Intrinsic Stabilization of Output Rates by Spike-Based Hebbian Learning

Richard Kempter
Keck Center for Integrative Neuroscience,
University of California at San Francisco,
San Francisco, CA 94143-0732, USA
kempter@phy.ucsf.edu

Wulfram Gerstner
Swiss Federal Institute of Technology Lausanne
Center of Neuromimetic Systems,
CH-1015 Lausanne EPFL, Switzerland
Wulfram.Gerstner@epfl.ch

J. Leo van Hemmen
Physik Department, Technische Universität München,
D-85747 Garching bei München, Germany
Leo.van.Hemmen@ph.tum.de

Submitted to Neural Comput. Dec. 1999, revised Sep. 2000.

Abstract: We study analytically a model of long-term synaptic plasticity where synaptic changes are triggered by presynaptic spikes, postsynaptic spikes, and the time differences between pre- and postsynaptic spikes. We show that plasticity can lead to an intrinsic stabilization of the mean firing rate of the postsynaptic neuron. Subtractive normalization of the synaptic weights (summed over all presynaptic inputs converging on a postsynaptic neuron) follows if, in addition, the mean input rates and the mean input correlations are identical at all synapses. If the integral over the learning window is positive, firing-rate stabilization requires a non-Hebbian component, whereas such a component is not needed, if the integral of the learning window is negative. A negative integral corresponds to 'anti-Hebbian' learning in a model with slowly varying firing rates. For spike-based learning, a strict distinction between Hebbian and 'anti-Hebbian' rules is questionable since learning is driven by correlations on the time scale of the learning window. The correlations between presynaptic and postsynaptic firing are evaluated for a piecewise-linear Poisson model and for a noisy spiking neuron model with refractoriness. Whereas a negative integral over the learning window leads to intrinsic rate stabilization, the positive part of the learning window picks up spatial and temporal correlations in the input.

1 Introduction

"Hebbian" learning (Hebb 1949), i.e., synaptic plasticity driven by correlations between pre- and postsynaptic activity, is thought to be an important mechanism for the tuning of neuronal connections during development and thereafter, in particular, for the development of receptive fields and computational maps; see e.g. (von der Malsburg 1973, Sejnowski 1977, Kohonen 1984, Linsker 1986, Sejnowski and Tesauro 1989, MacKay and Miller 1990, Wimbauer et al. 1994, Shouval and Perrone 1995, Miller 1996a); for a review see, e.g., Wiskott and Sejnowski (1998). It is well-known that simple Hebbian rules may lead to diverging synaptic weights so that weight normalization turns out to be an important topic (Kohonen 1984, Oja 1982, Miller and MacKay 1994, Miller 1996b). In practice, normalization of the weights w_i is imposed either by an explicit rescaling of all weights after a learning step, or by a constraint on the summed weights (e.g., $\sum_i w_i = c$, or $\sum_i w_i^2 = c$ for some constant c), or by an explicit decay term proportional to the weight itself.

In recent simulation studies of spike-based learning (Gerstner et al. 1996, Kempter et al. 1999a, Song et al. 2000, Xie and Seung 2000), however, *intrinsic* normalization properties of synaptic plasticity have been found. Neither an explicit normalization step nor a constraint on the summed weights was needed. So we face the question: How can we understand these findings? Some preliminary arguments as to why intrinsic normalization occurs in spike-based learning rules have been given (Gerstner 1998, Kempter et al. 1999a, Song et al. 2000). In this paper, we study normalization properties in more detail. In particular, we show that in spike-based plasticity with realistic learning windows the procedure of continuously strengthening and weakening the synapses can *automatically* lead to a normalization of the total input strength to the postsynaptic neuron in a competitive self-organized process and, hence, to a stabilization of the output firing rate.

In order to phrase the problem of normalization in a general framework, let us start with a rate-based learning rule of the form

$$\begin{aligned} \tau_w \frac{\mathrm{d}}{\mathrm{d}t} w_i(t) &= a_0 + a_1^{\mathrm{in}}\, \lambda_i^{\mathrm{in}}(t) + a_1^{\mathrm{out}}\, \lambda^{\mathrm{out}}(t) + a_2^{\mathrm{corr}}\, \lambda_i^{\mathrm{in}}(t)\, \lambda^{\mathrm{out}}(t) \\ &+ a_2^{\mathrm{in}}\, \left[\lambda_i^{\mathrm{in}}(t)\right]^2 + a_2^{\mathrm{out}}\, \left[\lambda^{\mathrm{out}}(t)\right]^2 \;, \end{aligned} \tag{1}$$

which can be seen as an expansion of local adaptation rules up to second order in the rates λ_i^{in} ($1 \le i \le N$) and λ^{out}; see e.g. (Bienenstock et al. 1982, Linsker 1986, Sejnowski and Tesauro 1989). (A definition of 'rate' will be given below). The time scale of learning is set by τ_w. The coefficients

a_0, $a_1^{\rm in}$, $a_1^{\rm out}$, $a_2^{\rm corr}$, $a_2^{\rm in}$, and $a_2^{\rm out}$ may, and in general will, depend on the weights w_i. For example, in Oja's rule (Oja 1982) we have $a_2^{\rm corr} = 1$ and $a_2^{\rm out}(w_i) = -w_i$ while all other coefficients vanish; for linear neurons, the weight vector converges to a unit vector. Here we study learning rules of the form (1) where the coefficients do *not* depend on the weights; we do, however, introduce upper and lower bounds for the weights (e.g., ${\rm d}w_i/{\rm d}t = 0$ if $w_i \geq w^{\rm max}$ or $w_i \leq 0$ for an excitatory synapse) so as to exclude run-away of individual weight values.

As a first issue of this paper, we argue that the focus on normalization of the *weights*, be it in the linear form $\sum_i w_i$ or in the quadratic form $\sum_i w_i^2$, is too narrow. A broad class of learning rules will naturally stabilize the *output rate* rather than the weights. As an example, let us consider a learning rule of the form

$$\tau_w \frac{\rm d}{{\rm d}t} w_i(t) = - \left[\lambda^{\rm out}(t) - \lambda^c\right] \lambda_i^{\rm in} \tag{2}$$

with constant rates $\lambda_i^{\rm in}$, $(1 \leq i \leq N)$. Equation (2) is a special case of Eq. (1). For excitatory synapses, $\lambda^{\rm out}$ increases with w_i; more precisely $\lambda^{\rm out}(w_1, \ldots, w_N)$ is an increasing function of each of the w_i. Hence learning stops if $\lambda^{\rm out}$ approaches $\lambda^c > 0$, and the fixed point $\lambda^{\rm out} = \lambda^c$ is stable. In the special case that all input lines $\lambda_i^{\rm in}$ have an equal rate $\lambda_i^{\rm in} = \lambda^{\rm in}$ independent of i, stabilization of the mean output rate implies the normalization of the sum $\sum_i w_i$. Section 3 of this paper will make the above argument more precise.

A disadvantage of the learning rule in Eq. (2) is that is has a correlation coefficient $a_2^{\rm corr} < 0$. Thus, in a pure rate description, the learning rule would be classified as anti-Hebbian rather than Hebbian. As a second issue of the paper, we will show in Section 4 that, for spike-based learning rules, the strict distinction between Hebbian and anti-Hebbian learning rules becomes questionable. A learning rule with a realistic time window (Levy and Stewart 1983, Markram et al. 1997, Zhang et al. 1998, Debanne et al. 1998, Bi and Poo 1998) can be anti-Hebbian in a time-averaged sense and still pick up positive "Hebbian" correlations between input and output spikes (Gerstner et al. 1996, Kempter et al. 1996, 1999a, 1999b). These correlations between input and output spikes that enter the learning dynamics will be calculated in this paper for a linear Poisson neuron and for a noisy spiking neuron with refractoriness.

Spike-based rules open the possibility of a direct comparison of model parameters with experiments (Senn et al. 1997, 2000). A theory of spike-based learning rules has been developed in (Gerstner et al. 1993, Häfliger et al. 1997, Ruf and Schmitt 1997, Senn et al. 1997, Eurich et al. 1999, Kempter et al. 1999a, Kistler and van Hemmen 2000, Roberts and Bell 2000, Senn et al. 2000, Xie and Seung 2000); for a review see, e.g., (van Hemmen 2000). Spike-based learning rules are closely related to rules for sequence learning (Herz et al. 1988, Herz et al. 1989, van Hemmen et al. 1990, Gerstner et al. 1993, Abbott and Blum 1996, Gerstner and Abbott 1997), where the idea of asymmetric learning windows is exploited. In Section 4 a theory of spike-based learning is outlined and applied to the problem of weight normalization via rate stabilization.

2 Input Scenario

We consider a single neuron that receives input from N synapses with weights w_i where $1 \leq i \leq N$ is the index of the synapse. At each synapse i, input spikes arrive stochastically. In the spike-based description of Section 4 we will model the input spike train at synapse i as an inhomogeneous Poisson process with rate $\lambda_i^{\text{in}}(t)$. In the rate description of Section 3 the input spike train is replaced by the continuous rate variable $\lambda_i^{\text{in}}(t)$. Throughout the paper we focus on two different input scenarios. In both scenarios, input rates are defined as statistical ensembles that we now describe.

2.1 Static-Pattern Scenario

The static-pattern scenario is the standard framework for the theory of unsupervised Hebbian learning (Hertz et al. 1991). The input ensemble contains p patterns defined as vectors $\mathbf{x}^\mu \in \mathbb{R}^N$ where $\mathbf{x}^\mu = (x_1^\mu, x_2^\mu, \ldots, x_N^\mu)^T$ with pattern label $1 \leq \mu \leq p$. Time is discretized in intervals Δ_t. After each time interval Δ_t, a new pattern is chosen randomly from the ensemble of patterns and applied at the input, i.e., during application of pattern μ the input rate at synapse i is $\lambda_i^{\text{in}} = x_i^\mu$. Temporal averaging over a time T yields the average input rate at each synapse

$$\overline{\lambda_i^{\text{in}}(t)} := \frac{1}{T} \int\limits_{t-T}^{t} \mathrm{d}t' \, \lambda_i^{\text{in}}(t') \; . \tag{3}$$

For long intervals $T \gg p\,\Delta_t$ the average input rate is constant, $\overline{\lambda_i^{\text{in}}(t)} = \overline{\lambda_i^{\text{in}}}$, and we find

$$\overline{\lambda_i^{\text{in}}} = \frac{1}{p} \sum_{\mu=1}^{p} x_i^\mu \; . \tag{4}$$

An overbar will always denote temporal averaging over long intervals T. The normalized correlation coefficient of the input ensemble is

$$C_{ij}^{\text{static}} = \frac{1}{p} \sum_{\mu=1}^{p} \frac{(x_i^\mu - \overline{\lambda_i^{\text{in}}})\,(x_j^\mu - \overline{\lambda_j^{\text{in}}})}{\overline{\lambda_i^{\text{in}}} \cdot \overline{\lambda_j^{\text{in}}}} = C_{ji}^{\text{static}} \,. \tag{5}$$

An explicit example of the static-pattern scenario is given in the Appendix. Input correlations play a major role in the theory of unsupervised Hebbian learning (Hertz et al. 1991).

2.2 Translation-Invariant Scenario

In developmental learning, it is often natural to assume a translation-invariant scenario; e.g., (Miller 1996a). In this scenario, the mean input rates are the same at all synapses and (again) constant in time, i.e.,

$$\overline{\lambda_i^{\text{in}}(t)} = \overline{\lambda^{\text{in}}} \quad \text{for all } i. \tag{6}$$

At each synapse i, the actual rate $\lambda_i(t)$ is time-dependent and fluctuates, on a time scale Δ_{corr} around its mean. In other words, the correlation function, defined as

$$C_{ij}(s) := \frac{\overline{\Delta\lambda_i^{\text{in}}(t)\,\Delta\lambda_j^{\text{in}}(t-s)}}{\overline{\lambda_i^{\text{in}}} \cdot \overline{\lambda_j^{\text{in}}}\,,} = C_{ji}(-s) \tag{7}$$

is supposed to vanish for $|s| \gg \Delta_{\text{corr}}$. Here $\Delta\lambda_i^{\text{in}}(t) := \lambda_i^{\text{in}}(t) - \overline{\lambda_i^{\text{in}}}$ is the deviation from the mean.

The essential hypothesis is that the correlation function is translation invariant, i.e., $C_{ij} = C_{|i-j|}$. Let us suppose, for example, that the ensemble of stimuli consists of objects or patterns of some typical size spanning l input pixels so that the correlation function has a spatial width of order l. If the objects are moved randomly with equal probability in an arbitrary direction across the stimulus array, then the temporal part of the correlation function will be symmetric in that $C_{|i-j|}(s) = C_{|i-j|}(-s)$.

Using the static-pattern scenario or the translation-invariant scenario we are able to simplify the analysis of Hebbian learning rules considerably, i.e., rate-based learning in Section 3 and spike-based learning in Section 4.

3 Plasticity in Rate Description

In this section we start with some preliminary considerations formulated on the level of rates and then analyze the rate description proper. Our aim is to show that the learning dynamics defined in equation (1) can yield a stable fixed point of the output *rate*. In order to keep the discussion as simple as possible, we focus mainly on a linearized rate model and focus on the input scenarios of Section 2. The learning rule is that of Eq. (1) with $a_2^{\rm in} = a_2^{\rm out} = 0$. All synapses are assumed to be of the same type so that the coefficients $a_1^{\rm in}$, $a_1^{\rm out}$, and $a_2^{\rm corr}$ do not depend on the synapse index i.

As is usual in the theory of Hebbian learning, we assume that the time scale τ_w of learning and that of fluctuations in the input and neuronal output can be separated. In other words, we assume $\tau_w \gg \Delta_t$ in the static-pattern scenario or $\tau_w \gg \Delta_{\rm corr}$ in the translation-invariant scenario. The right-hand side of (1) is then 'self-averaging', i.e., synaptic changes are driven by temporally averaged quantities

$$\begin{aligned} \tau_w \frac{\mathrm{d}}{\mathrm{d}t} w_i(t) &= a_0 + a_1^{\rm in}\, \overline{\lambda_i^{\rm in}} + a_1^{\rm out}\, \overline{\lambda^{\rm out}(t)} \\ &+ a_2^{\rm corr}\, \overline{\lambda_i^{\rm in}} \cdot \overline{\lambda^{\rm out}(t)} + a_2^{\rm corr}\, \overline{\Delta\lambda_i^{\rm in}(t)\ \Delta\lambda^{\rm out}(t)}\ . \end{aligned} \tag{8}$$

'Hebbian' learning of a synapse should be driven by the correlations of pre- and postsynaptic neuron, i.e., the last term on the right-hand side of (8); cf. (Sejnowski 1977, Sejnowski and Tesauro 1989, Hertz et al. 1991). This term is of second order in $\Delta\lambda^{\rm in}$ and it might therefore seem it is small compared to the term proportional to $\overline{\lambda_i^{\rm in}} \cdot \overline{\lambda^{\rm out}}$ – but is it really? Let us speculate for the moment that in an 'ideal' neuron the mean output rate $\overline{\lambda^{\rm out}}$ is attracted towards an operating point

$$\lambda_{\rm FP}^{\rm out} = -\frac{a_0 + a_1^{\rm in}\, \overline{\lambda^{\rm in}}}{a_1^{\rm out} + a_2^{\rm corr}\, \overline{\lambda^{\rm in}}}\ , \tag{9}$$

where $\overline{\lambda^{\rm in}}$ is the typical mean input rate which is identical at all synapses. Then the dominant term in the right-hand side of (8) would indeed be the covariance term $\propto \overline{\Delta\lambda_i^{\rm in}\, \Delta\lambda^{\rm out}}$ because all other terms in the right-hand side of Eq. (8) cancel each other. The idea of a fixed point of the rate $\overline{\lambda^{\rm out}} = \lambda_{\rm FP}^{\rm out}$ will be made more precise in the following two subsections.

3.1 Piecewise Linear Neuron Model

As a first example, let us study the piecewise linear neuron model

$$\lambda^{\text{out}}(t) = \left[\lambda_0 + \gamma_0 \, \frac{1}{N} \sum_{i=1}^{N} w_i(t) \, \lambda_i^{\text{in}}(t)\right]_+ \tag{10}$$

where λ_0 is a constant and $\gamma_0 > 0$ is the slope of the activation function[1]. In case $\lambda_0 < 0$, we must ensure explicitly that λ^{out} is non-negative. To do so, we have introduced the notation $[.]_+$ with $[x]_+ = x$ for $x > 0$ and $[x]_+ = 0$ for $x \leq 0$. In the following we will always assume that the argument inside the square brackets is positive, drop the brackets, and treat the model as strictly linear. Only at the end of the calculation we check for consistency, i.e., for $\lambda^{\text{out}} \geq 0$. At the end we also check for lower and upper bounds, $0 \leq w_i \leq w^{\max}$ for all i. [2]

Since learning is assumed to be slow ($\tau_w \gg \Delta_{\text{corr}}$), the weights w_i do not change on the fast time scale of input fluctuations. Hence the correlation term $\overline{\Delta\lambda_i^{\text{in}} \, \Delta\lambda^{\text{out}}}$ in (8) can be evaluated for constant w_i. Using (10), the definition $\Delta\lambda^{\text{out}}(t) := \lambda^{\text{out}}(t) - \overline{\lambda^{\text{out}}(t)}$, and the definition of the input correlations C_{ij} as given by Eq. (7) we obtain

$$\overline{\Delta\lambda_i^{\text{in}}(t) \, \Delta\lambda^{\text{out}}(t)} = \overline{\lambda_i^{\text{in}}} \, \gamma_0 \, \frac{1}{N} \sum_{j=1}^{N} w_j(t) \, \overline{\lambda_j^{\text{in}}} \, C_{ij}(0) \ . \tag{11}$$

We now want to derive the conditions under which the mean output rate $\overline{\lambda^{\text{out}}}$ approaches a fixed point $\lambda_{\text{FP}}^{\text{out}}$. Let us suppose that the 'average correlation' has a value

$$C := \frac{\sum_{i=1}^{N} \left(\overline{\lambda_i^{\text{in}}}\right)^2 \, C_{ij}(0)}{\sum_{i=1}^{N} \left(\overline{\lambda_i^{\text{in}}}\right)^2} \quad \text{for all } j \ , \tag{12}$$

i.e., we require that the average correlation is independent of the index j. This condition may look artificial but it is, in fact, a rather natural assumption. In particular, for the translation-invariant scenario Eq. (12)

[1] The normalization by $1/N$ ensures that the mean output rate stays the same if the number N of synaptic inputs (with identical input rate λ^{in}) is increased.

[2] For a discussion of the influence of lower and upper bounds for individual weights on the stabilization of the output rate see Section 5.3.

always holds. We use Eqs. (10), (11), and (12) in Eq. (8), multiply by $\gamma_0 \, N^{-1} \, \overline{\lambda_i^{\mathrm{in}}}$ and sum over i. Rearranging the result we obtain

$$\tau \frac{\mathrm{d}}{\mathrm{d}t} \overline{\lambda^{\mathrm{out}}(t)} = \lambda^{\mathrm{out}}_{\mathrm{FP}} - \overline{\lambda^{\mathrm{out}}(t)} \, . \tag{13}$$

with fixed point

$$\lambda^{\mathrm{out}}_{\mathrm{FP}} := \frac{\tau}{\tau_w} \frac{\gamma_0}{N} \left[a_0 \sum_{i=1}^{N} \overline{\lambda_i^{\mathrm{in}}} + a_1^{\mathrm{in}} \sum_{i=1}^{N} \left(\overline{\lambda_i^{\mathrm{in}}} \right)^2 - C \, a_2^{\mathrm{corr}} \, \lambda_0 \sum_{i=1}^{N} \left(\overline{\lambda_i^{\mathrm{in}}} \right)^2 \right] \, , \tag{14}$$

and time constant

$$\tau := -\tau_w \, \frac{N}{\gamma_0} \left[a_1^{\mathrm{out}} \sum_{i=1}^{N} \overline{\lambda_i^{\mathrm{in}}} + (1 + C) \, a_2^{\mathrm{corr}} \sum_{i=1}^{N} \left(\overline{\lambda_i^{\mathrm{in}}} \right)^2 \right]^{-1} \, . \tag{15}$$

The fixed point $\lambda^{\mathrm{out}}_{\mathrm{FP}}$ is asymptotically stable if and only if $\tau > 0$. For the translation-invariant scenario the mean input rates are the same at all synapses so that the fixed point (14) reduces to

$$\lambda^{\mathrm{out}}_{\mathrm{FP}} = - \frac{a_0 + a_1^{\mathrm{in}} \overline{\lambda^{\mathrm{in}}} - C a_2^{\mathrm{corr}} \lambda_0 \overline{\lambda^{\mathrm{in}}}}{a_1^{\mathrm{out}} + a_2^{\mathrm{corr}} \overline{\lambda^{\mathrm{in}}} \, (1 + C)} \, . \tag{16}$$

This is the generalization of Eq. (9) to the case of non-vanishing correlations.

We require $\lambda^{\mathrm{out}}_{\mathrm{FP}} > 0$ so as to assure that we are in the linear regime of our piecewise linear neuron model. For excitatory synapses we require in addition that the weights w_i are positive. Hence a realizable fixed point should have a rate $\lambda^{\mathrm{out}}_{\mathrm{FP}} > \max\{0, \lambda_0\}$. On the other hand, the weight values being bounded by $w^{\max}$, the fixed point must be smaller than $\lambda_0 + \gamma_0 \, w^{\max} \, \overline{\lambda^{\mathrm{in}}}$ since otherwise it cannot be reached. The learning dynamics would stop once all weights have reached their upper bound. Finally, the stability of the fixed point requires $\tau > 0$. The above requirements define conditions on the parameters a_0, a_1^{in}, a_1^{out}, a_2^{corr}, and λ_0 that we now explore in two examples. We note that a fixed point of the output rate implies a fixed point of the average weight $w := \sum_j \overline{\lambda_j^{\mathrm{in}}} \, w_j / \sum_j \overline{\lambda_j^{\mathrm{in}}}$; cf. Eq. (10).

Example 1 (No linear terms) *In standard correlation-based learning there are usually no linear terms. Let us therefore set $a_0 = a_1^{\mathrm{in}} = a_1^{\mathrm{out}} = 0$. We assume that the average correlation is non-negative, $C \geq 0$. Since we have $\gamma_0 > 0$ the term inside the square brackets of (15) must be negative in order to guarantee asymptotic stability of the fixed point. Since $\overline{\lambda_i^{\mathrm{in}}} > 0$ for all i,* *a stable fixed point* $\lambda_{\rm FP}^{\rm out}$ *can therefore be achieved only if* $a_2^{\rm corr} < 0$*, so that we end up with* anti-*Hebbian learning. The fixed point is*

$$\lambda_{\rm FP}^{\rm out} = \lambda_0 \, \frac{C}{1+C} \, . \tag{17}$$

As it should occur for positive rates we must require $\lambda_0 > 0$*. In addition, because* $\lambda_{\rm FP}^{\rm out} < \lambda_0$ *some of the weights* w_i *must be negative at the fixed point which contradicts the assumption of excitatory synapses; see Eq. (10). We note that for* $C \to 0$*, the output rate approaches zero; that is, the neuron becomes silent. To summarize this example, without linear terms a rate model of correlation-based learning must have a negative correlation coefficient in order to show output rate stabilization, and the model would be classified as anti-Hebbian learning.*

Example 2 (Hebbian learning) *In standard Hebbian learning the correlation term should have positive sign. We therefore set* $a_2^{\rm corr} > 0$*. As before we assume* $C > 0$*. From* $\tau > 0$ *in (15) we then obtain the condition that* $a_1^{\rm out}$ *be sufficiently negative in order to make the fixed point stable. In particular, we must have* $a_1^{\rm out} < 0$*. Hence firing rate stabilization in rate-based Hebbian learning requires a linear 'non-Hebbian' term* $a_1^{\rm out} < 0$*.*

3.2 Discussion

So far we have focused on a piecewise linear rate model. Can we generalize the above arguments to a non-linear neuron model? To keep the arguments transparent, we restrict our analysis to the translation-invariant scenario and assume identical mean input rates, $\overline{\lambda_i^{\rm in}} = \overline{\lambda^{\rm in}}$ for all i. As a specific example we consider a neuron model with a sigmoidal activation (or gain) function $\lambda^{\rm out} = g(u)$ where $u(t) = N^{-1} \sum_{i=1}^N w_i(t) \, \lambda_i^{\rm in}(t)$ and the derivative $g'(u)$ is positive for all u. For vanishing input, $u = 0$, we assume a spontaneous output rate $\lambda^{\rm out} = \lambda(0) > 0$. For $u \to -\infty$ the rate vanishes. For $u \to \infty$ the function g approaches the maximum output rate $\lambda^{\rm out} = \lambda^{\rm max}$; see Fig. 1.

Let us set $w = N^{-1} \sum_j w_j$. For a given mean weight w, we obtain a mean input $\overline{u} = w \, \overline{\lambda^{\rm in}}$. If the fluctuations $\Delta\lambda_i^{\rm in}(t) := \lambda_i^{\rm in}(t) - \overline{\lambda^{\rm in}}$ are small, we can linearize $g(u)$ around the mean input $\overline{u}$, i.e., $g(u) = g(\overline{u}) + g'(\overline{u}) \, (u - \overline{u})$; cf. Fig. 1. The linearization leads us back to (10),

$$\lambda^{\rm out} = \lambda_0(w) + \gamma_0(w) \, \frac{1}{N} \sum_{i=1}^{N} w_i \, \lambda_i^{\rm in} \; , \tag{18}$$

with $\lambda_0(w) = g(\overline{u}) - g'(\overline{u})\,\overline{u}$ and $\gamma_0(w) = g'(\overline{u})$.

The discussion of output rate stabilization proceeds as in the linear case. Does a fixed point exist? The condition $\mathrm{d}\overline{\lambda^{\rm out}}/\mathrm{d}t = 0$ yields a fixed point that is given by Eq. (16) with λ_0 replaced by $\lambda_0(w)$. Thus Eq. (16) becomes an *implicit* equation for $\lambda_{\rm FP}^{\rm out}$. A sigmoidal neuron can reach the fixed point if the latter lies in the range $0 < \lambda_{\rm FP}^{\rm out} < \lambda^{\rm max}$. The local stability analysis does not change. As before, a fixed point of $\lambda^{\rm out}$ implies a fixed point of the mean weight w. The statement follows directly from (18) since $\mathrm{d}\overline{\lambda^{\rm out}}/\mathrm{d}t = g'(u)\,\overline{\lambda^{\rm in}}N^{-1}\sum_i \mathrm{d}w_i/\mathrm{d}t$ and $g' > 0$. That is to say, stabilization of the output rate implies a stabilization of the weights.

Let us summarize this section. The value of the fixed point $\lambda_{\rm FP}^{\rm out}$ is identical for both linear and the linearized rate model and given by Eq. (16). It is stable if τ in Eq. (15) is positive. If we kept higher-order terms in the expansion of Eq. (18), the fixed point may shift slightly, but as long as the nonlinear corrections are small, the fixed point remains stable [due to the continuity of the differential equation (1)]. This highlights the fact that the learning rule (8) with an appropriate choice of parameters achieves a *stabilization* of the output *rate* rather than a true normalization of the synaptic weights.

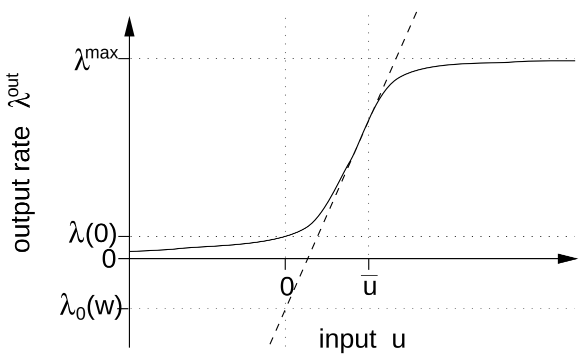

Figure 1: *Sigmoidal activation function g.* We plot the output rate $\lambda^{\rm out}$ (solid line) as a function of the neuron's input $u := N^{-1}\sum_i w_i\,\lambda_i^{\rm in}$. The spontaneous output rate at $u = 0$ is $\lambda(0)$ and the maximum output rate is $\lambda^{\rm max}$. The dashed line is a linearization of g around a mean input $\overline{u} > 0$. We note that here $\lambda_0(w) < 0$; cf. (18) and Examples 1 and 4.

4 Plasticity Driven by Spikes

In this section, we generalize the concept of intrinsic rate stabilization to spike-time dependent plasticity. We start by reviewing a learning rule where synaptic modifications are driven by the relative timing of pre- and postsynaptic spikes and defined by means of a learning window (Gerstner et al. 1996, Kempter et al. 1996); there is extensive experimental evidence supporting this idea (Levy and Stewart 1983, Bell et al. 1997, Markram et al. 1997, Zhang et al. 1998, Debanne et al. 1998, Bi and Poo 1998, Bi and Poo 1999, Abbott and Munro 1999). The comparison of the spike-based learning rule developed below with a rate-based learning rule discussed above allows us to give the coefficients a_1^{in}, a_1^{out}, and a_2^{corr} in (8) a more precise meaning. In particular, we show that a_2^{corr} corresponds to the integral over the learning window. Anti-Hebbian learning in the above rate model (Example 1), can thus be identified with a negative integral over the learning window Gerstner et al. 1996, 1997, Kempter et al. 1999a, 1999b).

4.1 Learning Rule

We are going to generalize Eq. (1) to spike-based learning. Changes in synaptic weights w_i are triggered by input spikes, output spikes, and the time-differences between input and output spikes. To simplify the notation, we write the sequence $\{t_i^1, t_i^2, \dots\}$ of spike arrival times at synapse i in the form of a 'spike train' $S_i^{\text{in}}(t) := \sum_m \delta(t - t_i^m)$ consisting of δ functions (real spikes have of course a finite width but what counts here are the events 'spike'). Similarly, the sequence of output spikes is denoted by $S^{\text{out}}(t) := \sum_n \delta(t - t^n)$. These notations allow us to formulate the spike-based learning rule as (Kempter et al. 1999a, Kistler and van Hemmen 2000, van Hemmen 2000)

$$\begin{aligned} \tau_w \frac{\mathrm{d}}{\mathrm{d}t} w_i(t) &= a_0 + a_1^{\text{in}}\, S_i^{\text{in}}(t) + a_1^{\text{out}}\, S^{\text{out}}(t) \\ &+ S_i^{\text{in}}(t) \int_{-\infty}^{t} \mathrm{d}t'\, W(t - t')\, S^{\text{out}}(t') \\ &+ S^{\text{out}}(t) \int_{-\infty}^{t} \mathrm{d}t'\, W(-t + t')\, S_i^{\text{in}}(t') \; . \end{aligned} \tag{19}$$

As in Eq. (1), the coefficients a_0, a_1^{in}, a_1^{out}, and the learning window W in general depend on the current weight value w_i. They may also depend on other local variables such as the membrane potential or the calcium concentration. For the sake of clarity of presentation, we drop these dependencies and assume constant coefficients. We can, however, assume upper and lower bounds for the synaptic efficacies w_i, i.e., weight changes are zero, if $w_i > w^{\max}$ or $w_i < 0$; see also the discussion in Section 5.3. Again for the sake of clarity, in the remainder of this section we presume that all weights are within the bounds.

How can we interpret the terms on the right in Eq. (19)? The coefficient a_0 is a simple decay or growth term that we will drop. $S_i^{\rm in}$ and $S^{\rm out}$ are sums of δ functions. We recall that integration of a differential equation $\mathrm{d}x/\mathrm{d}t = f(x,t) + a\,\delta(t)$ with arbitrary f yields a discontinuity at zero: $x(0_+) - x(0_-) = a$. Thus, an input spike arriving at synapse i at time t changes w_i at that time by a constant amount $\tau_w^{-1}\,a_1^{\rm in}$ and a variable amount $\tau_w^{-1}\int_{-\infty}^{t}\mathrm{d}t'\,W(t-t')\,S^{\rm out}(t')$, which depends on the sequence of output spikes occurring at times earlier than t. Similarly, an output spike at time t results in a weight change $\tau_w^{-1}\,[a_1^{\rm out} + \int_{-\infty}^{t}\mathrm{d}t'\,W(t'-t)\,S_i^{\rm in}(t')]$.

Our formulation of learning assumes discontinuous and instantaneous weight changes at the times of the occurrence of spikes which may not seem very realistic. The formulation can be generalized to continuous and delayed weight changes that are triggered by pre- and postsynaptic spikes. Provided the gradual weight change terminates after a time that is small compared to the time scale of learning, all results derived below would basically be the same.

Throughout what follows we assume that discontinuous weight changes are small compared to typical values of synaptic weights, meaning that learning is by small increments, which can be achieved for large τ_w. If τ_w is large, we can separate the time scale of learning from that of the neuronal dynamics, and the right-hand side of Eq. (19) is 'self-averaging' (Kempter et al. 1999a, van Hemmen 2000). The evolution of the weight vector (19) is then

$$\tau_w\,\frac{\mathrm{d}}{\mathrm{d}t}w_i(t) = a_1^{\rm in}\,\overline{\langle S_i^{\rm in}\rangle(t)} + a_1^{\rm out}\,\overline{\langle S^{\rm out}\rangle(t)} + \int_{-\infty}^{\infty}\mathrm{d}s\,W(s)\,\overline{\langle S_i^{\rm in}(t)\,S^{\rm out}(t-s)\rangle}\;. \tag{20}$$

Our notation with angular brackets plus horizontal bar is intended to emphasize that we average both over the spike statistics given the rates (angular brackets) *and* over the ensemble of rates (horizontal bar). An example will be given below in Section 4.4. Because weights change slowly and the statistical input ensemble is assumed to have stationary correlations, the two integrals in (19) have been replaced by a single integral in (20).

The formulation of the learning rule in Eq. (20) allows a direct link to rate-based learning as introduced in Eqs. (1) and (8). Indeed, averaged spike trains $\overline{\langle S_i^{\rm in}\rangle(t)}$ and $\overline{\langle S^{\rm out}\rangle(t)}$ are identical to the mean firing rates $\overline{\lambda_i^{\rm in}(t)}$ and $\overline{\lambda^{\rm out}(t)}$, respectively, as defined in Sections 2 and 3. In order interpret the correlation terms we have to be more careful. The terms $\overline{\langle S_i^{\rm in}\, S^{\rm out}\rangle}$ describe the correlation between input and output *on the level of spikes*. Below in Subsection 4.3 we will give a detailed treatment of the correlation term. In the next subsection we simplify the correlation term under the assumption of slowly changing firing rates.

4.2 Correlations in Rate-Based Learning

In this subsection we will relate the correlation term $a_2^{\rm corr}$ in Eq. (8) to the integral over the learning window W in Eq. (20). In order to make the transition from Eq. (20) to Eq. (8), two approximations are necessary (Kempter et al. 1999a, 1999b). First, we have to neglect correlations between input and output *spikes* apart from the correlations contained in the *rates*. That is, we make the approximation $\overline{\langle S_i^{\rm in}(t)\, S^{\rm out}(t-s)\rangle} \approx \overline{\lambda_i^{\rm in}(t)\, \lambda^{\rm out}(t-s)}$, the horizontal overbar denoting an average over the learning time. Second, we assume these rates change slowly as compared to the width of the learning window W (cf. Fig. 2a). Consequently we set $\lambda^{\rm out}(t-s) \approx \lambda^{\rm out}(t) - s\,[{\rm d}/{\rm d}t\, \lambda^{\rm out}(t)]$ in the integrand of (20). With the above two assumptions we obtain

$$
\begin{aligned}
\tau_w \frac{\rm d}{{\rm d}t}\, w_i(t) = \quad & a_1^{\rm in}\, \overline{\lambda_i^{\rm in}} + a_1^{\rm out}\, \overline{\lambda^{\rm out}(t)} + \overline{\lambda_i^{\rm in}(t)\, \lambda^{\rm out}(t)} \int_{-\infty}^{\infty} {\rm d}s\, W(s) \\
& - \overline{\lambda_i^{\rm in}(t)\, \frac{\rm d}{{\rm d}t}\, \lambda^{\rm out}(t)} \int_{-\infty}^{\infty} {\rm d}s\, s\, W(s)\ . \qquad (21)
\end{aligned}
$$

The last term on the right in (21) has been termed *differential*-Hebbian (Xie and Seung 2000). Equation (21) is equivalent to (8), if we neglect the differential-Hebbian term and if we set $a_2^{\rm corr} = \int {\rm d}s\, W(s)$ and $a_0 = 0$. We note, however, that $\overline{\lambda_i^{\rm in}(t)\, \lambda^{\rm out}(t)} = \overline{\lambda_i^{\rm in}} \cdot \overline{\lambda^{\rm out}(t)} + \overline{\Delta\lambda_i^{\rm in}(t)\, \Delta\lambda^{\rm out}(t)}$ still contains the correlations between the fluctuations in the input and output rates (but we had to assume that these correlations are slow compared to the width of the learning window). The condition of slowly changing rates will be dropped in the next section.

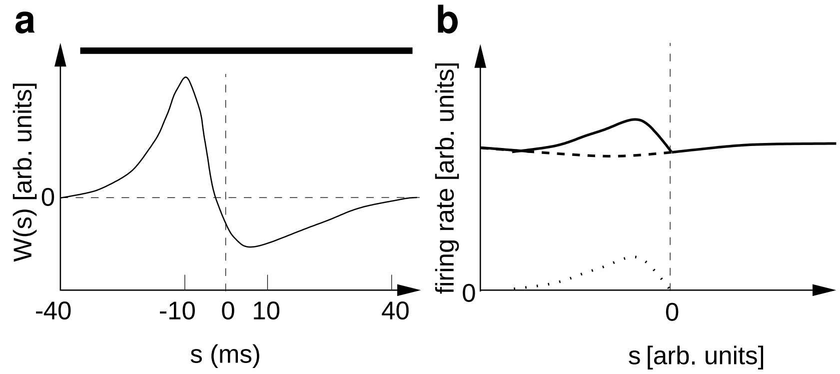

Figure 2: a) *Learning window* W (arbitrary units) as a function of the delay $s = t_i^m - t^n$ between presynaptic spike arrival time t_i^m and postsynaptic firing time t^n (schematic). If $W(s)$ is positive (negative) for some s, the synaptic efficacy w_i is increased (decreased). The increase of w_i is most efficient, if a presynaptic spike arrives a few milliseconds *before* the postsynaptic neuron starts firing. For $|s| \to \infty$ we have $W(s) \to 0$. The bar denotes the width of the learning window. The qualitative time course of the learning window is supported by experimental results (Levy and Stewart 1983, Markram et al. 1997, Zhang et al. 1998, Debanne et al. 1998, Bi and Poo 1998); see also the reviews by Brown and Chattarji (1994) and Linden (1999). The learning window can be described theoretically by a phenomenological model with microscopic variables (Senn et al. 1997, Senn et al. 2000, Gerstner et al. 1998). b) *Spike-spike correlations* (schematic). The correlation function $\mathrm{Corr}_i(s, \mathbf{w})/\overline{\lambda_i^{\mathrm{in}}}$ (full line) defined in (28) is the sum of two terms: (i) The causal contribution of an input spike at synapse i at time s to the output firing rate at time $s = 0$ is given by the time-reverted EPSP, i.e., $\gamma_0 \, N^{-1} \, w_i \, \epsilon(-s)$ (dotted line); (ii) the correlations between the mean firing rates $\overline{\lambda_i^{\mathrm{in}}(t) \; \lambda^{\mathrm{out}}(t-s)} \, / \, \overline{\lambda_i^{\mathrm{in}}}$ is indicated by the dashed line. For independent inputs with stationary mean, the rate contribution would be constant and equal to λ^{out}. In the figure we have sketched the situation where the rates themselves vary with time.

We have seen in Example 1 that, with rate-based learning, intrinsic normalization of the output rate is achievable once $a_2^{\mathrm{corr}} < 0$. We now argue that $a_2^{\mathrm{corr}} < 0$, which has been coined "anti-Hebbian" in the framework of rate-based learning, can still be "Hebbian" in terms of spike-based learning. The intuitive reason is that the learning window $W(s)$ as sketched in Fig. 2a has two parts, a positive and a negative one. Even if the integral over the learning window is negative, the size of weight changes in the positive part can be large enough to pick up correlations. In other words, the two approximations necessary to make the transition from Eq. (20) to Eq. (21) are, in general, not justified. An example is given in the the Appendix. In order to make our argument more precise we have to return to spike-based learning.

4.3 Spike-Spike Correlations

In the previous section, we have replaced $\overline{\langle S_i^{\mathrm{in}}(t)\, S^{\mathrm{out}}(t-s)\rangle}$ by the temporal correlation between slowly changing rates. This replacement leads to an oversimplification of the situation because it does not take into account the correlations on the level of spikes or fast changes of the rates. Here we present the correlation term in its general form.

Just as in Section 3, we require learning to be a slow process so that $\tau_w \gg \Delta_t$ or $\tau_w \gg \Delta_{\mathrm{corr}}$ (see Section 2 for a definition of Δ_t and Δ_{corr}). The correlation term can then be evaluated for constant weights w_i, $1 \le i \le N$. For input S_i^{in} with some given statistical properties that do not change on the slow time scale of learning, the correlations between $S_i^{\mathrm{in}}(t)$ and $S^{\mathrm{out}}(t-s)$ depend only on the time difference s and on the current weight vector $\mathbf{w}(t)$,

$$\mathrm{Corr}_i[s, \mathbf{w}(t)] := \overline{\langle S_i^{\mathrm{in}}(t)\, S^{\mathrm{out}}(t-s)\rangle}\,. \tag{22}$$

The horizontal bar refers, as before, to the temporal average introduced through the separation of time scales; cf. Eq. (3). Substituting (22) into (20) we obtain the dynamics of weight changes,

$$\tau_w \frac{\mathrm{d}}{\mathrm{d}t} w_i(t) = a_1^{\mathrm{in}}\, \overline{\lambda_i^{\mathrm{in}}} + a_1^{\mathrm{out}}\, \overline{\lambda^{\mathrm{out}}(t)} + \int_{-\infty}^{\infty} \mathrm{d}s\, W(s)\, \mathrm{Corr}_i[s, \mathbf{w}(t)] \ . \tag{23}$$

Equation (23) is completely analogous to (8) except that the correlations between rates have been replaced by those between spikes. We may summarize Eq. (23) by saying that *learning is driven by correlations on the time scale of the learning window.*

4.4 Piecewise Linear Poisson Neuron

To proceed with the analysis of Eq. (23), we need to determine the correlations Corr_i between input spikes at synapse i and output spikes. The correlations depend strongly on the neuron model under consideration. As a generalization to the linear rate-based neuron model in Section 3.1 we study an inhomogeneous Poisson model that generates spikes at a rate $\lambda^{\mathrm{out}}(t)$. For a membrane potential $u(t) \leq \vartheta$, the model neuron is quiescent. For $u > \vartheta$ the rate $\lambda^{\mathrm{out}}(t)$ is proportional to $u - \vartheta$. We call this model the piecewise linear Poisson neuron or, for short, the *Poisson neuron*.

Definition

The input to the neuron consists of N Poissonian spike trains with time-dependent intensities $\langle S_i^{\mathrm{in}} \rangle(t) = \lambda_i^{\mathrm{in}}(t)$, $1 \leq i \leq N$; for the mathematics of an inhomogeneous Poisson process we refer to Appendix A of (Kempter et al. 1998). As in the rate model of Section 3, the normalized correlation function for the *rate* fluctuations at synapse i and j is defined by Eq. (7). The input scenarios are the same as in Section 2, i.e., static-pattern scenario or translation-invariant scenario.

A spike arriving at t_i^f at synapse i evokes a postsynaptic potential (PSP) with time course $w_i\, \epsilon(t - t_i^f)$ which we assume to be excitatory (EPSP). We impose the condition $\int_0^\infty \mathrm{d}s\, \epsilon(s) = 1$ and require causality, i.e., $\epsilon(s) = 0$ for $s < 0$. The amplitude of the EPSP is given by the synaptic efficacy $w_i > 0$. The membrane potential u of the neuron is the linear superposition of all contributions,

$$u(t) = \frac{1}{N} \sum_{i=1}^{N} w_i(t) \sum_f \epsilon(t - t_i^f) = \frac{1}{N} \sum_{i=1}^{N} w_i(t) \int_{-\infty}^{+\infty} \mathrm{d}s\, \epsilon(s)\, S_i^{\mathrm{in}}(t - s) \qquad (24)$$

Since ϵ is causal only spike times $t_i^f < t$ count. The sums run over all spike arrival times t_i^f of all synapses. $S_i^{\mathrm{in}}(t) = \sum_f \delta(t - t_i^f)$ is the spike train at synapse i.

In the Poisson neuron, output spikes are generated stochastically with a time-dependent rate $\lambda_{S^{\mathrm{in}}}^{\mathrm{out}}$ that depends linearly on the membrane potential whenever u is above the threshold $\vartheta = -\lambda_0/\gamma_0$,

$$\lambda_{S^{\mathrm{in}}}^{\mathrm{out}}(t) = [\lambda_0 + \gamma_0\, u(t)]_+ = \left[\lambda_0 + \frac{\gamma_0}{N} \sum_{i=1}^{N} w_i(t) \sum_f \epsilon(t - t_i^f) \right]_+ . \qquad (25)$$

Here $\gamma_0 > 0$ and λ_0 are constants. As before, the notation with square brackets $[.]_+$ ensures that $\lambda^{\text{out}}_{S^{\text{in}}}$ be non-negative. The brackets will be dropped in the following. The subscript S^{in} of $\lambda^{\text{out}}_{S^{\text{in}}}$ indicates that the output rate depends on the specific realization of the input spike trains. We note that the spike-generation process is independent of previous *output* spikes. In particular, the Poisson neuron only allows refractoriness to the extent as it is included by the very nature of the Poisson process: the probability of more than one spike occurring in an interval of length δt is $o(\delta t)$. We drop this restriction in Section 4.6.

Expectation Values

In order to evaluate Eqs. (22) and (23), we first have to calculate the expectation values $\langle . \rangle$ (i.e., perform averages over the spike statistics of both input *and* output given the rates) and then average over time. By definition of the rate of an inhomogeneous Poisson process, the expected input is $\langle S^{\text{in}}_i \rangle = \lambda^{\text{in}}_i$. Taking advantage of Eqs. (24) and (25) we find that the expected output rate $\langle S^{\text{out}} \rangle = \langle \lambda^{\text{out}}_{S^{\text{in}}} \rangle = \lambda^{\text{out}}$, which is the rate contribution of *all* synapses to an output spike at time t, is given by

$$\langle S^{\text{out}} \rangle(t) = \lambda^{\text{out}}(t) = \lambda_0 + \frac{\gamma_0}{N} \sum_{i=1}^{N} w_i(t) \int \mathrm{d}s\, \epsilon(s)\, \lambda^{\text{in}}_i(t-s) \ , \qquad (26)$$

where the final term is a convolution of ϵ with the input rate.

Next we consider correlations between input and output, $\langle S^{\text{in}}_i(t)\, S^{\text{out}}(t-s) \rangle$, i.e., the expectation to find an input spike at synapse i at time t *and* and output spike at $t-s$. Since this expectation is defined as a *joint* probability density, it equals the probability density $\lambda^{\text{in}}_i(t)$ for an input spike at time t times the *conditional* probability density $\langle S^{\text{out}}(t-s) \rangle_{\{i,t\}}$ of observing an output spike at time $t-s$ given the input spike at synapse i at t; that is, $\langle S^{\text{in}}_i(t)\, S^{\text{out}}(t-s) \rangle = \lambda^{\text{in}}_i(t)\, \langle S^{\text{out}}(t-s) \rangle_{\{i,t\}}$. Within the framework of the *linear* Poisson neuron, the term $\langle S^{\text{out}}(t-s) \rangle_{\{i,t\}}$ equals the *sum* of the expected output rate $\lambda^{\text{out}}(t-s)$ in (26) and the specific contribution $\gamma_0\, N^{-1} w_i\, \epsilon(-s)$ of a single input spike at synapse i; cf. Eq. (25). To summarize, we get (Kempter et al. 1999a),

$$\langle S^{\text{in}}_i(t)\, S^{\text{out}}(t-s) \rangle = \lambda^{\text{in}}_i(t) \left[\frac{\gamma_0}{N}\, w_i(t)\, \epsilon(-s) + \lambda^{\text{out}}(t-s) \right] \ . \qquad (27)$$

Due to causality, the first term on the right, inside the square brackets of (27), must vanish for $s > 0$.

Temporal Average

In order to evaluate the correlation term (22), we need to take the temporal average of (27), i.e.,

$$\mathrm{Corr}_i[s, \mathbf{w}(t)] = \overline{\lambda_i^{\mathrm{in}}} \, \frac{\gamma_0}{N} \, w_i(t) \, \epsilon(-s) + \overline{\lambda_i^{\mathrm{in}}(t) \, \lambda^{\mathrm{out}}(t-s)} \; . \tag{28}$$

For excitatory synapses, the first term gives for $s < 0$ a positive contribution to the correlation function — as it should be (Fig. 2b). We recall that $s < 0$ means that a presynaptic spike precedes postsynaptic firing. The second term on the right in Eq. (28) describes the correlations between the input and output rates. These rates and correlations between them can, in principle, vary on an arbitrary fast time scale; for modeling papers, see e.g. (Gerstner et al. 1996, Xie and Seung 2000). We assume, however, for reasons of transparency that the *mean* input rates $\overline{\lambda_i^{\mathrm{in}}(t)} = \overline{\lambda_i^{\mathrm{in}}}$ are constant for all i; see also the input scenarios in Section 2. The mean output rate $\overline{\lambda^{\mathrm{out}}(t)}$ and the weights $w_i(t)$ may vary on the slow time scale of learning.

Learning Equation

Substituting (7), (26), and (28) into (23) we find

$$\begin{aligned} \tau_w \frac{\mathrm{d}}{\mathrm{d}t} w_i(t) \;&=\; a_1^{\mathrm{in}} \, \overline{\lambda_i^{\mathrm{in}}} + a_1^{\mathrm{out}} \, \overline{\lambda^{\mathrm{out}}(t)} + \overline{\lambda_i^{\mathrm{in}}} \cdot \overline{\lambda^{\mathrm{out}}(t)} \int_{-\infty}^{\infty} \mathrm{d}s \, W(s) \\ &+\; \frac{\gamma_0}{N} \sum_{j=1}^{N} w_j(t) \left[Q_{ij} \, \overline{\lambda_i^{\mathrm{in}}} \cdot \overline{\lambda_j^{\mathrm{in}}} + \delta_{ij} \, \overline{\lambda_i^{\mathrm{in}}} \int_{-\infty}^{0} \mathrm{d}s \, W(s) \, \epsilon(-s) \right] \end{aligned} \tag{29}$$

where

$$Q_{ij} := \int_{-\infty}^{\infty} \mathrm{d}s \, W(s) \int_{0}^{\infty} \mathrm{d}s' \, \epsilon(s') \, C_{ij}(s+s') \; . \tag{30}$$

The important factor in Eq. (30) is the input correlation function C_{ij} [as it appears in Eq. (7)] (low-pass) filtered by learning window W and the postsynaptic potential ϵ. For all input ensembles with $C_{ij}(\tau) = C_{ij}(-\tau)$ (viz. temporal symmetry), we have $Q_{ij} = Q_{ji}$. Hence the matrix Q can be diagonalized and has real eigenvalues. In particular, this remark applies to the 'static-pattern scenario' and the 'translation-invariant scenario' introduced in Section 2. It does not apply, for example, to sequence learning;

cf. (Herz et al. 1988, Herz et al. 1989, van Hemmen et al. 1990, Gerstner et al. 1993, Abbott and Blum 1996, Gerstner and Abbott 1997).

To get some preliminary insight into the nature of solutions to Eq. (29), let us suppose that all inputs have the same mean $\overline{\lambda_i^{\rm in}} = \overline{\lambda^{\rm in}}$. As we will see below, for a broad parameter range the mean output rate will be attracted towards a fixed point so that the first three terms in the right-hand side of Eq. (29) almost cancel each other. The dynamics of pattern formation in the weights w_i is then dominated by the largest positive eigenvalue of the matrix $[Q_{ij} + \delta_{ij}\,(\overline{\lambda^{\rm in}})^{-1} \int {\rm d}s\, W(s)\, \epsilon(-s)]$. The eigenvectors of this matrix are identical to those of the matrix (Q_{ij}). The eigenvalues of (Q_{ij}) can be positive even though $a_2^{\rm corr} = \int {\rm d}s\, W(s)$ is negative. A simple example is given in the Appendix.

To summarize this section, we have solved the dynamics of spike-time dependent plasticity defined by Eq. (20) for the Poisson neuron, a (piecewise) linear spiking neuron model. In particular, we have succeeded to evaluate the spike-spike correlations between presynaptic input and postsynaptic firing.

4.5 Stabilization in Spike-Based Learning

We are aiming at a stabilization of the output rate in analogy to Eqs. (14) and (15). To arrive at a differential equation for $\overline{\lambda^{\rm out}}$, we multiply (29) by $\overline{\lambda_i^{\rm in}}\,\gamma_0/N$ and sum over i. Similarly to (12) we assume that

$$Q := \frac{1}{N} \sum_{i=1}^{N} \left(\overline{\lambda_i^{\rm in}}\right)^2 Q_{ij} \tag{31}$$

is independent of the index j. To keep the arguments simple, we assume in addition that all inputs have the same mean $\overline{\lambda_i^{\rm in}} = \overline{\lambda^{\rm in}}$. For the translation-invariant scenario, both assumptions can be verified. With these simplifications, we arrive at a differential equation for the mean output rate $\overline{\lambda^{\rm out}}$ which is identical to (13), i.e., $\tau \frac{\rm d}{{\rm d}t}\overline{\lambda^{\rm out}} = \lambda_{\rm FP}^{\rm out} - \overline{\lambda^{\rm out}}$ with fixed point

$$\lambda_{\rm FP}^{\rm out} = \frac{\tau}{\tau_w}\,\gamma_0 \left[a_1^{\rm in} \left(\overline{\lambda^{\rm in}}\right)^2 - (Q + \beta)\,\lambda_0 \right] \tag{32}$$

and time constant

$$\tau = -\frac{\tau_w}{\gamma_0} \left[a_1^{\rm out}\overline{\lambda^{\rm in}} + \left(\overline{\lambda^{\rm in}}\right)^2 \int_{-\infty}^{\infty} {\rm d}s\, W(s) + Q + \beta \right]^{-1} \tag{33}$$

where

$$\beta := \frac{\overline{\lambda^{\text{in}}}}{N} \int_{-\infty}^{0} \mathrm{d}s\, W(s)\, \epsilon(-s) \tag{34}$$

is the contribution that is due to the spike-spike correlations between pre- and postsynaptic neuron. A comparison with Eqs. (14) and (15) shows that we may set $a_0 = 0$ and $a_2^{\text{corr}} = \int \mathrm{d}s\, W(s)$. The average correlation is to be replaced by

$$C = \frac{Q + \beta}{(\overline{\lambda^{\text{in}}})^2 \int \mathrm{d}s\, W(s)} \; . \tag{35}$$

All previous arguments regarding intrinsic normalization apply. The only difference is the new definition of the factor C in (35) as compared to (12). This difference is, however, important. First, the sign of $\int \mathrm{d}s\, W(s)$ enters the definition of C in Eq. (35). Furthermore, an additional term β appears. It is due to the extra spike-spike correlations between presynaptic input and postsynaptic firing. Note that β is of order N^{-1}. Thus it is small compared to the first term in (35) whenever the number of non-zero synapses is large (Kempter et al. 1999a). If we neglect the second term (i.e., $\beta = 0$) and if we set $\epsilon(s) = \delta(s)$ and $W(s) = a_2^{\text{corr}}\, \delta(s)$, Eq. (35) is identical to (12). On the other hand, β is of order $\overline{\lambda^{\text{in}}}$ in the input rates whereas Q is of order $(\overline{\lambda^{\text{in}}})^2$. Thus, β becomes important whenever the mean input rates are low. As shown in the Appendix, β is, for a realistic set of parameters, of the same order of magnitude as Q.

Let us now study the intrinsic normalization properties for two combinations of parameters.

Example 3 (Positive integral) *We consider the case $\int_{-\infty}^{+\infty} \mathrm{d}s\, W(s) > 0$ and assume that correlations in the input on the time scale of the learning window are positive so that $Q > 0$. Furthermore, for excitatory synapses, it is natural to assume $\beta > 0$. In order to guarantee the stability of the fixed point we must require $\tau > 0$, which yields the condition*

$$a_1^{\text{out}} < -\overline{\lambda^{\text{in}}} \int_{-\infty}^{\infty} \mathrm{d}s\, W(s) - \frac{Q + \beta}{\overline{\lambda^{\text{in}}}} < 0 \, . \tag{36}$$

Hence, if a_1^{out} is sufficiently negative, output rate stabilization is possible even if the integral over the learning window has positive sign. In other words, for a positive integral over the learning window, a non-Hebbian 'linear' term is necessary.

Example 4 (No linear terms) *In standard Hebbian learning, all linear terms vanish. Let us therefore set $a_1^{\text{in}} = a_1^{\text{out}} = 0$. As before in Example 3, we assume $Q + \beta > 0$. To guarantee the stability of the fixed point, we must now require*

$$\int_{-\infty}^{\infty} \mathrm{d}s\, W(s) < -\frac{Q+\beta}{(\overline{\lambda^{\text{in}}})^2} < 0\,, \tag{37}$$

Thus a stable fixed point of the output rate is possible, if the integral over the learning window is sufficiently negative. The fixed point is

$$\lambda_{\text{FP}}^{\text{out}} = \lambda_0 \,\frac{Q+\beta}{Q+\beta+(\overline{\lambda^{\text{in}}})^2 \int \mathrm{d}s\, W(s)} \tag{38}$$

Note that the denominator is negative because of Eq. (37). For a fixed point to exist we need $\lambda_{\text{FP}}^{\text{out}} > 0$, hence $\lambda_0 < 0$; cf. Fig. 1. We emphasize that, in contrast to Example 1, all weights at the fixed point can be positive, as should be the case for excitatory synapses.

To summarize the above two examples, a stabilization of the output rate is possible for positive integral with a non-Hebbian term $a_1^{\text{out}} < 0$ or for a negative integral and vanishing non-Hebbian terms.

4.6 Spiking Model with Refractoriness

In this subsection, we generalize the above results to a more realistic model of a neuron, viz. the spike response model (SRM) (Gerstner and van Hemmen 1992, Gerstner et al. 1996b, Gerstner 2000). Two aspects change with respect to the Poisson neuron. First, we include arbitrary refractoriness into the description of the neuron in Eq. (24). In contrast to a Poisson process, events in disjoint intervals are not independent. Second, we replace the linear firing intensity in Eq. (25) by a nonlinear one (and linearize only later on).

Let us start with the neuronal membrane potential. As before, each input spike evokes an excitatory postsynaptic potential described by a response kernel $\epsilon(s)$ with $\int_0^\infty \mathrm{d}s\, \epsilon(s) = 1$. After each output spike, the neuron undergoes a phase of refractoriness which is described by a further response kernel η. The total membrane potential is

$$u(t|\hat{t}) = \eta(t-\hat{t}) + \frac{1}{N}\sum_{i=1}^{N} w_i(t) \sum_f \epsilon(t - t_i^f)\ . \tag{39}$$

where $\hat{t}$ is the *last* output spike of the postsynaptic neuron. As an example, let us consider the refractory kernel

$$\eta(s) = -\eta_0 \exp\left(-\frac{s}{\tau_\eta}\right) \Theta(s) \, , \tag{40}$$

where $-\eta(0) =: \eta_0 > 0$ and $\tau_\eta > 0$ are parameters and $\Theta(s)$ is the Heaviside function, i.e., $\Theta(s) = 0$ for $s \le 0$ and $\Theta(s) = 1$ for $s > 0$. With this definition of η, the SRM is related to the standard integrate-and-fire model. A difference is that in the integrate-and-fire model the voltage is reset after each firing to a *fixed* value whereas, in Equations (39) and (40), the membrane potential is reset at time $t = \hat{t}$ by an amount $-\eta_0 - \eta(t - t')$, which depends on the time t' of the previous spike.

In analogy to Eq. (25), the probability of spike firing depends on the momentary value of the membrane potential. More precisely, the stochastic intensity of spike firing is a (nonlinear) function of $u(t|\hat{t})$, viz.,

$$\lambda^{\text{out}}(t|\hat{t}) = f[u(t|\hat{t})] \tag{41}$$

with $f(u) \to 0$ for $u \to -\infty$. For example, we may take $f(u) = \lambda_0 \exp[\gamma_0 \, (u - \vartheta)/\lambda_0]$ with parameters $\lambda_0, \gamma_0, \vartheta > 0$. For $\gamma_0 \to \infty$, spike firing becomes a threshold process and occurs whenever u reaches the threshold ϑ from below. For $\vartheta = 0$ and $|\gamma_0 \, u/\lambda_0| \ll 1$ we are led back to the linear model of Eq. (25). A systematic study of escape functions f can be found in (Plesser and Gerstner 2000).

After the definition of the model, we now turn to the calculation of the correlation function Corr_i as defined in Eq. (22). More precisely, we are interested in finding the generalization of Eq. (28) (for Poisson neurons) to spiking neurons with refractoriness. Due to the dependence of the membrane potential in Eq. (39) upon the last firing time $\hat{t}$, the first term on the right-hand side of Eq. (28) will become more complicated. Intuitively speaking, we have to average over the firing times $\hat{t}$ of the postsynaptic neuron. The correct averaging can be found from the theory of population dynamics studied in (Gerstner 2000). To keep the formulas as simple as possible, we consider constant input rates $\lambda_i^{\text{in}}(t) \equiv \lambda^{\text{in}}$ for all i. If many input spikes arrive on average within the postsynaptic integration time set by the kernel ϵ, the membrane potential fluctuations are small and we can linearize the dynamics about the mean trajectory of the membrane potential

$$\tilde{u}(t|\hat{t}) = \eta(t - \hat{t}) + \overline{\lambda^{\text{in}}} \frac{1}{N} \sum_{i=1}^{N} w_i \; . \tag{42}$$

To linear order in the membrane potential fluctuations, the correlations are (Gerstner 2000)

$$\mathrm{Corr}_i(s, \mathbf{w}) = \overline{\lambda^{\mathrm{in}}}\, P_i(-s) + \overline{\lambda^{\mathrm{in}}\, \lambda^{\mathrm{out}}} \tag{43}$$

where

$$P_i(s) = \overline{\lambda^{\mathrm{out}}}\, \frac{w_i}{N}\, \frac{\mathrm{d}}{\mathrm{d}s} \int_0^\infty \mathrm{d}x\, \mathcal{L}(x)\, \epsilon(s-x) + \int_0^s \mathrm{d}\hat{t}\;\; f[u(s|\hat{t})]\, S(s|\hat{t})\, P_i(\hat{t})\,. \tag{44}$$

The survivor function S is defined as

$$S(s|\hat{t}) = \exp\left\{ - \int_{\hat{t}}^{s} \mathrm{d}t'\, f[u(t'|\hat{t})] \right\}, \tag{45}$$

the mean output rate is

$$\overline{\lambda^{\mathrm{out}}(t)} = \left[\int_0^\infty \mathrm{d}t\, S(t|0) \right]^{-1}, \tag{46}$$

and the 'filter' $\mathcal{L}$ is

$$\mathcal{L}(x) = \int_x^\infty \mathrm{d}\xi\, \frac{\mathrm{d}}{\mathrm{d}u} f[u(\xi|x)]\, S(\xi|0)\,. \tag{47}$$

In Eqs. (44)–(47) the membrane potential u is given by the mean trajectory $\tilde{u}$ in (42).

Equations (43) and (44) are our main result. All considerations regarding normalization as discussed in the preceding subsection are valid with the function Corr_i defined in equations (43) and (44) and an output rate $\overline{\lambda^{\mathrm{out}}}$ defined in (46). In passing we note that Eq. (46) defines the gain function of the corresponding rate model, $\overline{\lambda^{\mathrm{out}}} = g[\overline{\lambda^{\mathrm{in}}}\, N^{-1} \sum_i w_i]$.

Let us now discuss Eq. (44) in more detail. The first term in the right-hand side of Eq. (44) describes the immediate influence of the EPSP ϵ on the firing probability of the postsynaptic neuron. The second term describes the reverberation of the primary effect that occurs one interspike interval later. If the interspike interval distribution is broad, the second term on the right in Eq. (44) is small and may be neglected (Gerstner 2000). In order to understand the relation of Eqs. (43) and (44) to Eq. (28), let us study two examples.

Example 5 (No refractoriness) *We want to show that in the limit of no refractoriness the spiking neuron model becomes identical with a Poisson model. If we neglect refractoriness* ($\eta \equiv 0$)*, the mean membrane potential in (42) is* $\tilde{u} = \overline{\lambda^{\text{in}}}\, N^{-1} \sum_i w_i$*. The output rate is* $\overline{\lambda^{\text{out}}} = f(\tilde{u})$ *[cf. (41)], and we have* $S(s|0) = \exp[-\overline{\lambda^{\text{out}}}\, s]$*; see (45). In (47) we set* $\gamma_0 = \mathrm{d}f/\mathrm{d}u$ *evaluated at* $\tilde{u}$ *and find* $P_i(s) = \epsilon(s)\, w_i\, \gamma_0/N$*. Hence, Eq. (43) reduces to Eq. (28), as it should be.*

Example 6 (High- and low-noise limit) *If we take an exponential escape rate* $f(u) = \lambda_0 \exp[\gamma_0\,(u - \vartheta)/\lambda_0]$ *the high-noise limit is found for* $|\gamma_0\,(u - \vartheta)/\lambda_0| \ll 1$ *and the low-noise limit for* $|\gamma_0\,(u - \vartheta)/\lambda_0| \gg 1$*. In the high-noise limit it can be shown that the filter* $\mathcal{L}$ *is relatively broad. More precisely, it starts with an initial value* $\mathcal{L}(0) > 0$ *and decays slowly* $\mathcal{L}(x) \to 0$ *for* $x \to \infty$ *(Gerstner 2000). Since decay is small, we set* $\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{L}(x) \approx 0$ *for* $x > 0$*. This yields* $P_i(s) = w_i\, N^{-1}\, \overline{\lambda^{\text{out}}}\, \mathcal{L}(0)\, \epsilon(s) + \int_0^s \mathrm{d}\hat{t}\, f[u(s|\hat{t})]\, S(s|\hat{t})\, P_i(\hat{t})$*. Thus the correlation function* Corr_i *contains a term proportional to* $\epsilon(-s)$ *as in the linear model. In the low-noise limit* ($\gamma_0 \to \infty$)*, we retrieve the threshold process and the filter* $\mathcal{L}$ *approaches a* δ*-function, viz.* $\mathcal{L}(x) = d\,\delta(x)$ *with some constant* $d > 0$ *(Gerstner 2000). In this case, the correlation function* Corr_i *contains a term proportional to the* derivative *of the EPSP, i.e.,* $\mathrm{d}\epsilon(s)/\mathrm{d}s$*.*

To summarize this section, we have shown that it is possible to calculate the correlation function $\mathrm{Corr}_i(s, \mathbf{w})$ for a spiking neuron model with refractoriness. Once we have obtained the correlation function, we can use it in the learning dynamics (23). We have seen that the correlation function depends on the noise level. This theoretical result is in agreement with experimental measurements on motoneurons (Fetz and Gustafsson 1983, Poliakov et al. 1996). It was found that, for high noise levels, the correlation function contained a peak with a time course roughly proportional to the postsynaptic potential. For low noise, however, the time course of the peak was similar to the derivative of the postsynaptic potential Thus the (piecewise-linear) Poisson neuron introduced in Section 4.4 is a valid model in the high-noise limit.

5 Discussion

In this paper we have compared learning rules at the level of spikes with those at the level of rates. The learning rules can be applied to both linear and nonlinear neuron models. Let us discuss our results in the context of the existing experimental and theoretical literature.

5.1 Experimental Results

Rate normalization has been found in experiments performed by Turrigiano et al. (1998,1999). They blocked GABA-mediated inhibition in a cortical culture, which initially raised activity. After about two days, the firing rate returned close to the control value and, at the same time, all synaptic strengths decreased. Conversely, a pharmacologically induced firing rate reduction leads to synaptic strengthening. Again, the output *rate* was normalized. Turrigiano et al. suggest that rate normalization is achieved by multiplying all weights by the same factor. For further mechanisms contributing to the activity-dependent stabilization of firing rates see e.g. (Desai et al. 1999) and references therein.

In our model several scenarios for intrinsic rate normalization are possible, depending on the choice of parameters for the Hebbian and non-Hebbian terms. Let us discuss each of them in turn.

(i) *Sign of correlation term.* Rate normalization in our model is most easily achieved, if the integral of the learning window W is negative, even though this is not necessary (see Examples 1 and 4). On the basis of neurophysiological data known at present (Markram et al. 1997, Zhang et al. 1998, Debanne et al. 1998, Bi and Poo 1998), the sign of the integral over the learning window cannot be decided; see also the review by (Linden 1999). We have shown that a negative integral corresponds, in terms of rate coding, to a negative coefficient a_2^{corr} and hence to an anti-Hebbian rule. Nevertheless the learning rule can pick up positive correlations, if the input contains temporal correlations C_{ij} *on the time scale of the learning window.* An example is given in the Appendix.

(ii) *Role of non-Hebbian terms.* We have seen in Examples 1 and 4 that the linear terms in the learning rule are not necessary for rate normalization. Nevertheless, we see from Example 2 or 3 that a negative coefficient a_1^{out} helps. A value $a_1^{\mathrm{out}} < 0$ means that in the absence of presynaptic activation postsynaptic spiking alone induces heterosynaptic long-term depression. This has been seen, for example, in hippocampal slices (Pockett et al. 1990, Christofi et al. 1993); see also the reviews of (Artola and Singer

1993, Brown and Chattarji 1994, Linden 1999). For cortical slices, both $a_1^{\mathrm{out}} > 0$ and $a_1^{\mathrm{out}} < 0$ have been reported (Volgushev et al. 1994).

On the other hand, a positive effect of presynaptic spikes on the synapses in the absence of postsynaptic spiking (i.e., $a_1^{\mathrm{in}} > 0$) helps to keep the fixed point in the range of positive rates $\lambda_{\mathrm{FP}}^{\mathrm{out}} > 0$; see Eq. (14). Some experiments indeed suggest that presynaptic activity alone results in homosynaptic long-term potentiation ($a_1^{\mathrm{in}} > 0$) (Bliss and Collingridge 1993, Urban and Barrionuevo 1996, Bell et al. 1997); see also (Brown and Chattarji 1994).

(iii) *Zero-order term.* As we have seen from Eq. (14), spontaneous weight growth $a_0 > 0$ helps go keep the fixed point of the rate in the positive range. Turrigiano et al. (1998) chronically blocked cortical culture activity and found an increase of synaptic weights, supporting $a_0 > 0$.

5.2 Spike-Time Dependent Learning Window

In Figure 2a we have indicated a learning window with two parts, a positive part (potentiation) and a negative one (depression). Such a learning window is in accordance with experimental results (Markram et al. 1997, Zhang et al. 1998, Debanne et al. 1998, Bi and Poo 1998). We may ask, however, are there *theoretical* reasons for this time-dependence of the learning window? For excitatory synapses, a presynaptic input spike that precedes postsynaptic firing may be the cause of the postsynaptic activity or, at least, "takes part in firing it" (Hebb 1949). Thus, a literal understanding of the Hebb rule suggests that for excitatory synapses the learning window $W(s)$ is positive for $s < 0$, if s is the difference between the times of occurrence of an input and an output spike. (Recall that $s < 0$ means that a presynaptic spike *precedes* postsynaptic spiking; see also Fig. 2b.) In fact, a positive learning phase (potentiation) for $s < 0$ has been found an important ingredient in models of sequence learning (Herz et al. 1988, Herz et al. 1989, van Hemmen et al. 1990, Gerstner et al. 1993, Abbott and Blum 1996, Gerstner and Abbott 1997). If the positive phase is followed by a negative phase (depression) for $s > 0$ as in Fig. 2a, the learning rule acts as a temporal contrast filter enhancing the detection of temporal structure in the input (Kempter et al. 1996, Gerstner et al. 1996, Kempter et al. 1999a, Song et al. 2000, Xie and Seung 2000). A two-phase learning rule with both potentiation and depression was a theoretical 'prediction' in a strict sense. The advantages of such a learning rule have been realized in models even *before* detailed experimental results on the millisecond time scale became available (Markram et al. 1997, Zhang et al. 1998, Debanne et al. 1998, Bi and Poo 1998).

5.3 Structural Stability of Output Rates

We have seen that synaptic plasticity with an appropriate combination of parameters, pushes the output rate $\overline{\lambda^{\rm out}}$ towards a fixed point. To explicitly show the existence and stability of the fixed point we had to require that the factor C defined by Eq. (12) is independent of the synapse index j. The standard example that we have used throughout the paper is an input scenario that has the same mean rate at each synapse ($\overline{\lambda_i^{\rm in}} = \overline{\lambda^{\rm in}}$ for all i) and is translation invariant $C_{ij} = C_{|i-j|}$. In this case, rate normalization is equivalent to a stable fixed point of the mean weight $N^{-1}\sum_i w_i$ (see Section 3.2).

We may wonder what happens, if Eq. (12) is not exactly true but holds only approximately. To answer this question, let us write the linear model of Eq. (29) in vector notation $\mathrm{d}\mathbf{w}/\mathrm{d}t = k_1\,\mathbf{e}_1 + M\mathbf{w}$ with some matrix M and the unit vector $\mathbf{e}_1 = N^{-1/2}\,(1,1,\ldots,1)^T$. A stable fixed point of the mean weight implies that $\mathbf{w} = \mathbf{e}_1$ is an eigenvector of M with *negative* eigenvalue. If such a negative eigenvalue $\lambda^M < 0$ exists under the condition (12), then, due to (local) continuity of solutions to of the differential equation (29) in dependence upon parameters, the eigenvalue will stay negative in some regime where (12) only holds approximately. The eigenvector corresponding to this negative eigenvalue will be close to, but not identical with $\mathbf{e}_1$. In other words, the output rate remains stable but the weight vector $\mathbf{w}$ is no longer exactly normalized.

The normalization properties of our learning rule are thus akin to, but not identical with, subtractive normalization (Miller and MacKay 1994, Miller 1996b). From our point of view, the basic property of such a rule is a stabilization of the output *rate*. The (approximate) normalization of the mean weight $N^{-1}\sum_i w_i = c$ with some constant c is a consequence. The normalization is exact, if C in Eq. (12) is independent of the synapse index j and mean input rates $\overline{\lambda_j^{\rm in}}$ are the same at all synapses.

Finally, the learning rule (1) with *constant* coefficients is typically unstable because weights receiving strongest enhancement grow without bounds at the expense of synapses receiving less or no enhancement. Unlimited growth can be avoided by explicitly introducing upper and lower bounds for *individual* weights. As a consequence, most of the weights saturate at these bounds. Saturated weights no longer participate in the learning dynamics and the stabilization arguments of Sections 3 and 4 can be applied to the set of remaining weights. For certain models it can be shown explicitly that the fixed point of the output rate remains unchanged compared to the case without bounds (Kempter et al. 1999a). We simply have to check that the fixed point lies within the parameter range allowed by the bounds on the weights.

5.4 Principal Components and Covariance

Whereas the output rate converges to a fixed point so that the mean weight adapts to a constant value $N^{-1}\sum_i w_i = c$, some synapses may grow and others decay. It is this synapse-specific change which leads to learning in its proper sense (in contrast to mere adaptation). Spike-based learning is dominated by the eigenvector of Q_{ij} with the largest eigenvalue (Kempter et al. 1999a, 1999b). Thus learning 'detects' the principal component of the matrix Q_{ij} defined in Eq. (30).

A few remarks are in order. First, we emphasize that, for spike-time dependent learning, a clear-cut distinction between Hebbian and anti-Hebbian rules is difficult. Since Q_{ij} is sensitive to the covariance of the input on the time scale of the learning window, the same rule can be considered as anti-Hebbian for one type of stimulus and as Hebbian for another one; see the Appendix.

Second, in contrast to Oja's rule (Oja 1982) it is not necessary to require that the mean input $\overline{\lambda_i^{\rm in}}$ vanishes at each synapse. Intrinsic rate normalization implies that the mean input level does not play a role. After an initial adaptation phase, neuronal plasticity has automatically 'subtracted' the mean input and becomes sensitive to the covariance of the input; see also the discussion around Eqs. (9) and (16). Thus, *after* convergence to the fixed point, the learning rate becomes equivalent to Sejnowski's covariance rule (Sejnowski 1977, Sejnowski and Tesauro 1989, Stanton and Sejnowski 1989).

5.5 Subthreshold Regime and Coincidence Detection

We have previously exploited the normalization properties of learning rules with negative integral of the learning window in a study of the barn owl auditory system (Gerstner et al. 1996, 1997). Learning led to a structured delay distribution with sub-millisecond time resolution albeit the width of the learning window was in the millisecond range. Intrinsic normalization of the output rate was used to keep the postsynaptic neuron in the subthreshold regime where the neuron functions as a coincidence detector (Kempter et al. 1998). In (Song et al. 2000) the same mechanism has been used to keep the neuron in the subthreshold regime where the neuron may show large output fluctuations. In this regime, rate stabilization induces stabilization of the value of the coefficient of variation (CV) over a broad input regime.

Due to stabilization of output rates, we speculate that a coherent picture of neural signal transmission may emerge where each neuron in a chain of several processing layers is equally active, once averaged over the complete stimulus ensemble. In that case a set of normalized input rates $\overline{\lambda_i^{\rm in}}$ in one 'layer' is transformed, on average, into a set of equal rates in the next layer. Ideally the target value of the rates to which the plasticity rule converges would be in the regime where the neuron is most sensitive, viz., the subthreshold regime. This is the regime, where signal transmission is maximal (Kempter et al. 1998). The plasticity rule could thus be an important ingredient to optimize the neuronal transmission properties (Brenner et al. 1998, Kempter et al. 1998, Stemmler and Koch 1999, Song et al. 2000).

Appendix

We illustrate the static-pattern scenario for the case of a learning window with negative integral. At each time step of length Δ_t, we present a pattern vector $\mathbf{x}^\mu$ which is chosen stochastically from some data base $\{\mathbf{x}^\mu; 1 \le \mu \le p\}$. The input rates are

$$\lambda_i^{\mathrm{in}}(t) = x_i^\mu \quad \text{for} \quad \mu\,\Delta_t < t < (\mu+1)\,\Delta_t \;. \tag{48}$$

The 'static' correlation between patterns is defined to be

$$C_{ij}^{\mathrm{static}} = \frac{1}{p} \sum_{\mu=1}^{p} \frac{\left(x_i^\mu - \overline{\lambda_i^{\mathrm{in}}}\right)\left(x_j^\mu - \overline{\lambda_j^{\mathrm{in}}}\right)}{\overline{\lambda_i^{\mathrm{in}}} \cdot \overline{\lambda_j^{\mathrm{in}}}} \tag{49}$$

where $\overline{\lambda_i^{\mathrm{in}}} := p^{-1} \sum_{\mu=1}^{p} x_i^\mu$ now denotes the average over all input patterns. We assume that there is no intrinsic order in the sequence of presentation of patterns. The 'time-dependent' correlation in the input is then (see Fig. 3)

$$C_{ij}(s) = \begin{cases} C_{ij}^{\mathrm{static}}\,(1 - |s|/\Delta_t) & \text{for } -\Delta_t \le s \le \Delta_t \\ 0 & \text{else.} \end{cases}$$

Let us consider a postsynaptic potential $\epsilon(s) = \delta(s - \Delta_t)$, i.e., the form of the potential is approximated by a delayed localized pulse, a (Dirac) delta function. As a learning window we take (see Fig. 3)

$$W(s) = \begin{cases} 1/\Delta_t & \text{for} \quad -2\,\Delta_t < s \le 0 \\ -1/\Delta_t & \text{for} \quad 0 < s \le 3\,\Delta_t \end{cases}$$

with $\int \mathrm{d}s\, W(s) = -1 =: a_2^{\mathrm{corr}}$. Hence the learning rule could be classified as 'anti-Hebbian'. On the other hand

$$Q_{ij} = \int \mathrm{d}s\, W(s) \int \mathrm{d}s'\, \epsilon(s')\, C_{ij}(s+s') = C_{ij}^{\mathrm{static}} \tag{50}$$

gives the static correlations. Since $C_{ij} = C_{ji}$ also $Q_{ij} = Q_{ji}$ holds and the matrix (Q_{ij}) is Hermitian. Furthermore, $\sum_{ij} C_{ij}^{\mathrm{static}} > 0$ and $\int \mathrm{d}s\, W(s)\,(1 - |s + \Delta_t|/\Delta_t) > 0$ so that $\sum_{ij} Q_{ij} c_i^* c_j > 0$ for all vectors $\mathbf{c} = (c_i) \neq 0$. Hence Q is a positive-definite matrix whose eigenvalues are positive (> 0). The dynamics of learning is dominated by the eigenvector of $(C_{ij}^{\mathrm{static}})$ with the largest eigenvalue just as in standard Hebbian learning (Hertz et al. 1991). We note that, in contrast to standard Hebbian learning, we need not impose

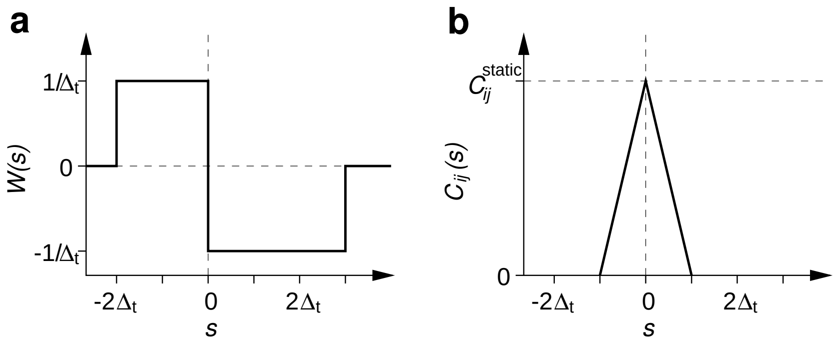

Figure 3: (a) Example of a learning window W and (b) a correlation function C_{ij}. Both are plotted as a function of the time difference s.

$\overline{\lambda_i^{\text{in}}} = 0$; the spike-time dependent learning rule automatically subtracts the mean.

Let us now assume that both $\overline{\lambda_j^{\text{in}}}$ and $Q := (\overline{\lambda^{\text{in}}})^2 \, N^{-1} \sum_i Q_{ij}$ are independent of j; see also Section 4.5. We want to compare the value of Q with that of $\beta = N^{-1} \overline{\lambda^{\text{in}}} \int_{-\infty}^{0} \mathrm{d}s \, W(s) \, \epsilon(-s)$ in Eq. (35). We find

$$\frac{Q}{\beta} = N \, \overline{\lambda^{\text{in}}} \, \Delta_t \, \left[\frac{1}{N} \sum_{i=1}^{N} C_{ij}^{\text{static}} \right] \tag{51}$$

Let us substitute some numbers so as to estimate the order of magnitude of this fraction. The learning window has a duration (width) of, say, the order of $\Delta_t = 10\,\text{ms}$. The mean rate, averaged over the whole stimulus ensemble, is about $\overline{\lambda^{\text{in}}}$ =10 Hz. A typical number of synapses is $N = 1000$. Let us assume that each input channel is correlated with 10 percent of the others with a value of $C_{ij}^{\text{static}} = 0.1$ and uncorrelated with the remaining 90 percent. Thus $N^{-1} \sum_i C_{ij}^{\text{static}} = 0.01$. This yields $Q/\beta = 1$. Hence β has the same order of magnitude as Q. For a lower value of the mean rate, Q/β will decrease, for a larger number of synapses it will increase.

Acknowledgments

We thank Werner Kistler for a critical reading of a first version of the manuscript. Richard Kempter has been supported by the Deutsche Forschungsgemeinschaft (FG Hörobjekte and Ke 788/1-1).

References

Abbott L. F., Blum K. I., 1996, Functional significance of long-term potentiation for sequence learning and prediction. *Cereb. Cortex*, 6:406–416.

Abbott L. F., Munro P., 1999, *Neural Information Processing Systems (NIPS). Workshop on spike timing and synaptic plasticity.* (Breckenridge, CO), http://www.pitt.edu/~pwm/LTP_LTD_99/.

Artola A., Singer W., 1993, Long-term depression of excitatory synaptic transmission and its relationship to long-term potentiation. *Trends Neurosci.*, 16:480–487.

Bell C. C., Han V. Z., Sugawara Y., Grant K., 1997, Synaptic plasticity in a cerebellum-like structure depends on temporal order. *Nature*, 387:278–281.

Bi G.-q., Poo M.-m., 1998, Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.*, 18:10464–10472.

Bi G.-q., Poo M.-m., 1999, Distributed synaptic modification in neural networks induced by patterned stimulation. *Nature*, 401:792–796.

Bienenstock E. L., Cooper L. N., Munro P. W., 1982, Theory for the development of neuron selectivity; orientation specificity and binocular interaction in visual cortex. *J. Neurosci.*, 2:32–48.

Bliss T. V. P., Collingridge G. L., 1993, A synaptic model of memory: long-term potentiation in the hippocampus. *Nature*, 361:31–39.

Brenner N., Agam O., Bialek W., de Ruyter van Steveninck R.R, 1998, Universal statistical behavior of neuronal spike trains. *Phys. Rev. Lett.*, 81(18):4000–4003.

Brown T. H., Chattarji S., 1994, Hebbian synaptic plasticity: Evolution of the contemporary concept. In Domany E., van Hemmen J. L., Schulten K., editors, *Models of Neural Networks II*, pp. 287–314, New York. Springer.

Christofi G., Nowicky A. V., Bolsover S. R., Bindman L. J., 1993, The postsynaptic induction of nonassociative long-term depression of excitatory synaptic transmission in rat hippocampal slices. *J. Neurophysiol.*, 69:219–229.

Debanne D., Gähwiler B. H., Thompson S. M., 1998, Long-term synaptic plasticity between pairs of individual CA3 pyramidal cells in rat hippocampal slice cultures. *J. Physiol.*, 507:237–247.

Desai N. S., Rutherford L. C., Turrigiano G. G., 1999, Plasticity in the intrinsic excitability of cortical pyramidal neurons. *Nat. Neurosci.*, 2:515–520.

Eurich C. W., Pawelzik K., Ernst U., Cowan J. D., Milton J. G., 1999, Dynamics of self-organized delay adaption. *Phys. Rev. Lett.*, 82:1594–1597.

Fetz E. E. and Gustafsson B., 1983, Relation between shapes of post-synaptic potentials and changes in the firing probability of cat motoneurons. *J. Physiol.* 341:387-410.

Gerstner W., 1998, Hebbian learning of Pulse timing in the Barn Owl auditory system. In Maass W., Bishop C. M., editors, *Pulsed Neural Networks*, chapter 14, pp. 261-295. MIT Press, Cambridge, MA.

Gerstner W., 2000, Population dynamics of spiking neurons: fast transients, asynchronous states and locking. *Neural Computation*, 12:43–89.

Gerstner W., Abbott L. F., 1997, Learning navigational maps through potentiation and modulation of hippocampal place cells. *J. Comput. Neurosci.*, 4:79–94.

Gerstner W., van Hemmen J. L., 1992, Associative memory in a network of 'spiking' neurons. *Network*, 3:139–164.

Gerstner W., Ritz R., van Hemmen J. L., 1993, Why spikes? Hebbian learning and retrieval of time-resolved excitation patterns. *Biol. Cybern.*, 69:503–515.

Gerstner W., Kempter R., van Hemmen J. L., Wagner H., 1996, A neuronal learning rule for sub-millisecond temporal coding. *Nature*, 383:76–78.

Gerstner W., van Hemmen J. L., Cowan J. D., 1996b, What matters in neuronal locking. *Neural Comput.*, 8:1689–1712.

Gerstner W., Kempter R., van Hemmen J.L., Wagner H., 1997, A developmental learning rule for coincidence tuning in the barn owl auditory system. In Bower J., editor, *Computational Neuroscience: Trends in Research 1997*, pp. 665–669. Plenum Press, New York.

Gerstner W., Kempter R., van Hemmen J. L., 1998, Hebbian learning of pulse timing in the barn owl auditory system. In Maass W., Bishop C. M., editors, *Pulsed Neural Networks*, chapter 14, pp. 353–377. MIT Press, Cambridge, MA.

Häfliger P., Mahowald M., Watts L., 1997, A spike based learning neuron in analog VLSI. In Mozer M. C., Jordan M. I., Petsche T., editors, *Advances in Neural Information Processing Systems 9*, pp. 692–698. MIT Press, Cambridge, MA.

Hebb D. O., 1949, *The Organization of Behavior*. Wiley, New York.

van Hemmen J. L., 2000, Theory of synaptic plasticity. In Moss F. and Gielen S., editors, *Handbook of Biological Physics, Vol. 4: Neuro-Informatics, Neural Modelling*, Elsevier, Amsterdam, to appear.

van Hemmen J. L., Gerstner W., Herz A., Kühn R., Sulzer B., Vaas M., 1990, Encoding and decoding of patterns which are correlated in space and time. In Dorffner G., editor, *Konnektionismus in Artificial Intelligence und Kognitionsforschung*, pp. 153–162, Springer, Berlin.

Hertz J., Krogh A., Palmer R. G., 1991, *Introduction to the Theory of Neural Computation*. Addison–Wesley, Redwood City.

Herz A. V. M., Sulzer B., Kühn R., van Hemmen J. L., 1988, The Hebb rule: Representation of static and dynamic objects in neural nets. *Europhys. Lett.*, 7:663–669.

Herz A. V. M., Sulzer B., Kühn R., van Hemmen J. L., 1989, Hebbian learning reconsidered: Representation of static and dynamic objects in associative neural nets. *Biol. Cybern.*, 60:457–467.

Kempter R., Gerstner W., van Hemmen J. L., 1999a, Hebbian learning and spiking neurons. *Phys. Rev. E*, 59:4498–4514.

Kempter R., Gerstner W., van Hemmen J. L., 1999b, Spike-based compared to rate-based Hebbian learning. In Kearns M. S., Solla S. A., Cohn D. A., editors, *Advances in Neural Information Processing Systems 11*, pp. 125–131, MIT Press, Cambridge, MA.

Kempter R., Gerstner W., van Hemmen J. L., Wagner H., 1996, Temporal coding in the sub-millisecond range: Model of *barn owl* auditory pathway. In Touretzky D. S., Mozer M. C., Hasselmo M. E., editors, *Advances in Neural Information Processing Systems 8*, pp. 124–130. MIT Press, Cambridge, MA.

Kempter R., Gerstner W., van Hemmen J. L., Wagner H., 1998, Extracting oscillations: Neuronal coincidence detection with noisy periodic spike input. *Neural Comput.*, 10:1987–2017.

Kistler W. M. and van Hemmen J. L., 2000, Modeling Synaptic Plasticity in Conjunction with the Timing of Pre- and Postsynaptic Action Potentials *Neural Comput.*, 12:385-405.

Kohonen T., 1984, *Self-Organization and Associative Memory.* Springer, Berlin.

Levy W. B., Stewart D., 1983, Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus. *Neurosci.*, 8:791–797.

Linden D. J., 1999, The return of the spike: postsynaptic action potentials and the induction of LTP and LTD. *Neuron*, 22:661–666.

Linsker R., 1986, From basic network principles to neural architecture: Emergence of spatial-opponent cells. *Proc. Natl. Acad. Sci. USA*, 83:7508–7512.

MacKay D. J. C., Miller K. D., 1990, Analysis of Linsker's application of Hebbian rules to linear networks. *Network*, 1:257–297.

von der Malsburg C., 1973, Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, 14:85–100.

Markram H., Lübke J., Frotscher M., Sakmann B., 1997, Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, 275:213–215.

Miller K. D., 1996a, Receptive fields and maps in the visual cortex: Models of ocular dominance and orientation columns. In Domany E., van Hemmen J. L., Schulten K., editors, *Models of Neural Networks III*, pp. 55–78, New York. Springer.

Miller K. D., 1996b, Synaptic economics: competition and cooperation in correlation-based synaptic plasticity. *Neuron*, 17:371–374.

Miller K. D., MacKay D. J. C., 1994, The role of constraints in Hebbian learning. *Neural Comput.*, 6:100–126.

Oja E., 1982, A simplified neuron model as a principal component analyzer. *J. Math. Biol.*, 15:267–273.

Plesser H. E., Gerstner W., 2000, Noise in integrate-and-fire models: from stochastic input to escape rates. *Neural Computation*, 12:367-384.

Pockett S., Brookes N. H., Bindman L. J., 1990, Long-term depression at synapses in slices of rat hippocampus can be induced by bursts of postsynaptic activity. *Exp. Brain Res.*, 80:196–200.

Poliakov A. V., Powers R. K., Sawczuk A., Binder M. D., 1996, Effects of background noise on the response of rat and cat motoneurons to excitatory current transients. *J. Physiol.* 495:143-157

Roberts P. D. and Bell C. C., 2000, Computational consequences of temporally asymmetric learning rules: sensory image cancellation. *Computational Neuroscience*, 7:xx-xx (to appear)

Ruf B., Schmitt M., 1997, Unsupervised learning in networks of spiking neurons using temporal coding. In Gerstner W., Germond A., Hasler M., Nicoud J.-D., editors, *Proc. 7th Int. Conf. Artificial Neural Networks (ICANN'97)*, pp. 361–366. Springer, Heidelberg.

Sejnowski T. J., 1977, Storing covariance with nonlinearly interacting neurons *J. Mathematical Biology*, 4:303-321

Sejnowski T. J., Tesauro G., 1989, The Hebb rule for synaptic plasticity: algorithms and implementations. In Byrne J. H., Berry W. O., editors, *Neural Models of Plasticity: Experimental and Theoretical Approaches*, chapter 6, pp. 94–103. Academic Press, San Diego.

Stanton, P. K. and Sejnowski, T. J., 1989, Associative long-term depression in the hippocampus induced by Hebbian covariance, *Nature* 339:215-218.

Senn W., Tsodyks M., Markram H., 1997, An algorithm for synaptic modification based on exact timing of pre- and postsynaptic action potentials. In Gerstner W., Germond A., Hasler M., Nicoud J.-D., editors, *Proc. 7th Int. Conf. Artificial Neural Networks (I-CANN'97)*, pp. 121–126, Heidelberg. Springer.

Senn W., Markram H., Tsodyks M., 2000, An algorithm for modifying neurotransmitter release probability based on pre- and postsynaptic spike timing. *Neural Comput.*, xx:xx-xx (to appear).

Shouval H. Z., Perrone M. P., 1995, Post-Hebbian learning rules. In Arbib M. A., editor, *The Handbook of Brain Theory and Neural Networks*, pp. 745–748. MIT Press, Cambridge, MA.

Song S., Miller K. D., Abbott L. F., 2000, Competitive Hebbian learning through spike-time-dependent synaptic plasticity. *Nat. Neurosci.*, 3:919–926.

Stemmler M., Koch C., 1999, How voltage-dependent conductances can adapt to maximize the information encoded by neurons. *Nat. Neurosci.*, 2:521–527.

Turrigiano G. G., 1999, Homeostatic plasticity in neuronal networks: the more things change, the more they stay the same. *Trends Neurosci.*, 22:221–227.

Turrigiano G. G., Leslie K. R., Desai N. S., Rutherford L. C., Nelson S. B., 1998, Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature*, 391:892–896.

Urban N. N., Barrionuevo G., 1996, Induction of Hebbian and non-Hebbian mossy fiber long-term potentiation by distinct patterns of high-frequency stimulation. *J. Neurosci.*, 16:4293–4299.

Volgushev M., Voronin L. L., Chistiakova M., Singer W., 1994, Induction of LTP and LTD in visual cortex neurons by intracellular tetanization. *NeuroReport*, 5:2069–2072.

Wimbauer S., Gerstner W., van Hemmen J. L., 1994, Emergence of spatio-temporal receptive fields and its application to motion detection. *Biol. Cybern.*, 72:81–92.

Wiskott L., Sejnowski T., 1998, Constrained optimization for neural map formation: a unifying framework for weight growth and normalization. *Neural Comput.*, 10:671–716.

Xie X. and Seung S. H., 2000, Spike-based learning rules and stabilization of persistent neural activity. In Solla S. A., Leen T. K, Müller K.-R., editors, *Advances in Neural Information Processing Systems 12*, pp. xx–xx. MIT Press, Cambridge, MA.

Zhang L. I., Tao H. W., Holt C. E., Harris W. A., Poo M.-m., 1998, A critical window for cooperation and competition among developing retinotectal synapses. *Nature*, 395:37–44.