# Stability of negative-image equilibria in spike-timing-dependent plasticity

Alan Williams* and Patrick D. Roberts†
*Neurological Sciences Institute, Oregon Health and Science University, 505 NW 185th Avenue, Beaverton, Oregon 97006, USA*

Todd K. Leen‡
*Department of Computer Science and Engineering, OGI School of Science and Engineering, Oregon Health and Science University, 20000 NW Walker Road, Beaverton, Oregon 97006, USA*

(Received 15 April 2003; published 29 August 2003)

We investigate the stability of negative image equilibria in mean synaptic weight dynamics governed by spike-timing-dependent plasticity (STDP). The model architecture closely follows the anatomy and physiology of the electrosensory lateral line lobe (ELL) of mormyrid electric fish. The ELL uses a spike-timing-dependent learning rule to form a negative image of the reafferent signal from the fish's own electric discharge, thus improving detectability of external electric fields. We derive sufficient conditions for existence of the negative image and necessary and sufficient conditions for stability, for arbitrary postsynaptic potential functions and arbitrary learning rules. This significantly generalizes earlier investigations. We then apply the general result to several examples of biological interest, including a class of learning rules consistent with the rule observed experimentally in the mormyrid ELL.

DOI: 10.1103/PhysRevE.68.021923 PACS number(s): 87.19.La, 87.18.Sn, 75.10.Nr

## I. INTRODUCTION

Synaptic plasticity is thought to be a fundamental mechanism for learning and adaptation in biological neural networks [1]. The activity dependence of synaptic plasticity has been observed experimentally [2,3], but the precise nature of that dependence and its functional or computational consequences are still largely unknown. The purpose of the present paper is to derive clear functional consequences from specific forms of activity-dependent synaptic plasticity.

Current models of synaptic plasticity are of two main types: rate-based and timing-based. In rate-based models, changes in synaptic weight depend on the mean spike rate of presynaptic and postsynaptic cells, usually via correlations [4,5]. Since mean spike rates are averages over time windows containing many spikes, the timing of individual spikes is unimportant in rate-based models. Recent experimental studies [6–8] have shown that in some systems the precise timing of individual spikes can have a pronounced effect on synaptic plasticity. Models of such *spike-timing-dependent plasticity* (STDP) [9] calculate changes in synaptic weights by combining the effect of all pairs of presynaptic and postsynaptic spikes [10–15], where the effect of each pair is a function of the time between them (called the spike-timing-dependent learning rule).

One system in which STDP has been observed experimentally, and where its functional role is understood, is the electrosensory lateral line lobe (ELL) of mormyrid electric fish [7]. The mormyrid identifies objects in its environment by emitting a stereotyped electrical discharge and detecting the perturbations to the resulting electrical field at the skin surface due to external objects. To cancel the predictable sensory input due to its own discharge, the mormyrid sends to the ELL a sequence of time-delayed, time-locked copies of the motor command which initiates the discharge [Ref. [16], citation (a)]. In the ELL these signals innervate medium ganglion (MG) cells through plastic synapses. The MG cells also receive primary afferent input from electroreceptors on the skin. The plastic synapses onto MG cells enable formation and maintenance of a negative image [17] of the primary afferent signal, via a spike-timing-dependent learning rule. This negative image effectively nulls out the sensory effect of the fish's own discharge, thus improving detectability of perturbations due to external objects. Plasticity allows the negative image to be maintained despite changes in the precise form of the discharge that result from fluctuations in water conductivity or from changes in body shape over the fish's life span.

To be behaviorally useful to the fish, the set of synaptic weights which create the negative image must be a stable equilibrium for the synaptic dynamics induced by the spike-timing-dependent learning rule. Roberts [18] explored stability of such equilibria under restrictive conditions on the form of the learning rule and of the postsynaptic potential function. The approach developed here allows us to derive analytical criteria for both existence and stability of negative image equilibria for systems with *arbitrary* spike-timing-dependent learning rules and arbitrary postsynaptic potential functions.

The structure of the paper is as follows. In Sec. II we describe the architectural and dynamical features of the model, and in Sec. III we derive dynamical equations for the synaptic weights. In Sec. IV we derive conditions for existence of negative image equilibria, and in Sec. V conditions for stability of such equilibria. In Sec. VI we discuss a number of general properties and consequences of the existence and stability criteria, and the role played in those criteria by the various components of the model. In Sec. VII we explicitly evaluate the general stability criteria for several classes

*Electronic address: williaal@ohsu.edu
†Electronic address: robertpa@ohsu.edu
‡Electronic address: tleen@cse.ogi.edu

 

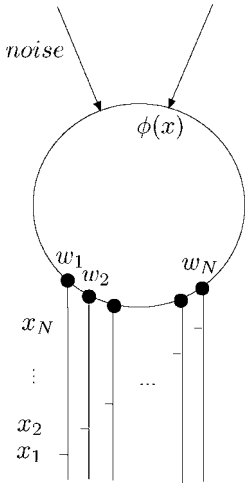


FIG. 1. Schematic of the architecture. The postsynaptic cell receives inputs from $N$ presynaptic neurons, a repeated external input $\phi(x)$, and unspecified noisy inputs. Presynaptic neuron $i$ spikes at time $x_i$ in each period of $\phi$, and has synaptic weight $w_i$ onto the postsynaptic cell.

of learning rule and postsynaptic potential function, and apply these results to the learning rule observed experimentally in the mormyrid ELL.

## II. FRAMEWORK

The model consists of a single postsynaptic cell (representing an MG cell) driven by the following inputs: an array of time-locked presynaptic cells (representing the efference copy of the motor command), a repeated external input (representing the postsynaptic potential in the MG cell due to primary afferents), and other unspecified inputs collectively modeled as noise [19–21] (Fig. 1). This architecture is based on the mormyrid ELL, but is general enough to capture the dynamics of other neural systems hypothesized to have an array of time-delayed, time-locked inputs [22,23].

For the spiking dynamics of the postsynaptic neuron we use the spike response (SR) model [24,25], without refractoriness. In such models the effect of presynaptic spikes on the postsynaptic cell is represented by a postsynaptic potential function (PSP), which is the change in the postsynaptic membrane potential due to the presynaptic spike, as a function of time. Spike response models have been shown to include leaky integrate-and-fire (LIF) models as a special case [26]; so while the formalism of SR models may appear more abstract than LIF, in fact there is no loss of biophysical realism in using SR. We do so here because the SR formalism is more convenient than LIF for the derivation of analytical results.

Each presynaptic cell $i$ spikes exactly once at a fixed time within each sweep of the repeated external input, causing a corresponding PSP in the postsynaptic cell.

The total membrane potential in the postsynaptic cell is the sum of these PSPs, weighted by synaptic efficacies (weights) $w_i$, and the two external inputs. This membrane potential induces the postsynaptic cell to spike at a certain (noisy) rate. Each presynaptic spike causes a constant (non-associative) change in the weight $w_i$, and each postsynaptic and presynaptic spike pair causes a change in $w_i$ according to a spike-timing-dependent learning rule, namely, a function of the time difference between the postsynaptic and presynaptic spikes (associative learning).

The repeated external input has the form of a brief stereotyped pulse with variable interpulse interval. The time-locked inputs occur for approximately the duration of the pulse, and are absent during interpulse intervals [7]. Hence the events which induce plasticity are restricted approximately to the duration of the pulses, provided the width of the learning rule is much less than the width of a pulse (a requirement we will impose below). For the purpose of calculating the weight changes due to plasticity we may therefore omit the interpulse intervals, and replace the repeated external input with a *periodic* input obtained by concatenating the stereotyped pulses.

Denoting the resulting period (pulse width) by $T$, we then use two time variables: $x \in [0,T)$ for the time within each repetition of the external input, and $t=nT$, $n \in \mathbb{Z}$ for the time of initiation of each period [20,21,27]. General dynamical quantities will be functions of the pair $(x,t)$. Let $x_i$ be the time within each period when presynaptic cell $i$ spikes, and $w_i(x,t)$ its corresponding weight. Since presynaptic spikes are time locked to the external input, $x_i$ is independent of $t$. Let $\mathcal{E}(s)$ be the PSP evoked by neuron $i$ at time $s$ after a spike. We assume $\mathcal{E}$ is causal: $\mathcal{E}(s)=0$ for $s<0$. Let $\alpha$ be the nonassociative weight change due to a presynaptic spike, and $\mathcal{L}(s)$ the associative weight change due to a postsynaptic spike time $s$ after a presynaptic spike. Let $\phi(x)$ be the periodic external input, and $U(x,t)$ the total postsynaptic potential due to the non-noisy inputs. We assume that for each $t$, the *mean* instantaneous postsynaptic spike rate density (in $x$) is given by $f(U(x,t))$ for some positive and strictly increasing function $f$.[1] The function $f$ can be thought of as the effective gain of the postsynaptic cell in the presence of the noisy inputs. High or low noise correspond to an $f$ with small or large maximum slope, respectively. No attempt is made to include a refractory period for postsynaptic spikes; and we will assume that the period of $\phi$ is greater than the refractory period of the presynaptic neurons, so that refractoriness on the presynaptic side is irrelevant.

Changes in weights will be implemented as discrete steps with no internal time course. In the present model there are two natural choices for the time at which weight changes occur: asynchronously (instantaneously, whenever a presynaptic or postsynaptic spike occurs) or synchronously (once per sweep of the repeated external input, updating all weights simultaneously). We adopt the latter strategy, updating weights at $x=0$ for each $t=nT, n \in \mathbb{Z}$. The value of $w_i$ in

[1]This simplified treatment of the noise is justified *post hoc* by the calculations to follow. Since we will assume that weight changes due to different spikes or spike pairs add linearly, we will find that the mean synaptic weight dynamics will depend only on the mean postsynaptic spike rate density, and not on any higher moments. Even the functional form of $f$ will turn out to be irrelevant to stability, provided it is strictly increasing.

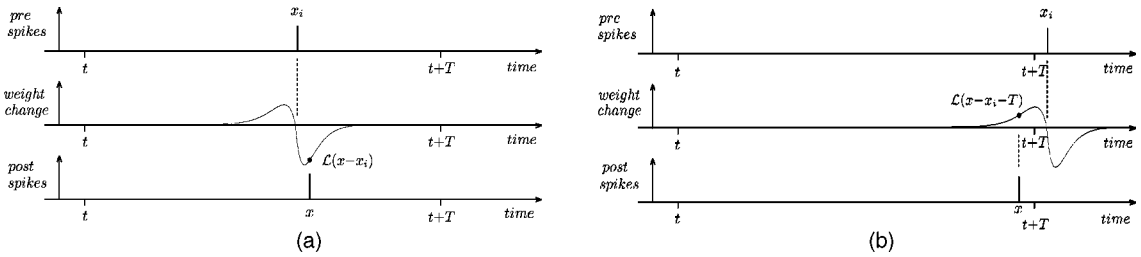


FIG. 2. Changes in weight due to pairing of presynaptic and postsynaptic spikes. (a) Pairing of a postsynaptic spike at time $(x,t)$ and presynaptic spike by neuron $i$ at time $(x_i,t)$ causes a change $\mathcal{L}(x-x_i)$ in weight $w_i$. (b) For $x$ within $\tau_L$ of a period edge, we must include pairing with presynaptic spikes in the neighboring period. Pairing of a postsynaptic spike at time $(x,t)$ and presynaptic spike by neuron $i$ at time $(x_i,t+T)$ causes a change $\mathcal{L}(x-x_i-T)$ in weight $w_i$. Arbitrary units.

the period beginning at $(0,t)$ is then independent of $x$, and will be denoted $w_i(t)$. For synchronous updating to be a reasonable approximation, we must assume that weight changes per cycle are small relative to the weights themselves (slow learning rate). Changes in weights due to different spikes or spike pairs are assumed to add linearly.

In biological systems, synaptic weights have bounded magnitude and do not change sign. Since the present paper is focused solely on the dynamics near equilibria, we impose no boundary conditions on the model. The results still apply to the biological case provided the weight equilibria are in the region enclosed by biological bounds.

We assume homogeneous parameters: the scalar $\alpha$ and the functions $\mathcal{E}$, $\mathcal{L}$ are the same for all presynaptic neurons, and the times $x_i$ are regularly spaced, $x_i = i\delta$, $i=0,1,\ldots,N-1$ for some $\delta>0$, $N=T/\delta \gg 1$.

For simplicity in the derivation of the weight dynamics, it will be convenient to assume that $\mathcal{E}(s),\mathcal{L}(s)$ are zero or negligible for $|s|>\tau_E,\tau_L$, respectively, with $\tau_E,\tau_L \ll T$. We will also require the learning rate to be slow: $T \ll \tau_w$, where $\tau_w$ is the time scale on which weights undergo significant relative change. For the existence of approximate negative image states we will need the spacing of presynaptic spike times much smaller than the widths of $\mathcal{E}$ and $\mathcal{L}$: $\delta \ll \tau_E,\tau_L$. These time-scale assumptions can be summarized as

$$\delta \ll (\tau_E,\tau_L) \ll T \ll \tau_w .$$

Typical values for the mormyrid ELL are $\delta<1$ ms [Ref. [16], citation (b)], $\tau_E \sim 20$ ms [7], $\tau_L \sim 40$ ms [7], $T \sim 80$ ms [Ref. [16], citation (b)], and $\tau_w \sim 10^2 T$ [7].

### III. WEIGHT DYNAMICS

To obtain the mean weight dynamics, we compute the mean value of $w_i(t+T)-w_i(t)$. The nonassociative change in $w_i(t)$ due to the single presynaptic spike at $(x_i,t)$ is $\alpha$. For the associative change due to presynaptic and postsynaptic spike pairs, consider the effect of a single postsynaptic spike at $(x,t)$. The pairing of this spike with the presynaptic spike at $(x_i,t)$ causes a change $\mathcal{L}(x-x_i)$ in $w_i$. To properly handle edge effects, we also include the pairing with presynaptic spikes at $(x_i,t-T)$ and $(x_i,t+T)$, for a total change of

$$\mathcal{L}(x-x_i-T)+\mathcal{L}(x-x_i)+\mathcal{L}(x-x_i+T). \tag{1}$$

For typical biological applications, where $\tau_L \ll T$, at most one of the above terms is non-negligible, but all must be included to handle cases where $x-x_i$ is within $\tau_L$ of $T$ or $-T$ (Fig. 2). In addition, $\tau_L \ll T$ allows us to approximate Eq. (1) by

$$\sum_{n=-\infty}^{\infty} \mathcal{L}(x-x_i-nT) = \mathring{\mathcal{L}}(x-x_i), \tag{2}$$

where $\mathring{\mathcal{L}}(s)=\Sigma_{n=-\infty}^{\infty}\mathcal{L}(s-nT)$ is the periodization of $\mathcal{L}$ with period $T$.

Quantity (2) is the change in $w_i(t)$ due to a single postsynaptic spike at $(x,t)$. Postsynaptic spikes between $t$ and $t+T$ occur at a mean rate density $f(U(x,t))$; hence the mean total change due to all postsynaptic spikes between $t$ and $t+T$ is

$$\int_0^T dx\ f(U(x,t))\mathring{\mathcal{L}}(x-x_i).$$

The mean total change in $w_i(t)$ due to both nonassociative and associative learning is therefore

$$\langle \Delta w_i(t)\rangle = \alpha + \int_0^T dx\ f(U(x,t))\mathring{\mathcal{L}}(x-x_i). \tag{3}$$

We now compute the postsynaptic potential $U(x,t)$. The contribution to $U(x,t)$ due to the presynaptic spike by neuron $i$ at $(x_i,t-nT)$ is $w_i(t+nT)\mathcal{E}(x-x_i+nT)$. For $\tau_E \ll T$ this quantity is non-negligible for at most one value of $n$, either $n=0$ (current period) or $n=-1$ (previous period). But to properly handle edge effects (Fig. 3) we include both, for a total contribution of

$$w_i(t-T)\mathcal{E}(x-x_i-T)+w_i(t)\mathcal{E}(x-x_i). \tag{4}$$

We assume that the learning rate is sufficiently slow so that we may approximate quantity (4) by

$$w_i(t)[\mathcal{E}(x-x_i-T)+\mathcal{E}(x-x_i)]. \tag{5}$$

Finally, $\tau_E \ll T$ allows us to approximate quantity (5) by

$$w_i(t)\sum_{n=-\infty}^{\infty} \mathcal{E}(x-x_i-nT) = w_i(t)\mathring{\mathcal{E}}(x-x_i), \tag{6}$$

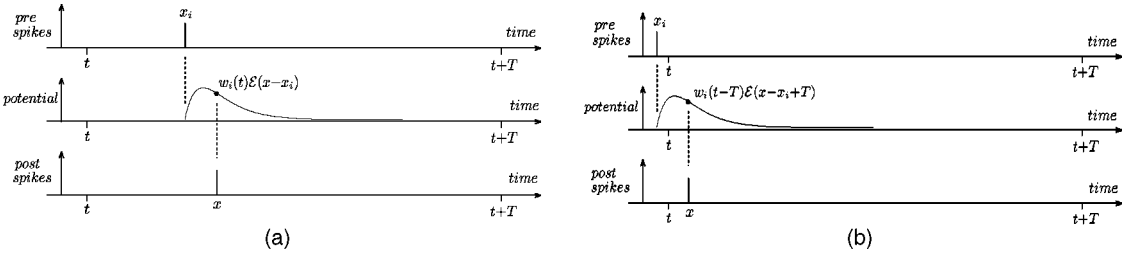


FIG. 3. Postsynaptic potential due to presynaptic spikes. (a) Potential at time $(x,t)$ due to presynaptic spike by neuron $i$ at time $(x_i,t)$ is $w_i(t)\mathcal{E}(x-x_i)$. (b) For $x$ within $\tau_E$ of 0, we must include the potential due to presynaptic spikes in the preceding period. The potential at time $(x,t)$ due to the presynaptic spike by neuron $i$ at time $(x_i,t-T)$ is $w_i(t-T)\mathcal{E}(x-x_i+T)$. Arbitrary units.

where $\mathring{\mathcal{E}}(s)=\Sigma_{n=-\infty}^{\infty}\mathcal{E}(s-nT)$ is the periodization of $\mathcal{E}$ with period $T$.

Quantity (6) is the contribution to $U(x,t)$ from neuron $i$. The total postsynaptic potential is the summed contribution from all presynaptic neurons, plus the repeated external input:

$$U(x,t)=\phi(x)+\sum_{j=1}^{N} w_j(t)\mathring{\mathcal{E}}(x-x_j). \tag{7}$$

Equations (3) and (7) define the mean weight dynamics. The common periodicity of the functions $\mathring{\mathcal{E}}$, $\mathring{\mathcal{L}}$, and $\phi$ is an important feature, allowing the systematic use of Fourier techniques.

## IV. THE NEGATIVE IMAGE

### A. Existence of negative image states

A set of weights $\{w_i\}$ for which the total postsynaptic potential $U(x,t)$ is approximately constant in $x$ will be referred to as an *approximate negative image* state. For such a state the contribution to the postsynaptic potential due to the presynaptic cells alone is, up to an additive constant $U_0$, an approximate negative image (Fig. 4) of the external input $\phi$:

$$\sum_{j=1}^{N} w_j\mathring{\mathcal{E}}(x-x_j)\simeq U_0-\phi(x). \tag{8}$$

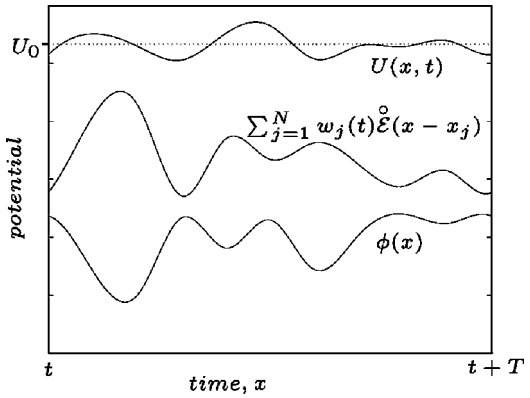


FIG. 4. An approximate negative image. If the postsynaptic potential $U(x,t)=\phi(x)+\Sigma_{j=1}^{N}w_j(t)\mathring{\mathcal{E}}(x-x_j)$ is approximately some constant $U_0$, then the potential $\Sigma_{j=1}^{N}w_j(t)\mathring{\mathcal{E}}(x-x_j)$ due to presynaptic spikes alone is approximately $U_0-\phi(x)$. Arbitrary units.

In the following, we first show that approximate negative image states exist provided a certain condition holds on the Fourier coefficients of the postsynaptic potential function $\mathring{\mathcal{E}}$ and the repeated external input $\phi$, and provided the presynaptic spike time spacing $\delta$ is sufficiently small. We then show that for a particular value of $U_0$ (depending on $\alpha$, $\mathring{\mathcal{L}}$, and $f$) there exists an approximate negative image state which is also an equilibrium (fixed point) for the weight dynamics.

For generic $\mathring{\mathcal{E}}$ and $\phi$, Eq. (8) cannot be made an exact equality for all $x$, because that would require solving infinitely many independent linear equations (one for each $x$) in only finitely many unknowns (the N weights $\{w_j\}$). But if we replace the discrete set of weights $w_j$ with a continuum weight density $\mathcal{W}$, then the analog of Eq. (8) can, under certain conditions, be made exact for all $x$. Given such a density, we then recover the biological case of discrete weights $\{w_j\}$ for which Eq. (8) is approximately true by defining the set $\{w_j\}$ to be a discrete approximation to $\mathcal{W}$.

Let $\mathcal{W}(y)$ be a weight density, with $\mathcal{W}(y)dy$ being the total weight for presynaptic spikes occurring between $y$ and $y+dy$, for $y\in[0,T)$. The continuum analog of Eq. (8), with exact equality for all $x$, is

$$\int_0^T dy\ \mathcal{W}(y)\mathring{\mathcal{E}}(x-y)=U_0-\phi(x). \tag{9}$$

To solve this equation for $\mathcal{W}$ we take the Fourier decomposition. Let $W_n=(1/T)\int_0^T dy\ e^{ik_n y}W(y)$ for $k_n=2\pi n/T$, $n\in\mathbb{Z}$ be the Fourier coefficients for $\mathcal{W}$, and let $E_n$, $\phi_n$ be the coefficients for $\mathring{\mathcal{E}}$ and $\phi$. Then Eq. (9) becomes

$$\begin{aligned}
&U_0-\sum_{n=-\infty}^{\infty}\phi_n e^{-ik_n x}\\
&\quad=\int_0^T dy\left(\sum_{n=-\infty}^{\infty}W_n e^{-ik_n y}\right)\left(\sum_{m=-\infty}^{\infty}E_m e^{-ik_m(x-y)}\right)\\
&\quad=\sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty}W_n E_m e^{-ik_m x}\int_0^T dy e^{i(k_m-k_n)y}\\
&\quad=T\sum_{n=-\infty}^{\infty}W_n E_n e^{-ik_n x}.
\end{aligned}$$

Hence $\mathcal{W}$ satisfies Eq. (9) if and only if

$$W_0 = \frac{U_0 - \phi_0}{TE_0},$$

$$W_n = \frac{-\phi_n}{TE_n}, \quad n \neq 0. \tag{10}$$

Given such a $\mathcal{W}$, we construct approximate negative image states with discrete weights as follows. Define $g(x)$ to be the deviation from a negative image:

$$g(x) = \phi(x) - U_0 - \sum_{j=1}^{N} w_j \mathring{\mathcal{E}}(x - x_j). \tag{11}$$

Then $\{w_j\}$ is an approximate negative image state if $g(x)$ is small relative to $U_0 - \phi(x)$, for all $x$. Consider the set of weights defined by

$$w_j = \delta \mathcal{W}(x_j),$$

where $\delta$ is the spacing of the $x_j$. These weights can be thought of as a discrete approximation to the weight density $\mathcal{W}(y)$. Substituting into Eq. (11) and using Eq. (9) gives

$$g(x) = \sum_{j=1}^{N} \delta \mathcal{W}(x_j) \mathring{\mathcal{E}}(x - x_j) - \int_0^T dy\ \mathcal{W}(y) \mathring{\mathcal{E}}(x - y).$$

This is the difference between a Riemann sum and the integral it approximates. The error theorem for Riemann sums then gives an upper bound for $g$:

$$|g(x)| \leqslant \delta \frac{T}{2} \max_y \left| \frac{d}{dy} [\mathcal{W}(y) \mathring{\mathcal{E}}(x - y)] \right|. \tag{12}$$

Hence, for $|g(x)|$ to be small, we need $\mathcal{W}(y)\mathring{\mathcal{E}}(x-y)$ to be differentiable in $y$, hence we need $\mathcal{W}(y)$ to be differentiable in $y$. A theorem of Fourier series [28] says that $\mathcal{W}(y)$ is differentiable if $\Sigma_{n=-\infty}^{\infty} |n W_n| < \infty$. By Eq. (10) this places a constraint on the Fourier coefficients of $\mathring{\mathcal{E}}$ and $\phi$:

$$\sum_{n=-\infty}^{\infty} \left| \frac{n \phi_n}{E_n} \right| < \infty. \tag{13}$$

This inequality requires $\phi_n$ to go to zero as $n \to \pm\infty$ more rapidly than $E_n / n^2$. In particular, the high frequency (large $|n|$) spectral content of $\phi$ must be less than the high frequency content of $\mathring{\mathcal{E}}$. Intuitively, in order for the convolution of $\mathring{\mathcal{E}}$ with a smooth weight density $\mathcal{W}$ to be able to "match" the high frequency components of $-\phi$, the high frequency content of $\phi$ cannot be too large.

If Eq. (13) is satisfied, and $\delta$ is sufficiently small, then from Eq. (12) the deviation $g(x)$ from an exact negative image is small, hence approximate negative image states exist.

### B. Existence of negative image equilibria

We now show that for a particular $U_0$ there exists an approximate negative image state that is an equilibrium for the weight dynamics. From Eq. (3), a weight state $\{w_j\}$ is an equilibrium if $U(x) = \phi(x) + \Sigma_{j=1}^{N} w_j \mathring{\mathcal{E}}(x - x_j)$ satisfies

$$\alpha + \int_0^T dx\ f(U(x)) \mathring{\mathcal{L}}(x - x_i) = 0 \quad \text{for all } i. \tag{14}$$

This is a system of $N$ equations in the $N$ unknowns $\{w_j\}$, but they are nonlinear equations for nonlinear $f$. In general such equations need not have solutions, but for approximate negative image states the nonlinearity is in some sense "small," and this will allow us to show that solutions exist provided $\delta$ is sufficiently small.

For an approximate negative image state we have $U(x) = U_0 + g(x)$ with $g(x) \ll U_0$, and we wish this $U(x)$ to satisfy Eq. (14). First define $U_0$ so that Eq. (14) would be satisfied if $g(x)$ were identically zero:

$$\alpha + \int_0^T dx\ f(U_0) \mathring{\mathcal{L}}(x - x_i) = 0 \quad \text{for all } i. \tag{15}$$

This requires

$$f(U_0) = \frac{-\alpha}{\int_0^T dx\ \mathring{\mathcal{L}}(x - x_i)} \quad \text{for all } i$$

$$= \frac{-\alpha}{\int_0^T dx\ \mathring{\mathcal{L}}(x)}, \tag{16}$$

where the independence of $i$ follows from the periodicity of $\mathring{\mathcal{L}}$. Hence, our desired $U_0$ exists and is given by

$$U_0 = f^{-1}\left( \frac{-\alpha}{\int_0^T dx\ \mathring{\mathcal{L}}(x)} \right), \tag{17}$$

provided $\alpha$, $\mathring{\mathcal{L}}$, and $f$ satisfy

$$\min_u f(u) < \frac{-\alpha}{\int_0^T dx\ \mathring{\mathcal{L}}(x)} < \max_u f(u). \tag{18}$$

From Eq. (15), $U(x) = U_0 + g(x)$ satisfies Eq. (14) if and only if

$$\int_0^T dx\ [f(U_0 + g(x)) - f(U_0)] \mathring{\mathcal{L}}(x - x_i) = 0 \quad \text{for all } i. \tag{19}$$

For brevity let $h(x) = f(U_0 + g(x)) - f(U_0)$ and $L_i(x) = \mathring{\mathcal{L}}(x - x_i)$. Then Eq. (19) can be written as

$$\langle h, L_i \rangle = 0 \quad \text{for all } i, \tag{20}$$

where $\langle \cdot, \cdot \rangle$ is the inner product defined by

$$\langle f_1, f_2 \rangle = \int_0^T dx\ f_1(x) f_2(x), \tag{21}$$

for $f_1$, $f_2$ in the space $X$ of smooth functions on the interval $[0,T]$.

Let $H$ be the set of functions $h$ corresponding to all possible values of the weights $\{w_j\}$:

$$H = \left\{ h : h(x) = f\left( \phi(x) - \sum_{j=1}^{N} w_j \mathring{\mathcal{E}}(x - x_j) \right) - f(U_0), w_j \in \mathbb{R}, j = 1, \ldots, N \right\},$$

where we have used $U_0 + g(x) = \phi(x) - \Sigma_{j=1}^{N} w_j \mathring{\mathcal{E}}(x - x_j)$ from Eq. (11). Let $S$ be the subspace of $X$ consisting of all linear combinations of the $\{L_i\}$, and $S^\perp$ be the (infinite-dimensional) subspace of $X$ orthogonal to $S$ in the inner product defined by Eq. (21). Then there exists an $h$ satisfying Eq. (20) if and only if $H$ and $S^\perp$ have nonempty intersection:

$$H \cap S^\perp \neq \varnothing. \tag{22}$$

We claim that condition (22) holds if $\delta$ is sufficiently small. If $\delta$ is small, then bound (12) implies that $g(x)$ is small for all $x$. In that case $h(x) = f(U_0 + g(x)) - f(U_0)$ is approximately its linearization in $g$, which we denote by $h_0(x)$:

$$h(x) \simeq h_0(x) = f'(U_0) g(x) = f'(U_0) \left[ \phi(x) - U_0 + \sum_{j=1}^{N} w_j \mathring{\mathcal{E}}(x - x_j) \right].$$

Let $H_0$ be the set of such $h_0$ corresponding to all possible values of the weights $\{w_j\}$:

$$H_0 = \left\{ h_0 : h_0(x) = f'(U_0) \left[ \phi(x) - U_0 + \sum_{j=1}^{N} w_j \mathring{\mathcal{E}}(x - x_j) \right], w_j \in \mathbb{R}, j = 1, \ldots, N \right\},$$

Then the condition that $H_0$ have nonempty intersection with $S^\perp$,

$$H_0 \cap S^\perp \neq \varnothing, \tag{23}$$

is equivalent to existence of $h_0 \in H_0$ such that $\langle h_0, L_i \rangle = 0$ for all $i$. This is equivalent to the linearization of system (19):

$$\int_0^T dx\ f'(U_0) \left[ \phi(x) - U_0 + \sum_{j=1}^{N} w_j \mathring{\mathcal{E}}(x - x_j) \right] \mathring{\mathcal{L}}(x - x_i) = 0 \quad \text{for all } i,$$

which can be rewritten as

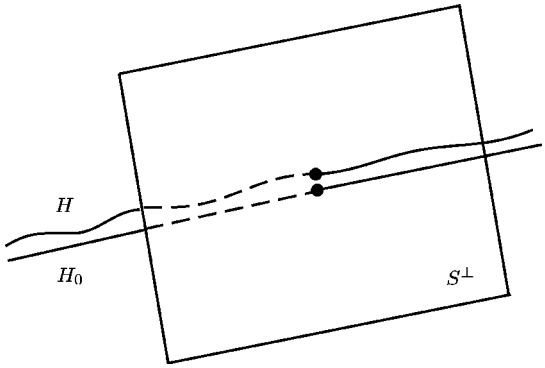


FIG. 5. Transversal intersection theorem. If $H_0$ has transversal intersection with $S^\perp$ and $H$ is sufficiently close to $H_0$, then $H$ intersects $S^\perp$.

$$\sum_{j=1}^{N} Q_{ij} w_j = \gamma, \tag{24}$$

where $\gamma = \int_0^T dx\ f'(U_0)[\phi(x) - U_0]$ and

$$Q_{ij} = f'(U_0) \int_0^T dx\ \mathring{\mathcal{E}}(x - x_j) \mathring{\mathcal{L}}(x - x_i). \tag{25}$$

This is a system of $N$ linear inhomogeneous equations in the $N$ unknowns $\{w_j\}$, which has a solution provided the coefficient matrix $Q$ is invertible. The eigenvalues of $Q$ will be calculated in the following section, and for generic $\mathcal{E}$ and $\mathcal{L}$ all eigenvalues are nonzero. Hence $Q$ is generically invertible, so that condition (23) holds. Furthermore, the intersection of $H_0$ with $S^\perp$ is generically transversal (not tangent).

Now as $\delta \to 0$, $H \to H_0$ in the metric induced by inner product (21). By the openness of transversal intersection (infinite-dimensional version [29]), any sufficiently small perturbation of $H_0$ also intersects $S^\perp$. Hence for sufficiently small $\delta$, $H$ intersects $S^\perp$ (Fig. 5), hence $h$ satisfying Eq. (20) exists. The corresponding weight state $\{w_j\}$ is an approximate negative image equilibrium.

## V. STABILITY CRITERION

We now derive a necessary and sufficient condition for the mean stability of approximate negative image equilibria, by examining the linearized weight dynamics around such states. Let $\{\hat{w}_j\}$ be an approximate negative image equilibrium satisfying Eq. (11) with

$$U(x) = \phi(x) + \sum_{j=1}^{N} \hat{w}_j \mathring{\mathcal{E}}(x - x_j) = U_0 + \hat{g}(x). \tag{26}$$

Solving for $\phi(x)$ in Eq. (26) and substituting into Eq. (7) yields

$$U(x,t) = U_0 + \hat{g}(x) + \sum_{j=1}^{N} v_j(t) \mathring{\mathcal{E}}(x - x_j),$$

where $v_j(t)=w_j(t)-\hat{w}_j$ is the deviation of weight $j$ from its equilibrium value and $\hat{g}(x)$ is the deviation from a negative image in the equilibrium state $\{\hat{w}_j\}$. To first order in $v_j$ we then have

$$f(U(x,t))\simeq f(U_0+\hat{g}(x))+f'(U_0+\hat{g}(x))\times\sum_{j=1}^{N} v_j(t)\mathring{\mathcal{E}}(x-x_j). \tag{27}$$

Substituting Eq. (27) into Eq. (3) and using $\triangle w_i(t)=\Delta v_i(t)$ yields

$$\langle\Delta v_i(t)\rangle=\alpha+\int_0^T dx\ f(U_0+\hat{g}(x))\mathring{\mathcal{L}}(x-x_i)+f'(U_0+\hat{g}(x))\sum_{j=1}^{N} v_j(t)\mathring{\mathcal{E}}(x-x_j)\mathring{\mathcal{L}}(x-x_i).$$

From the equilibrium condition, Eqs. (14) and (26), the term in Eq. (28) of zeroth order in $v_j$ vanishes. Hence

$$\langle\Delta v_i(t)\rangle\simeq\sum_{j=1}^{N} P_{ij}v_j(t),$$

where

$$P_{ij}=\int_0^T dx\ f'(U_0+\hat{g}(x))\mathring{\mathcal{E}}(x-x_j)\mathring{\mathcal{L}}(x-x_i).$$

Now assume $\delta$ is sufficiently small so that

$$\hat{g}(x)\ll\frac{f'(U_0)}{f''(U_0)}\quad\text{for all } x.$$

Then $f'(U_0+\hat{g}(x))\simeq f'(U_0)$ for all $x$, so that $P_{ij}\simeq Q_{ij}$, where $Q$ is the matrix defined in Eq. (25), and we obtain

$$\langle\Delta v_i(t)\rangle\simeq\sum_{j=1}^{N} Q_{ij}v_j(t).$$

Taking the mean on both sides and using $\langle\Delta v_i(t)\rangle=\Delta\langle v_i(t)\rangle$ yields

$$\Delta\langle v_i(t)\rangle\simeq\sum_{j=1}^{N} Q_{ij}\langle v_i(t)\rangle. \tag{28}$$

Equation (28) gives the linearized dynamics for $\langle v_i(t)\rangle$ near $\langle v_i(t)\rangle=0$, hence for $\langle w_i(t)\rangle$ near the approximate negative image equilibrium $\{\hat{w}_j\}$.

The system, Eq. (28), is stable if and only if all eigenvalues of $Q+I$ have norm less than 1. Due to periodicity of $\mathring{\mathcal{E}},\mathring{\mathcal{L}}$ and regular spacing of the $\{x_i\}$, the matrix $Q$ has the property that each of its rows equals the row above it shifted one entry to the right (and wrapped around at the edges). Such matrices are called *circulant* [30] and their eigenvectors and eigenvalues are easily found, as follows. If $u$ is the vector with components $u_i=e^{ikx_i}$, then

$$(Qu)_i=\sum_{j=1}^{N} Q_{ij}u_j=\sum_{j=1}^{N} e^{ikx_j}\int_0^T dx\ \mathring{\mathcal{E}}(x-x_j)\mathring{\mathcal{L}}(x-x_i),$$

so that

$$\frac{(Qu)_i}{u_i}=\sum_{j=1}^{N} e^{ik(x_j-x_i)}\int_0^T dx\ \mathring{\mathcal{E}}(x-x_j)\mathring{\mathcal{L}}(x-x_i). \tag{29}$$

By periodicity of $\mathring{\mathcal{E}}$ and $\mathring{\mathcal{L}}$, the integral in the above expression is a function of $x_j-x_i$ modulo $T$. The factor $e^{ik(x_j-x_i)}$ is also a function of $x_j-x_i$ modulo $T$ provided $e^{ikT}=1$. Now if the $\{x_i\}$ are regularly spaced, the sum over $j$ of a function of $x_j-x_i$ modulo $T$ is independent of $i$. In that case Eq. (29) would imply that $(Qu)_i/u_i$ is independent of $i$, hence $u$ is an eigenvector of $Q$, with eigenvalue equal to the right-hand side of Eq. (29). We get a complete set of such eigenvectors by taking $N$ values of $k$ such that $e^{ikT}=1$ and the functions $e^{ikx_i}$ are independent functions of $i$. Here we choose

$$k_n=\frac{2\pi n}{T},\quad n=0,1,\ldots,N-1.$$

The corresponding eigevalues of $Q$ are

$$\lambda_n=\sum_{j=1}^{N} e^{ik_n(x_j-x_i)}\int_0^T dx\ \mathring{\mathcal{E}}(x-x_j)\mathring{\mathcal{L}}(x-x_i).$$

Letting $z_j=x_j-x_i$, making the change of variables $y=x_j-x$, and using periodicity of $\mathring{\mathcal{E}}$, $\mathring{\mathcal{L}}$ gives

$$\lambda_n=\sum_{j=1}^{N} e^{ik_nz_j}\int_0^T dy\ \mathring{\mathcal{E}}(-y)\mathring{\mathcal{L}}(z_j-y)=\sum_{j=1}^{N} e^{ik_nx_j}\mathring{\mathcal{L}}*_T\tilde{\mathring{\mathcal{E}}}(x_j),$$

where $*_T$ is convolution on the interval $[0,T]$, $\tilde{}$ is horizontal reflection $[\tilde{\mathring{\mathcal{E}}}(y)=\mathring{\mathcal{E}}(-y)]$, and since $\{z_j\}=\{x_j\}$ we have replaced $z_j$ by $x_j$ in the sum.

The stability condition is $|1+\lambda_n|<1$ for all $n$:

$$\text{Stability:}\left|1+\sum_{j=1}^{N} e^{ik_nx_j}\mathring{\mathcal{L}}*_T\tilde{\mathring{\mathcal{E}}}(x_j)\right|<1,$$

$$k_n=\frac{2\pi n}{T},\quad n=0,1,\ldots,N-1. \tag{30}$$

In the biological setting two limiting regimes are of special interest: slow learning ($\mathring{\mathcal{L}}$ small) and dense spacing ($\delta$ small).

If $\mathring{\mathcal{L}}$ is small, then so are the eigenvalues of $Q$. If $\lambda_n=a_n+ib_n$ with $a_n,b_n$ real, we have

$$|1+\lambda_n|^2=(1+a_n)^2-b_n^2=1+2a_n+(a_n^2-b_n^2).$$

If $a_n$ and $b_n$ are sufficiently small, this quantity is less than 1 if and only if $a_n<0$. Hence for sufficiently small $\mathring{\mathcal{L}}$, all eigenvalues of $Q+I$ have norm less than 1 if and only if all eigenvalues of $Q$ have negative real part.[2] The stability condition then becomes $\mathrm{Re}\,\lambda_n<0$ for all $n$. Hence in the slow learning limit, Eq. (30) becomes

$$\text{Slow learning:} \quad \mathrm{Re}\sum_{j=1}^{N} e^{ik_n x_j}\mathring{\mathcal{L}}_{*T}\tilde{\mathring{\mathcal{E}}}(x_j)<0,$$

$$n=0,1,\ldots,N-1. \tag{31}$$

The dense spacing limit ($\delta\to 0$) is the continuum limit in $x_i$. The discrete weight density $w_i/T$ is replaced by a continuum weight density $W(x)$, sums over $x_j$ are replaced by integrals over $x$, and $N\to\infty$. This yields

$$\lambda_n=\int_0^T dx\; e^{ik_n x}\mathring{\mathcal{L}}_{*T}\tilde{\mathring{\mathcal{E}}}(x), \quad n=0,1,\ldots\ .$$

Hence $\lambda_n$ is just the $n$th Fourier coefficient of $\mathring{\mathcal{L}}_{*T}\tilde{\mathring{\mathcal{E}}}$. The Fourier convolution theorem then gives

$$\lambda_n=\mathring{L}_n\tilde{\mathring{E}}_n=\mathring{L}_n\overline{\mathring{E}_n},$$

where $\mathring{E}_n,\tilde{\mathring{E}}_n,\mathring{L}_n$ are the $n$th Fourier coefficients of $\mathring{E},\tilde{\mathring{E}},\mathring{L}$, respectively, and $\overline{z}$ is the complex conjugate of $z$. Substituting into the stability condition $|1+\lambda_n|<1$ for all $n$ gives the dense spacing limit of Eq. (30):

$$\text{Dense spacing:} \quad |1+\mathring{L}_n\overline{\mathring{E}_n}|<1, \quad n=0,1,\ldots. \tag{32}$$

Finally, with both slow learning and dense spacing the stability condition becomes

$$\text{Slow learning, dense spacing:} \quad \mathrm{Re}[\mathring{L}_n\overline{\mathring{E}_n}]<0,$$

$$n=0,1,\ldots. \tag{33}$$

A further simplification follows in the long period limit, $\tau_E,\tau_L\ll T$. Holding $\tau_E,\tau_L$ constant and taking $T\to\infty$, the Fourier series of $\mathring{\mathcal{E}},\mathring{\mathcal{L}}$ in Eq. (33) approach Fourier transforms of $\mathcal{E},\mathcal{L}$, respectively. The stability condition then becomes

Slow learning, dense spacing, long period:

$$\mathrm{Re}[\mathcal{F}[\mathcal{L}](k)\overline{\mathcal{F}[\mathcal{E}](k)}]<0, \quad k\in(-\infty,\infty). \tag{34}$$

For the calculation of examples we will work in the slow learning, dense spacing, long period limit, which is the limit of primary biological interest in the mormyrid ELL.

[2]The slow learning limit can thus be thought of as the continuous time (continuous $t$) limit. All eigenvalues of $Q$ having negative real part are equivalent to stability of the system $d\langle v\rangle/dt=Q\langle v\rangle$ and hence of $T\,d\langle v\rangle/dt=Q\langle v\rangle$, which is the continuous time version of Eq. (28).

## VI. GENERAL REMARKS

*The roles of nonassociative and associative learning*. Both nonassociative and associative learnings ($\alpha$ and $\mathcal{L}$, respectively) play a role in whether approximate negative image equilibria exist, via Eq. (18). They are also involved in determining the location of such equilibria, via Eq. (14). The interpretation of Eq. (14) is that at equilibrium, the mean change due to nonassociative learning ($\alpha$) must be precisely opposite to the mean change due to associative learning (the $\mathcal{L}$ term). If the postsynaptic spike rate density $f$ is bounded, this places a relative magnitude constraint on $\alpha$ and $\mathcal{L}$, namely, Eq. (18). If this constraint is violated then the mean changes due to associative and nonassociative learning are unable to balance one another, and no negative image equilibrium is possible.

By contrast, only associative learning plays a role in the stability of approximate negative image equilibria, via Eq. (30). The irrelevance of nonassociative learning for stability has an intuitive interpretation: near an approximate negative image equilibrium, the mean nonassociative change is canceled by the mean associative change due the constant postsynaptic potential $U_0$ around which $U(x,t)$ fluctuates. Only the deviations of $U(x,t)$ from $U_0$ cause a net change in the weights, and these changes are purely associative [due to postsynaptic spikes generated by $U(x,t)$]. Alternatively, nonassociative learning can be analogized to a constant externally applied force in a physical system. Such a force changes the location of equilibria, but has no effect on the dynamics around equilibria.

*The role of the repeated input*. For a given postsynaptic potential response $\mathcal{E}$, the repeated input $\phi$ plays a role in the existence of approximate negative image states via Eq. (13): $\phi$ cannot have too much high frequency content relative to $\mathcal{E}$ for such states to exist.

Assuming $\phi$ is such that approximate negative image states exist, it then plays an important role in determining the weight configurations in such states, and in particular, in approximate negative image equilibria, via Eq. (8).

But $\phi$ plays no role in the stability of the resulting negative image equilibrium. This is intuitively reasonable, since in approximate negative image states, $\phi$ is "nulled out" by the summed postsynaptic potentials due to time-locked presynaptic spikes.

*The role of noise*. The functional form of the mean postsynaptic spike rate $f$ affects the existence and location of the negative image equilibrium via Eqs. (18) and (14).

But provided $f$ is strictly increasing (so that $f'$ is positive), $f$ has no effect on the stability of the equilibrium. Hence the classification of learning rules as (mean) stable or unstable is, except for this mild monotonicity requirement, insensitive to the fine structure of the noise. This is a post hoc justification for not modeling the noise in more detail.

*Canonically stable learning rule*: $\mathcal{L}=-\mathcal{E}$. In the dense spacing and slow learning limit, suppose $\mathcal{L}=-\mathcal{E}$. The stability condition (33) is then

$$|\mathring{E}_n|^2>0 \quad \text{for all } n, \tag{35}$$

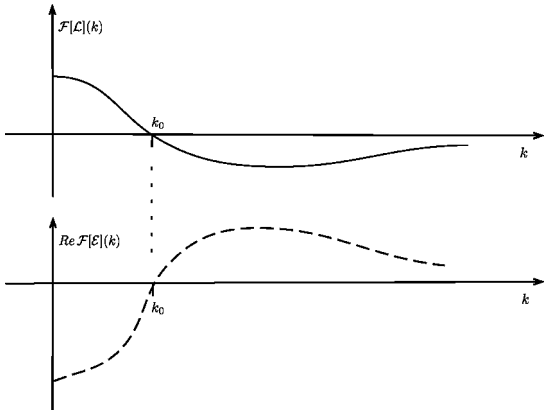


FIG. 6. If $\mathrm{Re}[\mathcal{F}[\mathcal{E}](k)]$ changes sign at $k_0$, then for the product $\mathcal{F}[\mathcal{L}](k)\mathrm{Re}[\mathcal{F}[\mathcal{E}](k)]$ to be negative around $k_0$ we must have $\mathcal{F}[\mathcal{L}](k)$ also change sign at $k_0$, in the opposite sense to $\mathrm{Re}[\mathcal{F}[\mathcal{E}]](k)$. Arbitrary units.

or in other words, $\mathring{E}_n \neq 0$ for all $n$. Since this is true for generic $\mathcal{E}$, the learning rule $\mathcal{L}=-\mathcal{E}$ is generically stable.

*Area sign condition.* In the dense spacing and slow learning limit, consider $n=0$ in Eq. (33). Since $\mathring{L}_0$ and $\overline{\mathring{E}_0}=\mathring{E}_0$ are just the areas under the functions $\mathring{\mathcal{L}}$ and $\mathring{\mathcal{E}}$, Eq. (33) says that for stability, these areas must be opposite in sign. If they are the same sign, then the negative image is unstable. In particular, if $\mathcal{E}$ and $\mathcal{L}$ are both nonnegative, the negative image is unstable. Hence, if $\mathcal{E}$ is an excitatory PSP and $\mathcal{L}$ is any pure potentiating learning rule, the negative image is unstable. Similarly, inhibitory PSPs with purely depressing learning rules are unstable.

*Symmetric and antisymmetric learning rules*. In the dense spacing, slow learning, long period limit, there is a nonempty, positive measure set of postsynaptic response functions for which purely symmetric or purely antisymmetric learning rules are generically unstable. This follows from the fact that the Fourier transforms of symmetric and antisymmetric functions are pure real and pure imaginary, respectively.

Suppose the real part of the Fourier transform of $\mathcal{E}$ has a zero:

$$\mathrm{Re}[\mathcal{F}[\mathcal{E}](k_0)]=0 \quad \text{for some } k_0. \tag{36}$$

Then, generically, $\mathrm{Re}[\mathcal{F}[\mathcal{E}](k)]$ changes sign at $k_0$. Suppose $\mathcal{L}$ is symmetric, so that $\mathcal{F}[\mathcal{L}](k)$ is pure real. Then

$$\mathrm{Re}[\mathcal{F}[\mathcal{L}](k)\overline{\mathcal{F}[\mathcal{E}](k)}]=\mathcal{F}[\mathcal{L}](k)\mathrm{Re}[\mathcal{F}[\mathcal{E}](k)].$$

Since $\mathrm{Re}[\mathcal{F}[\mathcal{E}](k)]$ changes sign at $k_0$, for the stability condition (34) to be satisfied for $k$ near $k_0$, we must have $\mathcal{F}[\mathcal{L}](k)$ change sign at $k_0$, in the opposite sense to $\mathrm{Re}[\mathcal{F}[\mathcal{E}](k)]$; see Fig. 6. But this forces $\mathcal{F}[\mathcal{L}](k_0)=0$, which is untrue for generic symmetric $\mathcal{L}$. Hence, generic symmetric learning rules are unstable for postsynaptic response functions satisfying Eq. (36).

Similarly, if the imaginary part of the Fourier transform of $\mathcal{E}$ has a zero:

$$\mathrm{Im}[\mathcal{F}[\mathcal{E}](k_0)]=0 \quad \text{for some } k_0, \tag{37}$$

then generic antisymmetric learning rules are unstable.

Pure antisymmetric learning rules have another difficulty: since they satisfy $\int_0^T dx\,\mathcal{L}(x)=0$, near a negative image equilibrium the mean weight change per cycle due to an antisymmetric $\mathcal{L}$ is zero, to first order in $g$. The total mean weight change per cycle is therefore approximately $\alpha$. Hence, negative image equilibria for pure antisymmetric learning rules are only possible if $\alpha=0$ (no nonassociative learning).

*Cooperative stability*. It follows from Eq. (30) that the sum of stable learning rules is stable; but it is also clear that given a generic $\mathcal{E}$, there exist pairs of learning rules $\mathcal{L}_1$ and $\mathcal{L}_2$, each individually unstable, for which the sum $\mathcal{L}_1+\mathcal{L}_2$ is stable. This is most easily seen by direct computation in the slow learning, dense spacing, long period limit, via Eq. (34) (see the examples calculated below).

*Duality principle*. Interchanging $\mathcal{L}$ and $\mathcal{E}$ in Eq. (25) transforms $Q_{ij}$ to $Q_{ji}$, hence $Q$ to $Q^T$, hence $Q+I$ to $(Q+I)^T$. The eigenvalues of a real matrix are unchanged by transposition. The stability condition, that all eigenvalues of $Q+I$ have norm less than 1, is thus invariant under interchange of $\mathcal{L}$ and $\mathcal{E}$. In other words, a PSP $\mathcal{E}$ and learning rule $\mathcal{L}$ are a stable pair if and only if the PSP $\mathcal{L}$ and learning rule $\mathcal{E}$ are a stable pair.

This has potential biological relevance if the functional forms of PSPs and associative learning rules overlap. The single-lobe exponential and $\alpha$ function learning rules treated in the examples, below, are also plausible PSPs, hence duality applies.

*Inversion principle*. Replacing $\mathcal{E}$ with $-\mathcal{E}$ and $\mathcal{L}$ by $-\mathcal{L}$ in Eq. (25) leaves $Q_{ij}$ invariant, hence $Q+I$ invariant. The stability condition is therefore invariant under inversion of both $\mathcal{E}$ and $\mathcal{L}$. In other words, a PSP $\mathcal{E}$ and learning rule $\mathcal{L}$ are a stable pair if and only if the PSP $-\mathcal{E}$ and learning rule $-\mathcal{L}$ are a stable pair.

In particular, the stable learning rules for an inhibitory PSP are just minus the stable learning rules for the corresponding excitatory PSP. Plasticity at inhibitory synapses was explored in Ref. [31], and preliminary experimental evidence was given in Ref. [32].

*Independence of normalization.* In the slow learning and dense spacing limit, the stability conditions, Eq. (33) or Eq. (34), are invariant under multiplication of $\mathcal{L}$ or $\mathcal{E}$ by positive constants. Hence, provided the magnitudes of $\mathcal{L}$ or $\mathcal{E}$ are not so large that the slow learning assumption is violated, stability does not depend on those magnitudes. In particular, in working with specific examples it is not necessary to give $\mathcal{L}$ or $\mathcal{E}$ any overall normalization.

## VII. EXAMPLES

Working in the slow learning, dense spacing, long period limit, we now compute explicit criteria for stability when $\mathcal{E}$ and $\mathcal{L}$ have functional forms commonly used in the spike-timing-dependent plasticity literature. The PSP $\mathcal{E}$ will be assumed excitatory and causal, and of exponential or alpha function form. The learning rule $\mathcal{L}$ will consist of one or two ''lobes'': a ''pre-before-post'' lobe (presynaptic spike before

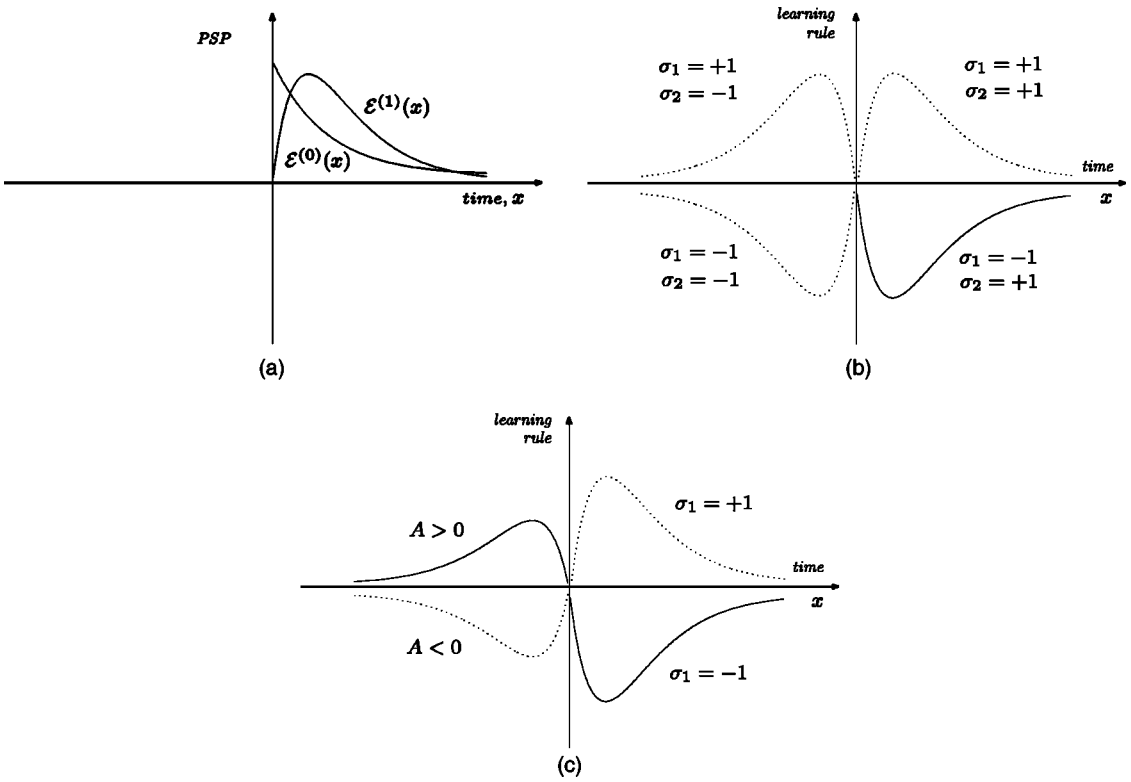


FIG. 7. PSPs and learning rules used in the examples. (a) The PSP $\mathcal{E}^{(p_E)}(x)$ is exponential for $p_E=0$ and an $\alpha$ function for $p_E=1$. (b) One-lobe learning rules of alpha function form, $\mathcal{L}_I^{(1)}(x)$, for the four possible combinations of $\sigma_1$ (potentiating or depressing) and $\sigma_2$ (pre-before-post or post-before-pre). (c) Two-lobe learning rules of alpha function form, $\mathcal{L}_{II}^{(1)}(x)$, for the four possible combinations of $\sigma_1$ (pre-before-post lobe potentiating or depressing) and the sign of $A$ (post-before-pre lobe potentiating or depressing). The area of the pre-before-post lobe is normalized to $\pm 1$, and the area of the post-before-pre lobe is $A$. Arbitrary units.

postsynaptic spike) and/or a "post-before-pre" lobe (postsynaptic spike before presynaptic spike). Each lobe will be of exponential or alpha function form, and either potentiating (positive) or depressing (negative). Such $\mathcal{E}$ and $\mathcal{L}$ can be written as follows:

$$\mathcal{E}^{(p_E)}(x)=x^{p_E}e^{-x/\tau_E}H(x),$$

$$\mathcal{L}_I^{(p_L)}(x)=\sigma_1(\sigma_2 x)^{p_L}e^{-\sigma_2 x/\tau_L}H(\sigma_2 x) \quad \text{(one lobe)},$$

$$\mathcal{L}_{II}^{(p_L)}(x)=\frac{\sigma_1}{\tau_{L_1}^{p_L+1}}x^{p_L}e^{-x/\tau_{L_1}}H(x)+\frac{A}{\tau_{L_2}^{p_L+1}}(-x)^{p_L} \times e^{x/\tau_{L_2}}H(-x), \quad \text{(two lobe)}$$

where $H$ is the Heaviside function

$$H(x)=\begin{cases}1, & x\geqslant 0\\ 0, & x<0.\end{cases}$$

The parameters (see Fig. 7) are as follows: $p_E, p_L=0$ for an exponential or 1 for an alpha function; $\tau_E$, $\tau_L$, $\tau_{L_1}$, and $\tau_{L_2}$ are positive time constants; $\sigma_1=+1$ for a potentiating lobe or $-1$ for a depressing lobe; $\sigma_2=+1$ for a pre-before-post lobe or $-1$ for a post-before-pre lobe; and for the two lobe $\mathcal{L}$, $A$ is the area of the post-before-pre lobe, with the area of the pre-before-post lobe normalized to $\pm 1$ . We impose no overall normalization on $\mathcal{E}$ or $\mathcal{L}$, since this has no effect on stability.

We assume an excitatory PSP $\mathcal{E}$; to obtain the stable cases for the inhibitory PSP $-\mathcal{E}$, simply replace $\mathcal{L}$ with $-\mathcal{L}$ in the stable cases for $\mathcal{E}$ (i.e. replace $\sigma_1$ with $-\sigma_1$ and $A$ with $-A$).

For both the one lobe and two lobe $\mathcal{L}$, there are four possible combinations of $p_E$ and $p_L$: exponential or alpha function PSP with exponential or alpha function learning rule. We will refer to these four cases as ee, ea, ae, and aa, with the first letter in the pair indicating that the PSP is exponential or alpha function and the second letter referring to the learning rule.

The Fourier transforms of these $\mathcal{E}$ and $\mathcal{L}$ are rational functions in $k$, and the stability condition will reduce to the requirement that a certain polynomial in $k^2$, whose coefficients are themselves polynomials in the parameters of $\mathcal{E}$ and $\mathcal{L}$, be negative for all $k$. Since the algebra in all cases is essentially the same, differing only in the size and coefficients of the resulting polynomial, we present only one case (aa, one lobe) in full detail and for all other cases simply list the end results.

### A. Function PSP, one-lobe $\alpha$ function learning rule

For

$$\mathcal{E}(x)=xe^{-x/\tau_E}H(x),$$

$$\mathcal{L}(x)=\sigma_1\sigma_2 x e^{-\sigma_2 x/\tau_L}H(\sigma_2 x),$$

we have

$$\mathcal{F}[\mathcal{E}](k)=\frac{\tau_E^2}{(1-ikE)^2},$$

$$\mathcal{F}[\mathcal{L}](k)=\frac{\sigma_1\tau_L^2}{(1-\sigma_2 ik\tau_L)^2}$$

leading to

$$\begin{aligned}\mathrm{Re}\{\mathcal{F}[\mathcal{L}](k)\overline{\mathcal{F}[\mathcal{E}](k)}\}&=C\,\mathrm{Re}[\sigma_1(1+i\sigma_2 k\tau_L)^2(1-ik\tau_E)^2]\\&=C\sigma_1[\sigma_2^2\tau_L^2\tau_E^2k^4+(4\sigma_2\tau_L\tau_E-\sigma_2^2\tau_L^2\\&\quad-\tau_E^2)k^2+1],\end{aligned}$$

where $C=\tau_L^2\tau_E^2/[(1+\sigma_2^2\tau_L^2k^2)^2(1+\tau_E^2k^2)^2]$. Since $C>0$, the stability condition is then

$$\sigma_1[\sigma_2^2r^2k^4+(4\sigma_2 r-\sigma_2^2r^2-1)k^2+1]<0 \quad \text{for all } k, \tag{38}$$

where $r=\tau_L/\tau_E$. The expression on the left is a quadratic in $k^2$. The condition is impossible for $\sigma_1=+1$ (it fails at $k=0$) but for $\sigma_1=-1$ more work is required. The quadratic $ax^2+bx+c$ is negative for all $x\geqslant 0$ if and only if $a<0$ and ($b<0$ or $b^2-4ac<0$). Applying this condition to Eq. (38) with $\sigma_1=-1$ yields

$$r^2-4\sigma_2 r+1<0 \tag{39}$$

or

$$(r-\sigma_2)^2(r^2-6\sigma_2 r+1)<0 \tag{40}$$

For $\sigma_2=-1$ these give $-2-\sqrt{3}<r<-2+\sqrt{3}$ or $-3-2\sqrt{2}<r<-3+2\sqrt{2}$, both of which are impossible because $r>0$. For $\sigma_2=+1$ we get $2-\sqrt{3}<r<2+\sqrt{3}$ or $3-2\sqrt{2}<r<3+2\sqrt{2}$; the former is contained in the latter, giving stability if and only if

$$\sigma_1=-1,\ \ \sigma_2=+1,$$

$$3-2\sqrt{2}<\frac{\tau_L}{\tau_E}<3+2\sqrt{2}. \tag{41}$$

The only stable case is depressive and pre-before-post, with $\tau_L/\tau_E$ constrained to lie in a finite interval (Fig. 8). Note that this interval contains $\tau_L/\tau_E=1$, where $\mathcal{L}=-\mathcal{E}$ (the canonically stable learning rule).

Duality is also applicable here. Interchanging $\mathcal{E}$ and $\mathcal{L}$ in this example is equivalent to interchanging $\tau_E$ and $\tau_L$ and multiplying both $\mathcal{E}$ and $\mathcal{L}$ by $-1$. The multiplications offset and we are left with $r$ replaced by $1/r$. It follows that if the interval of stability for the pair $(\mathcal{E},\mathcal{L})$ is $s_1<r<s_2$, the corresponding interval for the pair $(\mathcal{L},\mathcal{E})$ is $s_1<1/r<s_2$. But by duality these intervals must coincide; hence we must have $s_1=1/s_2$. This is indeed the case for $s_1=3-2\sqrt{2}$ and $s_2=3+2\sqrt{2}$.

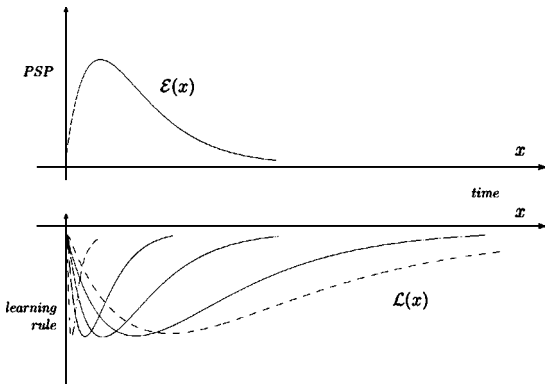


FIG. 8. Range of stable one lobe $\mathcal{L}$ for given $\mathcal{E}$, in case aa. The learning rule must be depressive and pre-before-post, with $3-2\sqrt{2}<\tau_L/\tau_E<3+2\sqrt{2}$. Stable examples are drawn with solid lines; endpoints of the stable interval are drawn with dashed lines. Arbitrary units.

The instability of the $\sigma_1=+1$ case for any $\tau_E$ and $\tau_L$, by the failure of the stability condition at $k=0$, is just the area sign condition.

### B. Summary of Results

For the one-lobe learning rules the stable parameter ranges are all easily calculated:

ee: $\sigma_1=-1,\sigma_2=+1$, all $\tau_L/\tau_E$

ea: $\sigma_1=-1,\sigma_2=+1$, $\tau_L/\tau_E<2$

ae: $\sigma_1=-1,\sigma_2=+1$, $\tau_L/\tau_E>1/2$

aa: $\sigma_1=-1,\sigma_2=+1$, $2-\sqrt{3}<\tau_L/\tau_E<2+\sqrt{3}$.

Note that in all four cases we get instability, for all $\tau_L$ and $\tau_E$, if $\mathcal{L}$ is not depressive and pre-before-post. For $\mathcal{L}$ depressive and pre-before-post, all four cases have some range of $\tau_L/\tau_E$ in which $\mathcal{L}$ is stable. The extent of that range depends critically on the precise functional form of $\mathcal{E}$ and $\mathcal{L}$; but for $1/2<\tau_L/\tau_E<2$ we have stability independent of the functional form of $\mathcal{E}$ and $\mathcal{L}$.

For the two-lobe learning rules the polynomial arising out of the stability condition has coefficients depending on $\sigma_1$ and on three continuous parameters: $r_1$, $r_2$, and $A$, where $r_1=\tau_{L_1}/\tau_E$, $r_2=\tau_{L_2}/\tau_E$. The polynomials are given in the Appendix.

In all four cases, $\sigma_1=+1$ is always unstable. For $\sigma_1=-1$, the boundaries of the stable region in $(r_1,r_2)$ for various values of $A$ are plotted numerically in Fig. 9.

For one-lobe learning rules we found that only depressive and pre-before-post permits stability. For two-lobe learning rules, the pre-before-post lobe must be depressive for stability, and the post-before-pre lobe cannot have area $A$ greater than 1. This is just the area sign condition: the area of the pre-before-post lobe is $-1$, and for stability when paired with an excitatory PSP $\mathcal{E}$ the total area under the learning rule $\mathcal{L}$ must be negative.

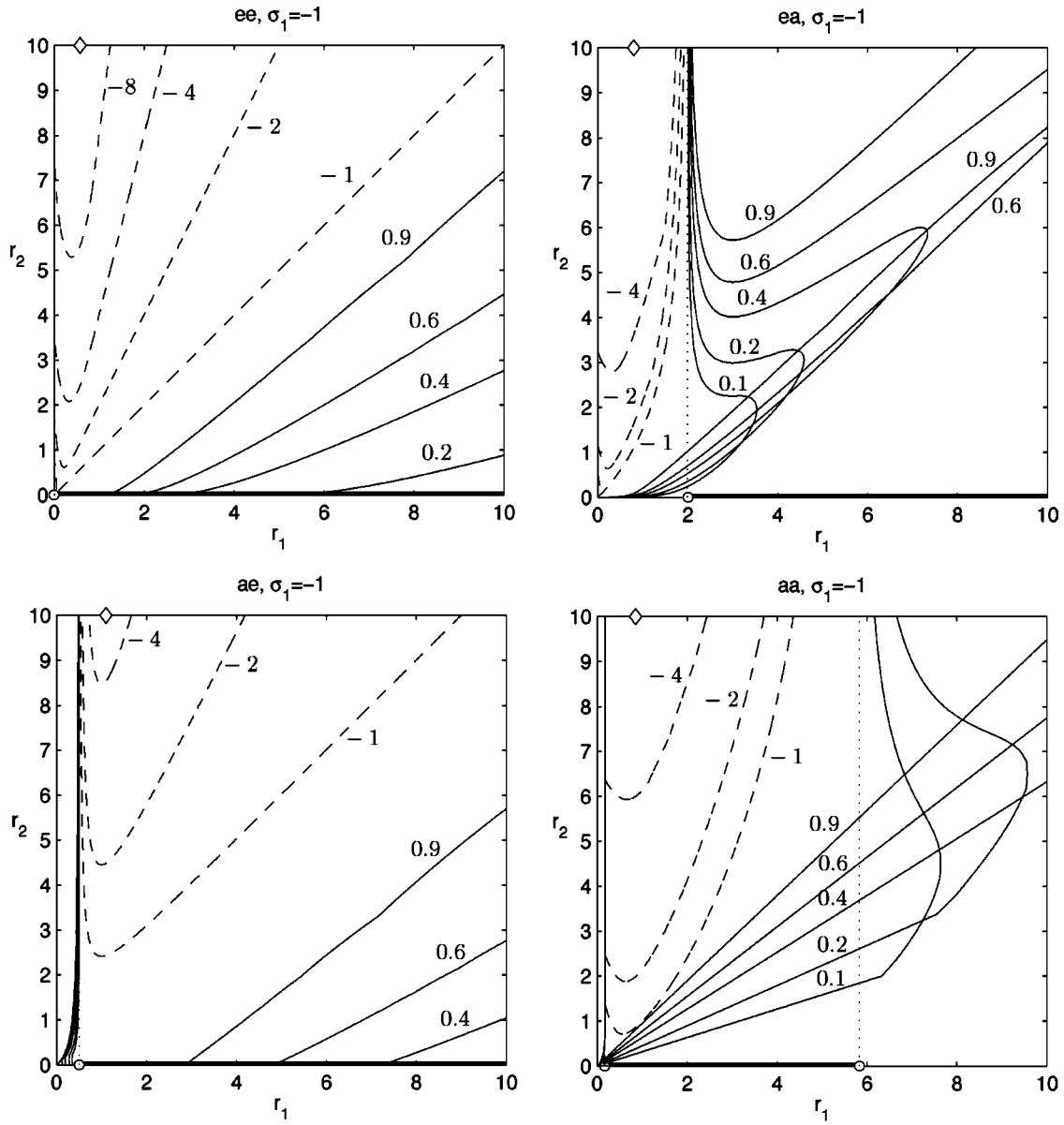


FIG. 9. Boundary curves of the stable region in $(r_1, r_2)$ for various values of $A$, for two lobe $\mathcal{L}$ with $\sigma_1 = -1$. Curves are labeled by $A$. Curves with $A > 0$ are drawn with solid lines, curves with $A < 0$ with dashed lines. In all cases the region of stability is on the side of the curve containing the diamond ( $\diamond$ ) in the upper left corner of the plot. The interval of stability for the corresponding one-lobe learning rules is the portion of the $r_1$ axis in bold.

For $A < 1$, the effect of the post-before-pre lobe shows the following general trends: in cases ee and ae, as the absolute area $|A|$ of the post-before-pre lobe increases, the stable region in the relative time constants $r_1$ and $r_2$ tends to shrink. Hence, the post-before-pre lobe can be thought of as destabilizing in such cases. In cases ee and ae the situation is less clear. Increasingly negative $A$ (larger depressive post-before-pre) is uniformly destabilizing, but increasingly positive $A$ (larger potentiating post-before-pre) appears to be destabilizing for small $r_2$ but stabilizing for large $r_2$.

Cooperative stability, in which a two-lobe rule is stable while each of its lobes individually would be unstable, occurs in cases ea, ae, and aa: any point $(r_1, r_2)$ in a stable region, with $r_1$ outside the interval in which the corresponding one-lobe rule is stable, is an example of cooperative stability.

Finally, the shape and extent of stable regions for two-lobe learning rules, or the extent of stable intervals for one-lobe learning rules, depend critically on whether $\mathcal{E}$ and $\mathcal{L}$ are exponential or $\alpha$ function in form. This suggests that in order to infer even such qualitative properties as stability or instability in a biological context, the learning rule must be known with considerable precision.

However, for particular values of some parameters the dependence on functional form may be such that useful conclusions can still be drawn in the absence of such precision; for example, the stability in one-lobe, depressive pre-before-post learning rules with $\tau_L / \tau_E \simeq 1$, independent of whether $\mathcal{E}$

or $\mathcal{L}$ are exponential or $\alpha$ function in form. This particular finding has direct relevance to the learning rule observed experimentally in mormyrid ELL [7]. The experimental data are not precise enough to suggest a particular functional form, but do indicate a one-lobe, depressive, pre-before-post rule, with a width of the same order of magnitude as the width of a PSP. Stability of such a rule is consistent with the analytic results derived above.

## VIII. SUMMARY

We have investigated the existence and stability of negative image equilibria in spike-timing-dependent plasticity. The network architecture of the neural model is based on the known anatomy of mormyrid ELL, a cerebellum-like structure that is the initial site of electrosensory processing in mormyrid electric fish.

We proved that two conditions must hold for the existence of negative image equilibria. First, the high frequency content of the Fourier transform of the repeated external input must be less than that of the postsynaptic potential function, Eq. (13). Second, the nonassociative and associative components of the learning rule must satisfy a relative magnitude constraint, Eq. (18).

We proved a necessary and sufficient condition for stability of negative image equilibria, Eq. (30). The condition involves the Fourier transforms of the associative component of the learning rule and the postsynaptic potential function. We found stability to be independent of the nonassociative component of the learning rule, independent of the form of the repeated external input, and independent of the form of the postsynaptic gain function (provided the latter is a strictly increasing function of the postsynaptic potential).

The general stability condition derived in this paper is consistent with stability of the experimentally observed spike-timing-dependent learning rule in mormyrid ELL.

## ACKNOWLEDGMENTS

We would like to thank Dr. Gerhard Magnus, Dr. Nathaniel Sawtell, and the members of Dr. Curtis Bell's laboratory for insightful discussions. This material is based upon work supported by the National Science Foundation under Grant No. IBN-0114558, and by the National Institute of Mental Health under Grant No. R01-MH60364.

## APPENDIX

For completeness we provide below the polynomial conditions for stability of the two-lobe learning rules treated in the examples.

ee: $ak^4+bk^2+c<0$ for all $k$,

$$a=\sigma_1 r_1 r_2^2 - A r_1^2 r_2,$$

$$b=\sigma_1(r_2^2+r_1)+A(r_1^2-r_2),$$

$$c=\sigma_1+A$$

ea: $ak^6+bk^4+ck^2+d<0$ for all $k$

$$a=\sigma_1 r_1 r_2^3(2r_2-1)-A r_1^3 r_2(2r_1+1),$$

$$b=\sigma_1 r_2(r_2^3-3r_1^2 r_2+6r_1 r_2+r_1)+A r_1(r_1^3-3r_1 r_2^2 - 6r_1 r_2+r_2),$$

$$c=\sigma_1(2r_2^2-r_1^2+2r_1)+A(2r_1^2-r_2^2-2r_2),$$

$$d=\sigma_1+A$$

ae: $ak^4+bk^2+c<0$ for all $k$

$$a=\sigma_1 r_2^2(2r_1-1)-A r_1^2(2r_2+1),$$

$$b=\sigma_1(r_2^2+2r_1-1)+A(r_1^2-2r_2-1),$$

$$c=\sigma_1+A$$

aa: $ak^8+bk^6+ck^4+dk^2+e<0$ for all $k$

$$a=\sigma_1 r_1^2 r_2^4+A r_1^4 r_2^2,$$

$$b=\sigma_1(-r_2^4-2r_1^2 r_2^3+4r_1 r_2^4-r_1 r_2^3+3r_1^2 r_2^2 + 2r_1 r_2^2)+A r_1(-r_1^4+2r_1^3 r_2^2-4r_1^4 r_2-r_1^3 r_2 = 2r_1^2 r_2+3r_1^2 r_2^2),$$

$$c=\sigma_1[-3r_1 r_2+(r_1^2+r_2^2)(r_2+1)^2+r_2^2(1+4r_1 - 4r_1 r_2)]+A[-3r_1 r_2+(r_1^2+r_2^2)(r_1-1)^2+r_1^2 \times(1-4r_2-4r_1 r_2)],$$

$$d=\sigma_1(2r_2^2-r_1^2+4r_1-1)+A(2r_1^2-r_2^2-4r_2-1),$$

$$e=\sigma_1+A$$

[1] D.O. Hebb, *The Organization of Behavior* (Wiley, New York, 1949).
[2] T. Lomo, Exp. Brain Res. **12**, 46 (1971).
[3] T.V. Bliss and T. Lomo, J. Physiol. (London) **232**, 331 (1973).
[4] T.J. Sejnowski, J. Theor. Biol. **69**, 385 (1977).
[5] E.L. Bienenstock, L.N. Cooper, and P.W. Munro, J. Neurosci. **2**, 32 (1982).
[6] H. Markram, J. Lübke, M. Frotscher, and B. Sakmann, Science **275**, 213 (1997).
[7] C.C. Bell, V. Han, Y. Sugawara, and K. Grant, Nature (Lon-

don) **387**, 278 (1997).
[8] Q. Bi and M. Poo, J. Neurosci. **18**, 10 464 (1998).
[9] L.F. Abbott and S.B. Nelson, Nat. Neurosci. **3**, 1178 (2000).
[10] W. Gerstner, R. Kempter, J.L. van Hemmen, and H. Wagner, Nature (London) **383**, 76 (1996).
[11] M.C.W. van Rossum, G.Q. Bi, and G.G. Turrigiano, J. Neurosci. **20**, 88128821 (2000).
[12] J. Rubin, D.D. Lee, and H. Sompolinsky, Phys. Rev. Lett. **86**, 364 (2001).
[13] M. Yoshioka, Phys. Rev. E **65**, 011903 (2002).
[14] V.P. Zhigulin, M.I. Rabinovich, R. Huerta, and H.D. Abarbanel, Phys. Rev. E **67**, 021901 (2003).
[15] H. Cateau and T. Fukai, Neural Comput. **15**, 597 (2003).
[16] (a) C.C. Bell, K. Grant, and J. Serrier, J. Neurophysiol. **68**, 843 (1992); (b) C.C. Bell (private communication).
[17] C.C. Bell, D. Bodznick, J. Montgomery, and J. Bastian, Brain Behav. Evol. **50**, 17 (1997).
[18] P.D. Roberts, Phys. Rev. E **62**, 4077 (2000).
[19] R. Kempter, W. Gerstner, and J.L. van Hemmen, Phys. Rev. E **59**, 4498 (1999).
[20] P.D. Roberts, J. Comput. Neurosci. **7**, 235 (1999).
[21] P.D. Roberts and C.C. Bell, J. Comput. Neurosci. **9**, 67 (2000).
[22] R.H. Hahnloser, A.A. Kozhevnikov, and M.S. Fee, Nature (London) **419**, 65 (2002).
[23] D. Ehrlich, J.H. Casseday, and E. Covey, J. Neurophysiol. **77**, 2360 (1997).
[24] W. Gerstner, R. Ritz, and J.L. van Hemmen, Biol. Cybern. **69**, 503 (1993).
[25] W. Gerstner and W.M. Kistler, *Spiking Neuron Models* (Cambridge University Press, Cambridge, 2002).
[26] W. Gerstner, Phys. Rev. E **51**, 738 (1995).
[27] P.D. Roberts, J. Neurophysiol. **84**, 2035 (2000).
[28] D.C. Champeney, *A Handbook of Fourier Theorems* (Cambridge University Press, Cambridge, 1987).
[29] R. Abraham and J. Robbin, *Transversal Mappings and Flows* (Benjamin, New York, 1967).
[30] P.J. Davis, *Circulant Matrices* (Wiley, New York, 1979).
[31] P.D. Roberts, Neurocomputing **32-33**, 243 (2000).
[32] V. Han, C.C. Bell, K. Grant, and Y. Sugawara, J. Comp. Neurol. **404**, 359 (1999).