzed how the structure of a neural network can constrain one aspect of a neuron's stimulus-response relationship: the location of the stimulus eliciting the maximum firing rate. For any given neuron whose connections with the periphery form a convergent network with more neurons in lower layers and nondegenerate synaptic weight matrices, no optimal stimulus can correspond to a strict maximum such that moving the stimulus away from the optimum in any direction in stimulus space decreases the firing rate. Instead, the stimulus that elicits the maximum response must always lie on the topological boundary of the set of permissible stimuli. Once we know that the optimal stimulus lies on the boundary, we can avoid the interior altogether and test only stimuli on the boundary to search for the optimal stimulus, which may potentially reduce the number of experimental trials because the dimension of the boundary is typically one less than the dimension of the full stimulus set.

These results hold not only for feedforward networks, but also for layered recurrent networks where neurons within each layer may connect arbitrarily with one another, provided that the recurrent network is globally stable in the sense that its final equilibrium state depends on only the stimulus, not its initial state. These stability conditions are biologically reasonable for sensory systems that are not inherently oscillatory or chaotic.

The conditions for the nonexistence of optimal stimulus at the interior also imply that each layer of the network is locally controllable in the sense that one can always find a stimulus to produce any desired pattern of perturbation to the activity of all the neurons in any given layer. The manipulation of holding constant the activities of all neurons in a layer except for one has been used as a technique in proving statement (b) of theorem 3. In a real experiment, similar perturbation might help dissociate the effects of the synaptic inputs arising from different groups of neurons.

To apply the theoretical results to real systems, one has to examine the topology of the stimulus set carefully, while keeping in mind that the input stimulus should be similar to the actual inputs to the bottom layer of the network. The examples of stimulus sets considered in this letter are simplistic and intended for illustration purposes only. Further work is needed to see whether the theory developed here can be applied to stimuli with complex high-level sensory features, especially those that occur in the natural environments.

The theoretical results obtained in this letter are probably more relevant to sensory networks that can be well approximated as a feedforward

network, especially in the periphery. We have generalized the theory to layered recurrent networks that allow lateral connections within each layer. A limitation of this approach is that when feedback connections are included, the equivalent layered network is reduced to having only two layers (see Figure 3D). The theory still holds true, but the results become more limited. For high-level brain areas where detailed anatomical network structure is not generally known or even well studied, our results might still help provide some information about the structure of the functional neural network, which connects a neuron to the periphery. For instance, given a neuron that exhibits a strict maximum in a stimulus parameter space that resembles the actual input to the periphery, then by the theorems derived in this letter, the functional subnetwork must violate at least one of the conditions of the theorems so that the network might be divergent or the synaptic weights might be degenerate (see also section 6.3). Otherwise the theorems would hold and lead to logical contradiction.

Since a neuron's response to a single fixed stimulus varies randomly from trial to trial, it would be useful to extend this work to a stochastic framework. For the theorems to apply, one needs a sufficient number of trials for each stimulus in order to reliably estimate the true mean firing rate. Another limitation of this study is that we have only considered globally asymptotically stable recurrent networks and have demonstrated only local controllability with sufficiently small change of the stimulus. It would be of interest to address issues of global controllability in continuous dynamical models, preferably also in a stochastic framework. This study is also limited in that it does not apply to transient responses and oscillatory systems.

Adaptation and plasticity can alter the weight matrix in a neural network and the slope and threshold of the gain functions. These time-varying factors can greatly reduce the effectiveness of online optimization procedures. The theorems in this letter could still hold as long as the gain functions remain monotonically increasing, and the synaptic changes do not make the weight matrix degenerate. In such situations, the optimal stimulus would still be confined to the boundary of the stimulus set, although its exact location could drift along the boundary as adaptation and plasticity are taking place. So the theory might still help us avoid searching the interior of the stimulus space.

Previous studies have employed various methods to find the optimal stimulus, such as reverse-correlation methods (de Boer & Kuyper, 1968; deCharms, Blake, & Merzenich, 1998), iterative online maximization of the firing rate (Harth & Tzanakou, 1974; Tzanakou, Michalak, & Harth, 1979; Nelken, Prut, Vaadia, & Abeles, 1994; Anderson & Micheli-Tzanakou, 2002; Bleeck, Patterson, & Winter, 2003; O'Connor, Petkov, & Sutter, 2005; Bandyopadhyay & Young, 2005), and maximization of mutual information between stimulus and response ensembles (Machens,

2002). The basic reverse-correlation method is based on linear stimulus-response relation, similar to a perceptron model (see section 6.3). As explained in section 6.3, the existence of an optimal stimulus in the interior of a restricted stimulus space does not necessarily violate the conditions of the theorems in this letter. Our results here apply only to firing rate maximization for systems whose underlying functional subnetworks are convergent. If the sensory stimuli are defined in a space isomorphic to the periphery, then one need only to search the boundary of this space to find the optimal stimulus. For maximizing criteria other than firing rate, the results in this letter may no longer hold true except for special cases when a criterion is monotonically related to the firing rate (see section 6.4).

Finally, this work might be applicable to the problem of neural network inversion, which is to find the level set consisting of all the inputs that lead to the same output in a trained neural network (Williams, 1986; Jensen et al., 1999). Finding the optimal stimulus for a neural network may be regarded as a special case of neural network inversion if the maximum response is known, and previous studies in both neuroscience and engineering have made use of this type of procedure (Lehky, Sejnowski, & Desimone, 1992; Jensen et al., 1999). The inversion may help analyzing network models derived from stimulus-response data of real neurons (Lehky et al., 1992; Lau, Stanley, & Dan, 2002; Prenger, Wu, David, & Gallant, 2004). Our work suggests that for a neural network with strictly increasing gain functions, like the sigmoid units in connectionist models, the search space for the optimal input for the network can be potentially reduced to the boundary of the input space.

## Appendix A: A Globally Stable Recurrent Network with a Degenerate Linearized System

The example that follows demonstrates that global stability of a recurrent network does not always guarantee that its linearized system is nondegenerate. A stable linear system is automatically nondegenerate. This is why condition 4 of theorem 4 explicitly requires the stability of the linearized system.

Consider a network of two neurons sending excitatory connections to themselves and to each other, all with the same weight $1/2$, and assume that each neuron receives the same stimulus input $x$. The dynamical equations are

$$\frac{dv_1}{dt} = -v_1 + \frac{1}{2}g(v_1) + \frac{1}{2}g(v_2) + x, \tag{A.1}$$

$$\frac{dv_2}{dt} = -v_2 + \frac{1}{2}g(v_1) + \frac{1}{2}g(v_2) + x, \tag{A.2}$$

where the gain function is $g(v) = \tanh(v)$. Subtracting the two equations yields

$$\frac{d}{dt}(v_1 - v_2) = -(v_1 - v_2). \tag{A.3}$$

Therefore, the states of the two neurons always become identical over time regardless of the initial states and the input $x$. Writing $v_1 = v_2 \equiv v$, we condense equations A.1 and A.2 into a single equation:

$$\frac{dv}{dt} = -v + g(v) + x. \tag{A.4}$$

For any input $x$, this system always approaches a single equilibrium state given by

$$-u + g(u) + x = 0, \tag{A.5}$$

which has a unique solution for any input $x$ because the line $u - x$ always has a unique intersection with the gain function $g(u) = \tanh u$ whose maximum slope is 1.

This network is globally stable because for any given stimulus $x$, the system always approaches a single equilibrium state given by equation A.5 regardless of the initial state. On the other hand, the matrix as defined by equation 5.4 is

$$\mathbf{M}(\mathbf{u}) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \frac{g'(u)}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \tag{A.6}$$

which is degenerate when $\mathbf{u} = (u, u) = (0, 0)$, which occurs for $x = 0$ in equation A.5. Equation 5.6 for the linearized system around $(0, 0)$ becomes

$$\frac{d}{dt} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = - \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \tag{A.7}$$

which is degenerate, with eigenvalues 0 and $-2$. By requiring that the linearized system be stable, as in condition 4 of theorem 4, it is guaranteed also to be nondegenerate, as needed in the proof of the theorem.

## Appendix B: Nonequivalence of Recurrent and Feedforward Networks

In the main text, we have summarized qualitatively similar results for two very different neural network architectures: feedforward networks and

globally stable recurrent networks. One possible explanation for the similarity is that the equilibrium state of a recurrent network in theorem 4 can be represented exactly by some equivalent feedforward network having the same number of layers and neurons, but possibly having different feedforward weights and gain functions. Here we provide a counterexample to show that this is not generally true.

Consider a two-layer recurrent network with two excitatory neurons connected reciprocally by the weight matrix

$$\mathbf{T} = \begin{bmatrix} 0 & 1 \\ \varepsilon & 0 \end{bmatrix}, \tag{B.1}$$

where $0 < \varepsilon < 1$, and the 0 diagonal entries mean no self-connection. The two neurons 1 and 2 receive distinct input $x_1$ and $x_2$, respectively, and obey the dynamical equations

$$\frac{dv_1}{dt} = -v_1 + v_2 + x_1, \tag{B.2}$$

$$\frac{dv_2}{dt} = -v_2 + \varepsilon \tanh v_1 + x_2, \tag{B.3}$$

with the gain functions $g_1(u) = \tanh u$ and $g_2(u) = u$. This network is globally asymptotic stable because matrix $\mathbf{S}$ defined by equation 5.14 is positive definite; that is,

$$\mathbf{S} \equiv \mathbf{D}\bar{\mathbf{G}}^{-1} - \frac{1}{2}(\mathbf{T} + \mathbf{T}^{\mathrm{T}}) = \begin{bmatrix} 1 & -\frac{1+\varepsilon}{2} \\ -\frac{1+\varepsilon}{2} & 1 \end{bmatrix} \tag{B.4}$$

has positive eigenvalues $1 \pm \frac{1+\varepsilon}{2}$ for $0 < \varepsilon < 1$. Here both the decay rate matrix $\mathbf{D}$ and the maximum gain slope matrix $\bar{\mathbf{G}} = \mathrm{diag}(\bar{g}_1', \bar{g}_2')$ are the identity matrix.

This recurrent network satisfies all the conditions of theorem 4. We show that its equilibrium state cannot be represented exactly by any feedforward network of the same size. Using $(u_1, u_2)$ instead of $(v_1, v_2)$ to denote the equilibrium state, we have

$$u_1 = u_2 + x_1, \tag{B.5}$$

$$u_2 = \varepsilon \tanh u_1 + x_2. \tag{B.6}$$

Substituting equation B.6 into B.5 yields

$$F(u_1, x) \equiv u_1 - \varepsilon \tanh u_1 - x = 0, \tag{B.7}$$

where $x \equiv x_1 + x_2$ is the combined input. Since $\partial F(u_1, x)/\partial u_1 = 1 - \varepsilon(\tanh u_1)' \neq 0$ for $0 < \varepsilon < 1$ and $0 < \tanh' \leq 1$, the implicit function theorem (Rudin, 1976) guarantees the existence of an input-output function for neuron 1 in the equilibrium state:

$$u_1 = f(x) = f(x_1 + x_2). \tag{B.8}$$

The input-output function for neuron 2 follows from equations B.8 and B.5:

$$u_2 = f(x_1 + x_2) - x_1. \tag{B.9}$$

Equations B.8 and B.9 can be implemented by a three-layer feedforward network (interpreting $f$ as a gain function), but not by a two-layer feedforward network.

Seeking a contradiction, suppose that a two-layer feedforward neural network with two neurons always has the same output as the recurrent network in the equilibrium state. Write the outputs of the two neurons as $\tilde{g}_1(w_{11}x_1 + w_{12}x_2)$ and $\tilde{g}_2(w_{21}x_1 + w_{22}x_2)$, where the weights $w_{ij}$ and gain functions $\tilde{g}_i$ can be freely chosen. Since the output of neuron 2 must match its recurrent network counterpart for the special case $x_1 = 0$, we have

$$\tilde{g}_2(w_{22}x_2) = g_2(u_2) = f(x_2), \tag{B.10}$$

where the last step follows from $g_2(u_2) = u_2$ and equation B.9. It follows from equation B.10 that the gain function $\tilde{g}_2$ must be of the form

$$\tilde{g}_2(z) = f\left(\frac{z}{w_{22}}\right). \tag{B.11}$$

Similarly, for $x_2 = 0$, we obtain $\tilde{g}_2(w_{21}x_1) = g_2(u_2) = f(x_1) - x_1$ and

$$\tilde{g}_2(z) = f\left(\frac{z}{w_{21}}\right) - \frac{z}{w_{21}}. \tag{B.12}$$

Consistency requires that equations B.11 and B.12 be identical, which is equivalent to that

$$f(kx) = f(x) - x \tag{B.13}$$

for all $x$ and for some fixed $k \neq 1$ (otherwise, $k = 1$ in equation B.13 would imply $x \equiv 0$). It follows from equation B.13 that $f(x) = a_0 + a_1 x$. To see this, assuming that the Taylor series $f(x) = \sum_{n=0}^{\infty} a_n x^n$ is convergent, we rewrite equation B.13 as $\sum_{n=0}^{\infty} a_n k^n x^n = \sum_{n=0}^{\infty} a_n x^n - x$. Since the coefficients of the same order on both sides of the equation must be equal, we have $a_n k^n = a_n$

for all $n \geq 2$, which implies $a_n = 0$ since $k \neq 1$. The linear function $f(x) = a_0 + a_1 x$ contradicts the fact that $f(x)$ must satisfy the nonlinear equation $f(x) - \varepsilon \tanh f(x) - x = 0$, according to equations B.7 and B.8. Hence this feedforward network cannot be equivalent to the recurrent network across all inputs.

## References


Adrian, E. D. (1928). *The basis of sensation: The action of the sense organs*. New York:. Norton.

Albrecht, D. G., & Hamilton, D. B. (1982). Striate cortex of monkey and cat: Contrast response function. *Journal of Neurophysiology*, *48*, 217–237.

Amari, S.-I. (1972). Characteristics of random nets of analog neuron-like elements. *IEEE Transactions on Systems, Man and Cybernetics*, *2*, 643–657.

Anderson, M. J., & Micheli-Tzanakou, E. (2002). Auditory stimulus optimization with feedback from fuzzy clustering of neuronal responses. *IEEE Trans. Inf. Technol. Biomed.*, *6*, 159–170.

Bandyopadhyay, S., & Young, E. (2005). *Spectral information fields of DCN principal neurons and their optimal spectral features.* Abstract 50 Cosyne 2005 meeting, Snowbird, UT.

Ben-Israel, A., & Greville, T. N. E. (2003). *Generalized inverses: Theory and applications* (2nd ed). New York: Springer.

Bhatia, N. P., & Szegö, G. (1970). *Stability theory of dynamical systems*. Berlin: Springer-Verlag.

Bleeck, S., Patterson, R. D., & Winter, I. M. (2003). Using genetic algorithms to find the most effective stimulus for sensory neurons. *J. Neurosci. Methods*, *125*, 73–82.

Cheng, K., Hasegawa, T., Saleem, K. S., & Tanaka, K. (1994). Comparison of neuronal selectivity for stimulus speed, length, and contrast in the prestriate visual cortical areas V4 and MT of the macaque monkey. *J. Neurophysiol.*, *71*, 2269–2280.

de Boer, R., & Kuyper, P. (1968). Triggered correlation. *IEEE Trans. Biomed. Eng.*, *15*, 169–179.

deCharms, R. C., Blake, D. T., & Merzenich, M. M. (1998). Optimizing sound features for cortical neurons. *Science*, *280*, 1439–1443.

Forti, M., & Tesi, A. (1995). New conditions for global stability of neural networks with application to linear and quadratic programming problems. *IEEE Transactions on Circuits and Systems*, *42*, 354–366.

Harary, F., Norman, R. Z., & Cartwright, D. (1965). *Structural models: An introduction to the theory of directed graphs*. New York: Wiley.

Harth, E., & Tzanakou, E. (1974). Alopex: A stochastic method for determining visual receptive fields. *Vision Res.*, *14*, 1475–1482.

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proc. Natl. Acad. Sci. U.S.A.*, *81*, 3088–3092.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.*, *160*, 106–154.

Jensen, C. A., Reed, R. D., Marks, R. J., El-Sharkawi, M. A., Jung, J. B., Miyamoto, R. T., et al. (1999). Inversion of feedforward neural networks: Algorithms and applications. *Proceedings of the IEEE*, *87*, 1536–1549.

Krantz, S. G., & Parks, H. R. (2002). *The implicit function theorem: History, theory and applications*. Boston: Birkhauser.

Lau, B., Stanley, G. B., & Dan, Y. (2002). Computational subunits of visual cortical neurons revealed by artificial neural networks. *Proc. Natl. Acad. Sci. U.S.A.*, *99*, 8974–8979.

Laurent, G., Stopfer, M., Friedrich, R. W., Rabinovich, M. I., Volkovskii, A., & Abarbanel, H. D. (2001). Odor encoding as an active, dynamical process: Experiments, computation, and theory. *Annu. Rev. Neurosci.*, *24*, 263–297.

Lehky, S. R., Sejnowski, T. J., & Desimone, R. (1992). Predicting responses of nonlinear neurons in monkey striate cortex to complex patterns. *J. Neurosci.*, *12*, 3568–3581.

Lu, W., & Chen, T. (2003). New conditions on global stability of Cohen-Grossberg neural networks. *Neural Comput.*, *15*, 1173–1189.

Machens, C. K. (2002). Adaptive sampling by information maximization. *Phys. Rev. Lett.*, *88*, 228104.

Maffei, L., & Fiorentini, A. (1973). The visual cortex as a spatial frequency analyser. *Vision Res.*, *13*, 1255–1267.

Minsky, M. L., & Papert, S. A. (1988). *Perceptrons: An introduction to computational geometry*. Cambridge, MA: MIT Press.

Munkres, J. R. (1999). *Topology* (2nd ed). Upper Saddle River, NJ: Prentice Hall.

Nelken, I., Prut, Y., Vaadia, E., & Abeles, M. (1994). In search of the best stimulus: An optimization procedure for finding efficient stimuli in the cat auditory cortex. *Hear. Res.*, *72*, 237–253.

O'Connor, K. N., Petkov, C. I., & Sutter, M. L. (2005). Adaptive stimulus optimization for auditory cortical neurons. *J. Neurophysiol.*, *94*, 4051–4067.

O'Kusky, J., & Colonnier, M. (1982). A laminar analysis of the number of neurons, glia, and synapses in the adult cortex (area 17) of adult macaque monkeys. *J. Comp. Neurol.*, *210*, 278–290.

Oram, M. W., Xiao, D., Dritschel, B., & Payne, K. R. (2002). The temporal resolution of neural codes: Does response latency have a unique role? *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, *357*, 987–1001.

Packer, O., Hendrickson, A. E., & Curcio, C. A. (1989). Photoreceptor topography of the retina in the adult pigtail macaque (*Macaca nemestrina*). *J. Comp. Neurol.*, *288*, 165–183.

Prenger, R., Wu, M. C., David, S. V., & Gallant, J. L. (2004). Nonlinear V1 responses to natural scenes revealed by neural network analysis. *Neural Netw.*, *17*, 663–679.

Reid, R. C., & Alonso, J. M. (1995). Specificity of monosynaptic connections from thalamus to visual cortex. *Nature*, *378*, 281–284.

Rieke, F., Warland, D., de Ruyter van Steveninck, R., & Bialek, W. (1997). *Spikes: exploring the neural code*. Cambridge, MA: MIT Press.

Rosenblatt, F. (1962). *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Washington DC: Spartan Books.

Rudin, W. (1976). *Principles of mathematical analysis* (3rd ed.). New York: McGraw-Hill.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.),

*Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1). Cambridge, MA: MIT Press.

Ryugo, D. K. (1992). The auditory nerve: Peripheral innervation, cell body morphology, and central projections. In D. B. Webster (Ed.), *Mammalian auditory fathway: Neuroanatomy* (pp. 23–65). New York: Springer-Verlag.

Schein, S. J., & de Monastero, F. M. (1987). Mapping of retinal and geniculate neurons onto striate cortex of macaque. *J. Neurosci.*, *7*, 996–1009.

Seung, H. S., & Sompolinsky, H. (1993). Simple models for reading neuronal population codes. *Proc. Natl. Acad. Sci. U.S.A.*, *90*, 10749–10753.

Spear, P. D., Kim, C. B., Ahmad, A., & Tom, B. W. (1996). Relationship between numbers of retinal ganglion cells and lateral geniculate neurons in the rhesus monkey. *Vis. Neurosci.*, *13*, 199–203.

Spivak, M. (1965). *Calculus on manifods*. New York: Benjamin.

Tanaka, K. (1983). Cross-correlation analysis of geniculostriate neuronal relationships in cats. *J. Neurophysiol.*, *49*, 1303–1318.

Tzanakou, E., Michalak, R., & Harth, E. (1979). The Alopex process: Visual receptive fields by response feedback. *Biol. Cybern.*, *35*, 161–174.

Williams, R. J. (1986). Inverting a connectionist network by backpropagation of error. In Proc. 8th Annu. Conf. Cognitive Science Society. Hillsdale, NJ: Erlbaum.

Wilson, H. R., & Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys. J.*, *12*, 1–24.

Young, E. D. (1998). Cochlear nucleus. In G. M. Shepherd (Ed.), *Synaptic organization of the brain* (4th ed.). New York: Oxford University Press.

Yu, J. J., & Young, E. D. (2000). Linear and nonlinear pathways of spectral information transmission in the cochlear nucleus. *Proc. Natl. Acad. Sci. U.S.A.*, *97*, 11780–1786.

Zhang, K., Anderson, M., & Young, E. D. (2004). *Are there optimal sounds in nonlinear auditory coding*? Society for Neuroscience Abstracts, no. 305.15.

---

Received November 29, 2005; accepted April 27, 2007.