Cognitive Sciences EPrint Archive (CogPrints) 5281, (2006)
http://cogprints.org/5281/

# Slowness: An Objective for Spike-Timing-Dependent Plasticity?

Henning Sprekeler, Christian Michaelis and Laurenz Wiskott
Institute for Theoretical Biology, Humboldt-Universität Berlin

December 11, 2006

**Abstract**

Slow Feature Analysis (SFA) is an efficient algorithm for learning input-output functions that extract the most slowly varying features from a quickly varying signal. It has been successfully applied to the unsupervised learning of translation-, rotation-, and other invariances in a model of the visual system, to the learning of complex cell receptive fields, and, combined with a sparseness objective, to the self-organized formation of place cells in a model of the hippocampus.

In order to arrive at a biologically more plausible implementation of this learning rule, we consider analytically how SFA could be realized in simple linear continuous and spiking model neurons. It turns out that for the continuous model neuron SFA can be implemented by means of a modified version of standard Hebbian learning. In this framework we provide a connection to the trace learning rule for invariance learning. We then show that for Poisson neurons spike-timing-dependent plasticity (STDP) with a specific learning window can learn the same weight distribution as SFA. Surprisingly, we find that the appropriate learning rule reproduces the typical STDP learning window. The shape as well as the timescale are in good agreement with what has been measured experimentally. This offers a completely novel interpretation for the functional role of spike-timing-dependent plasticity in physiological neurons.

## 1 Introduction

The formation of invariant representations is one of the major challenges the neural system of an organism faces during interaction with the environment. Certain objects have to be identified as being the same when encountered in different situations although the stimuli they produce may be quite different. This means that the processing of sensory information has to be invariant to certain aspects of the stimulus (those varying for the same object) and sensitive to others (those which do not vary and identify the object).

It is widely accepted that the invariances found in the brain are at least partially a product of learning. Because of the limited amount of information in the genome as well as the apparent flexibility of the neural development in different environments, it seems unlikely that the information needed to form invariant representations is already there at the beginning of the individual development. Some must be gathered from the sensory input experienced during interaction with the environment.

One powerful principle for the learning of invariances in an unsupervised fashion is temporal stability, or slowness. A scene that the eye views is very unlikely to change completely from one second to the next. Rather, there is a good chance that an object that can be seen now will also be present at the next instant of time. Therefore sensory signals that are more stable over time are more likely to carry useful information about the world than those that vary quickly. This idea is the basis of a whole class of learning algorithms (Földiak, 1991; Mitchison, 1991; Becker and Hinton, 1992; O'Reilly and Johnson, 1994; Stone and Bray, 1995; Wallis and Rolls, 1997; Peng et al., 1998).

Slow Feature Analysis (SFA,Wiskott and Sejnowski, 2002) is one of these unsupervised learning algorithms based on the slowness principle. Given a multidimensional input signal $\mathbf{x}(t)$ and a finite-dimensional function space $\mathcal{F}$, SFA finds the input-output function $g_1(\mathbf{x}(t))$ in $\mathcal{F}$ that generates the most slowly varying output signal $y_1(t) = g_1(\mathbf{x}(t))$. It is important to note that the function $g_1(\mathbf{x}(t))$ is required to be an instantaneous function of the input signal. Otherwise, slow output signals could be generated by simply lowpass filtering the input signal. As the goal of the slowness principle is to detect slowly varying features of the *input* signals, a mere lowpass filter would certainly generate slow output signals, but it would not serve the purpose.

As a measure of slowness or rather 'fastness' SFA uses the variance of the time derivative, $\langle \dot{y}_1(t)^2 \rangle_t$, which is the objective function to be minimized. Here, $\langle \cdot \rangle_t$ denotes temporal averaging. For mathematical convenience and to avoid the trivial constant response, $y_1(t) = \text{const}$, a zero-mean and unit variance constraint are imposed. Furthermore, it is possible to find a second function $g_2$ extracting $y_2(t) = g_2(\mathbf{x}(t))$ that again minimizes the given objective under the constraint of being uncorrelated with $y_1$, a third one uncorrelated with both $y_1$ and $y_2$ and so on, thereby generating a set of slow features of the input ordered by the degree of 'slowness'. However, in this paper, we will consider just one single unit.

SFA has been successfully applied to the learning of translation-, rotation- and other invariances in a model of the visual system (Wiskott and Sejnowski, 2002) and it has been shown that SFA applied to natural image sequences learns functions that reproduce a wide range of features of complex cells in primary visual cortex (Berkes and Wiskott, 2005). Iteration of the same principle in a hierarchical model in combination with a sparseness objective has been used to model the self-organized formation of spatial representations resembling place cells as found in the hippocampal formation of rodents (Franzius et al., 2006).

Thus on an abstract level SFA seems to capture an important aspect of cortical information processing. However, SFA as a technical algorithm is biologically rather implausible. There is in particular one step in its canonical formulation that seems especially odd compared to what neurons are normally thought to do. In this step the eigenvector that corresponds to the smallest eigenvalue of the covariance matrix of the time derivative of some multidimensional signal is extracted. The aim of this paper is to show how this kind of computation can be realized in a spiking model neuron.

In the first part we will consider a continuous model neuron and demonstrate that a modified Hebbian learning rule enables the neuron to learn the slowest (in the sense of SFA) linear combination of its inputs. In addition, we provide a link to the trace learning rule, which is another implementation of the slowness principle. We then examine if these findings also hold for a spiking model neuron and find that for a linear Poisson neuron spike-timing-dependent plasticity (STDP) can be interpreted as an implementation of the slowness principle.

# 2 Continuous model neuron

## 2.1 Linear model neuron and basic assumptions

First consider a linear continuous model neuron with an input-output function given by

$$a^{\text{out}}(t) = \sum_{i=1}^{n} w_i \, a_i^{\text{in}}(t) \,, \tag{1}$$

with $a_i^{\text{in}}$ indicating the input signals, $w_i$ the weights, and $a^{\text{out}}$ the output signal. For mathematical convenience, let $a_i^{\text{in}}$ and $a^{\text{out}}$ be defined on the interval $[-\infty, \infty]$ but differ from zero only on $[0, T]$, which could be the lifetime of the system. We assume that the input is approximately whitened on any sufficiently large interval $[t_a, t_b] \subseteq [0, T]$, i.e. each input signal has approximately zero mean and unit variance and is uncorrelated to other input signals:

$$\int_{t_a}^{t_b} a_i^{\text{in}}(t)\,\mathrm{d}t \approx 0 \quad \text{(zero mean)}, \tag{2}$$

$$\int_{t_a}^{t_b} \left(a_i^{\text{in}}(t)\right)^2 \mathrm{d}t \approx 1 \quad \text{(unit variance)}, \tag{3}$$

$$\int_{t_a}^{t_b} a_i^{\text{in}}(t)\, a_{j\neq i}^{\text{in}}(t)\,\mathrm{d}t \approx 0 \quad \text{(decorrelation)}\,. \tag{4}$$

This can be achieved by a normalization and decorrelation step of the units projecting to the considered unit. Furthermore, we assume that the output is normalized to unit variance, which for whitened input means that the weight vector is normalized to length one. In an online learning rule this could be implemented by either an activity dependent or a weight dependent normalization term. Thus for the output signal we have

$$\int_{t_a}^{t_b} a^{\text{out}}(t)\,\mathrm{d}t \overset{(1,2)}{\approx} 0 \quad \text{(zero mean)}, \tag{5}$$

$$\int_{t_a}^{t_b} \left(a^{\text{out}}(t)^2\right)\,\mathrm{d}t \overset{(1,3)}{\approx} \sum_{i=1}^{n} w_i^2 := 1 \quad \text{(unit variance)}. \tag{6}$$

In the following we will often consider filtered signals. Therefore we introduce abbreviations for the convolution $f \circ g$ and the cross-correlation $f \star g$ of two functions $f(t)$ and $g(t)$:

$$\text{Convolution:} \quad [f \circ g](t) := \int_{-\infty}^{\infty} f(\tau) g(t-\tau)\,\mathrm{d}\tau\,, \tag{7}$$

$$\text{Cross-correlation:} \quad [f \star g](t) := \int_{-\infty}^{\infty} f(\tau) g(t+\tau)\,\mathrm{d}\tau\,. \tag{8}$$

For convenience, we will often use windowed signals, indicated by a hat

$$\hat{s}(t) := \begin{cases} s(t) & \text{if } t \in [t_a, t_b] \\ 0 & \text{otherwise} \end{cases}\,, \tag{9}$$

which allows us to replace the integration of a signal $s(t)$ over $[t_a, t_b]$ by an integration of $\hat{s}(t)$ over $[-\infty, \infty]$. We assume that the interval $[t_a, t_b]$ is long compared to the width of the filters. In this case effects from the integration boundaries are negligible and we have

$$\int_{t_a}^{t_b} [f \circ s](t) h(t)\,\mathrm{d}t \approx \int_{-\infty}^{\infty} [f \circ \hat{s}](t) h(t)\,\mathrm{d}t\,. \tag{10}$$

Similar considerations hold for the cross-correlation (8).

Since convolution and cross-correlation are conveniently treated in Fourier space, we repeat also the definition of the Fourier transform $\mathcal{F}_s(\nu)$ and the power spectrum $\mathcal{P}_s(\nu)$ of a signal $s(t)$.

$$\text{Fourier transform:} \quad s(t) =: \int_{-\infty}^{\infty} \mathcal{F}_s(\nu)\, e^{2\pi i \nu t}\,\mathrm{d}\nu\,, \tag{11}$$

$$\text{Power spectrum:} \quad \mathcal{P}_s(\nu) := \mathcal{F}_s(\nu)\overline{\mathcal{F}_s}(\nu)\,. \tag{12}$$

Furthermore, we make the assumption that input signals (and hence also the output signals) do not have significant power above some reasonable frequency $\nu_{\max}$.

## 2.2 Reformulation of the slowness objective

SFA is based on the minimization of the second moment of the time derivative, $\int \dot{a}^{\mathrm{out}}(t)^2 \, \mathrm{d}t$. Even though there are neurons with transient responses to changes in the input, we believe it would be more plausible if we could derive an SFA-learning rule that does not depend on the time derivative, because it might be difficult to extract, especially for spiking neurons. It is possible to replace the time derivative by a low-pass filtering as follows.

$$
\begin{aligned}
& \text{minimize} && \int_{-\infty}^{\infty} \dot{a}^{\mathrm{out}}(t)^2 \, \mathrm{d}t && (13)\\
& && = \int_{-\infty}^{\infty} \mathcal{P}_{\dot{a}^{\mathrm{out}}}(\nu) \, \mathrm{d}\nu \qquad \text{(because of Parseval's theorem)} && (14)\\
& && = 4\pi^2 \int_{-\infty}^{\infty} \nu^2 \mathcal{P}_{a^{\mathrm{out}}}(\nu) \, \mathrm{d}\nu \qquad (\text{since } \mathcal{F}_{\dot{s}}(\nu) = 2\pi \mathrm{i} \nu \mathcal{F}_s(\nu)) && (15)\\
\iff & \text{maximize} && \int_{-\infty}^{\infty} -\nu^2 \mathcal{P}_{a^{\mathrm{out}}}(\nu) \, \mathrm{d}\nu && (16)\\
\iff & \text{maximize} && \int_{-\infty}^{\infty} (\nu_{\max}^2 - \nu^2) \mathcal{P}_{a^{\mathrm{out}}}(\nu) \, \mathrm{d}\nu && (17)\\
& && \left(\text{since } \int_{-\infty}^{\infty} \mathcal{P}_{a^{\mathrm{out}}}(\nu) \, \mathrm{d}\nu = \int_{-\infty}^{\infty} a^{\mathrm{out}}(t)^2 \, \mathrm{d}t \overset{(6)}{\approx} \text{const}\right) &&\\
& && = \int_{-\infty}^{\infty} \max(0, (\nu_{\max}^2 - \nu^2)) \mathcal{P}_{a^{\mathrm{out}}}(\nu) \, \mathrm{d}\nu && (18)\\
& && (\text{since } \mathcal{P}_{a^{\mathrm{out}}}(\nu) = 0 \text{ for } |\nu| > \nu_{\max} \text{ by assumption}) &&\\
& && = \int_{-\infty}^{\infty} \mathcal{P}_{f_{\mathrm{SFA}}}(\nu) \mathcal{P}_{a^{\mathrm{out}}}(\nu) \, \mathrm{d}\nu && (19)\\
& && (\text{with } f_{\mathrm{SFA}}(t) \text{ defined such that } \mathcal{P}_{f_{\mathrm{SFA}}} = \max(0, (\nu_{\max}^2 - \nu^2))) && (20)\\
& && = \int_{-\infty}^{\infty} \left[ f_{\mathrm{SFA}} \circ a^{\mathrm{out}} \right](t)^2 \, \mathrm{d}t \, . && (21)
\end{aligned}
$$

Thus, SFA can either be achieved by minimizing the variance of the time derivative of the output signal or by maximizing the variance of the appropriately filtered output signal. Figure 1 provides an intuition for this alternative. The filter $f_{\mathrm{SFA}}$ is obviously a low-pass filter, as one would expect, with a $(\nu_{\max}^2 - \nu^2)$-power spectrum below the limiting frequency $\nu_{\max}$. Because the phases are not determined, further assumptions are required to fully determine an SFA-filter. However, we will proceed without defining a concrete filter, since it is not required for the considerations below.

## 2.3 Hebbian learning on filtered signals

It is known that standard Hebbian learning under the constraint of a unit weight vector applied to a linear unit maximizes the variance of the output signal. We have seen in the previous section that SFA can be reformulated as a maximization problem for the variance of the low-pass filtered output-signal. To achieve this we simply apply Hebbian learning to the filtered input- and output-signals instead of the original signals.

Consider a hypothetical unit that receives low-pass filtered inputs and therefore, because of the linearity of the unit and the filtering, generates a low-pass filtered output

$$[f_{\mathrm{SFA}} \circ a^{\mathrm{out}}](t) \overset{(1)}{=} \left[ f_{\mathrm{SFA}} \circ \sum_{i=1}^{n} w_i \, a_i^{\mathrm{in}} \right](t) = \sum_{i=1}^{n} w_i \, [f_{\mathrm{SFA}} \circ a_i^{\mathrm{in}}](t) \, , \tag{22}$$

where $f_{\mathrm{SFA}}$ is the kernel of the linear filter applied. It is obvious that a *filtered Hebbian learning rule*

$$\dot{w}_i = \gamma \, [f^{\mathrm{in}} \circ a_i^{\mathrm{in}}](t) \, [f^{\mathrm{out}} \circ a^{\mathrm{out}}](t) \tag{23}$$

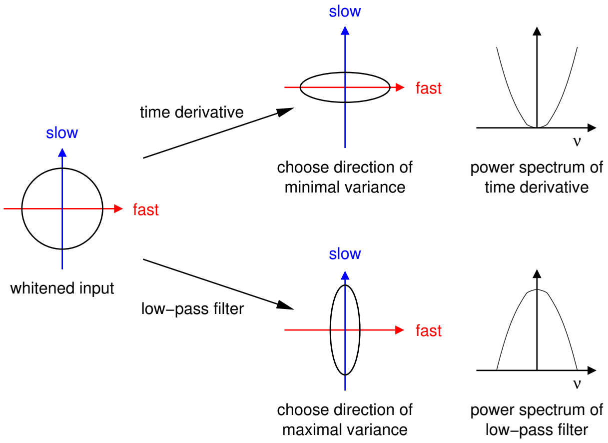


**Figure 1: Choosing slow directions of the input**. Finding the direction of least variance in the time derivative of the input (which is part of the SFA algorithm) can be replaced by finding the direction of maximum variance in an appropriately lowpass filtered version of the input signal.

with $f^{\text{in}} := f^{\text{out}} := f_{\text{SFA}}$ maximizes (21).

Remember that the input is white, i.e. the $a_i^{\text{in}}$ are uncorrelated and have unit variance, and the weight vector is normalized to norm one by some additional normalization rule, so that we know that the output signal $a^{\text{out}}$ has the same variance no matter what the direction of the weight vector is. Thus, the filtered Hebbian plasticity rule (together with the normalization rule not specified here) optimizes slowness (13) under the constraint of unit variance (6). Figure 2 illustrates this learning scheme. It also underlines the necessity for a clear distinction between processing and learning. Although the slowness principle does not allow lowpass filtering as a means of generating slow signals during processing, the learning rule may well make use of lowpass filtered signals in order to detect slowly varying features in the input signal. This distinction will become particularly important for the Poisson model neuron below, as it incorporates an EPSP that acts as a lowpass filter during processing. An implementation of the slowness principle in such a system must avoid that the system exploits the EPSP as a means of generating slow signals.

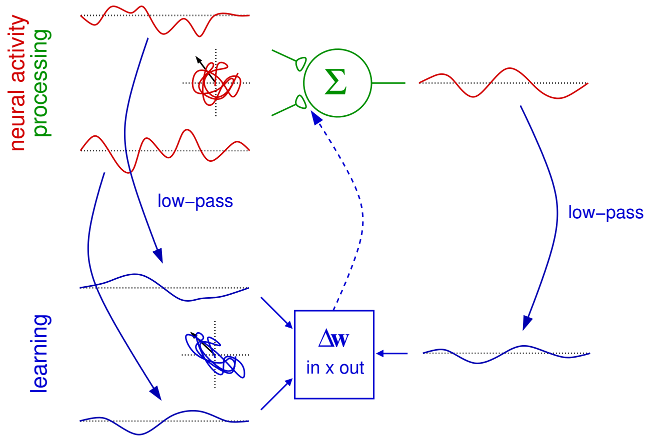


**Figure 2: 'Filtered Hebbian' learning rule**. Input and output signals are filtered (downward arrows). The weight change is the result of applying the Hebbian learning rule on the filtered signals (square box and upward arrow). Thereby, the variance of the filtered version of the output is maximized without actually filtering the output.

## 2.4 Alternative filtering procedures

If learning is slow, the total weight change over a time interval $[t_a, t_b]$ in this synapse can be written as

$$\begin{aligned}
\Delta w_i \quad &:= \quad \int_{t_a}^{t_b} \dot{w}_i \, \mathrm{d}t && (24)\\
&\overset{(23)}{=} \quad \gamma \int_{t_a}^{t_b} [f^{\mathrm{in}} \circ a_i^{\mathrm{in}}](t)\, [f^{\mathrm{out}} \circ a^{\mathrm{out}}](t)\, \mathrm{d}t && (25)\\
&\overset{(10)}{\approx} \quad \gamma \int_{-\infty}^{\infty} [f^{\mathrm{in}} \circ \hat{a}_i^{\mathrm{in}}](t)\, [f^{\mathrm{out}} \circ \hat{a}^{\mathrm{out}}](t)\, \mathrm{d}t && (26)\\
&= \quad \gamma \int_{-\infty}^{\infty} [[f^{\mathrm{out}} \star f^{\mathrm{in}}] \circ \hat{a}_i^{\mathrm{in}}](t)\, \hat{a}^{\mathrm{out}}(t)\, \mathrm{d}t && (27)\\
&= \quad \gamma \int_{-\infty}^{\infty} \hat{a}_i^{\mathrm{in}}(t)\, [[f^{\mathrm{in}} \star f^{\mathrm{out}}] \circ \hat{a}^{\mathrm{out}}](t)\, \mathrm{d}t && (28)\\
&= \quad \gamma \int_{-\infty}^{\infty} [f^{\mathrm{in}} \star f^{\mathrm{out}}](t)\, [\hat{a}^{\mathrm{out}} \star \hat{a}_i^{\mathrm{in}}](t)\, \mathrm{d}t\,. && (29)
\end{aligned}$$

Thus one can either convolve input and output signal with filters $f^{\mathrm{in}}$ and $f^{\mathrm{out}}$, respectively, the input signal with $f^{\mathrm{out}} \star f^{\mathrm{in}}$, or the output signal with $f^{\mathrm{in}} \star f^{\mathrm{out}}$. Note that $[f^{\mathrm{in}} \star f^{\mathrm{out}}](t) = [f^{\mathrm{out}} \star f^{\mathrm{in}}](-t)$. One can actually use any pair of filters $f^{\mathrm{in}}$ and $f^{\mathrm{out}}$ as long as $f^{\mathrm{in}} \star f^{\mathrm{out}}$ fulfills the condition

$$\mathcal{F}_{f^{\mathrm{in}} \star f^{\mathrm{out}}}(\nu) = \mathcal{P}_{f_{\mathrm{SFA}}}(\nu)\,. \tag{30}$$

## 2.5 Relation to other learning rules

Hebbian learning on lowpass-filtered signals is the basis of several other models for unsupervised learning of invariances (Földiak, 1991; O'Reilly and Johnson, 1994; Wallis and Rolls, 1997). These models essentially subject the output signal to an exponential temporal filter $f(t) := \theta(t)\gamma \exp(-\gamma t)$ and then use Hebbian learning to associate it with the input signal. Here $\theta(t)$ denotes the Heaviside step function, which is 0 for $t < 0$ and 1 for $t \geq 0$. This learning rule has been termed the 'trace rule'. The considerations in the last section provide a link between this approach and ours. We simply have to replace $f^{\mathrm{in}}(t)$ by a $\delta$-function and $f^{\mathrm{out}}(t)$ by $f(t)$. Equation (29) then takes the form

$$\Delta w_i = \gamma \sum_j \left[ \int_{-\infty}^{\infty} f(t)\,[\hat{a}_j^{\mathrm{in}} \star \hat{a}_i^{\mathrm{in}}](t)\,\mathrm{d}t \right] w_j\,, \tag{31}$$

since the output signal $a^{\mathrm{out}} = \sum_j w_j a_j^{\mathrm{in}}$ is a linear function of the input (1). In the previously mentioned applications of the 'trace rule', the statistics of the input signals was always reversible, so we will assume that all correlation functions $[a_i^{\mathrm{in}} \star a_j^{\mathrm{in}}](t)$ are symmetric in time. This implies that only the symmetric component of $f(t)$ is relevant for learning:

$$f^{\mathrm{sym}}(t) := \frac{1}{2}(f(t) + f(-t)) = \frac{\gamma}{2}\exp(-\gamma|t|). \tag{32}$$

It is easy to show that the learning rule (31) can be interpreted as a gradient ascent on the following objective function:

$$\begin{aligned} \Psi &= \int_{-\infty}^{\infty} f^{\mathrm{sym}}(t)[a^{\mathrm{out}} \star a^{\mathrm{out}}](t)\,\mathrm{d}t & (33)\\ &= \int_{-\infty}^{\infty} \mathcal{F}_{f^{\mathrm{sym}}}(\nu)\mathcal{P}_{a^{\mathrm{out}}}(\nu)\,\mathrm{d}\nu\,. & (34) \end{aligned}$$

By comparison with equation (19), it becomes clear that the 'trace rule' implements a very similar objective as our model. The only difference is that the power spectrum (20) is replaced by the Fourier transform of the filter $f^{\mathrm{sym}}$. Note that in order to be able to interpret $\Psi$ as an objective function, it should be real-valued. The replacement of $f$ by $f^{\mathrm{sym}}$ ensures that $\mathcal{F}_{f^{\mathrm{sym}}}$ is real-valued and symmetric, so $\Psi$ is real-valued as well. The Fourier transform of $f^{\mathrm{sym}}$ is given by

$$\mathcal{F}_{f^{\mathrm{sym}}}(\nu) = \frac{\gamma}{\gamma^2 + (2\pi\nu)^2}. \tag{35}$$

This shows that the only difference between the 'trace rule' and our model lies in the choice of the power spectrum for the lowpass filter. While we are using a parabolic power spectrum with a cutoff (20), the 'trace rule' uses a power spectrum with the shape of a Cauchy function (35).

From this perspective, one can interpret SFA as a quadratic approximation of the 'trace rule'. To what extent this approximation is valid depends on the power spectra of the input signals. If most of the input power is concentrated at low frequencies where the power spectrum resembles a parabola, the learning rules can be expected to learn very similar weight vectors. In fact any Hebbian learning rule that leads to an objective function of the shape of equation (19) with a lowpass filtering spectrum in the place of $\mathcal{P}_{f_{\mathrm{SFA}}}$ essentially implements the slowness principle, as among signals with the same variance, it will favor slower ones.

# 3 Spiking model neuron

Real neurons do not transmit information via a continuous stream of analog values like the model neuron considered in the previous section, but rather emit action potentials that carry information by means of their rate and probably also by their exact timing, a fact we will not consider here. How can the model developed so far be mapped onto this scenario?

## 3.1 The linear Poisson neuron

Again, we restrict our analysis to a simple case by modeling the spike train signals by inhomogeneous Poisson processes. First, sufficiently large constants $c_i^{\mathrm{in}}$ are added to the continuous and zero-mean signals $a_i^{\mathrm{in}}(t)$ to turn them into strictly positive signals that can be interpreted as rates

$$r_i^{\mathrm{in}}(t) \;:=\; c_i^{\mathrm{in}} + a_i^{\mathrm{in}}(t)\,. \tag{36}$$

The constants $c_i^{\mathrm{in}}$ represent mean firing rates, which are modulated by the input signals $a_i^{\mathrm{in}}$. From the input rates $r_i^{\mathrm{in}}(t)$ we then derive inhomogeneous Poisson spike trains $S_i^{\mathrm{in}}(t)$ drawn from ensembles $E_i^{\mathrm{in}}$ such that

$$\langle S_i^{\mathrm{in}}(t) \rangle_{E_i^{\mathrm{in}}} = r_i^{\mathrm{in}}(t)\,, \tag{37}$$

where $\langle \cdot \rangle_{E_i^{\mathrm{in}}}$ denotes the average over the ensemble $E_i^{\mathrm{in}}$.

The output rate is modeled as a weighted sum over the input spike trains convolved with an EPSP $\epsilon(t)$ plus a baseline firing rate $r_0$, which ensures that the output firing rate remains positive. This is necessary as we allow inhibitory synapses, i.e. negative weights.

$$m(t) := r_0 + \sum_{i=1}^{n} w_i \, [\epsilon \circ S_i^{\mathrm{in}}](t)\,. \tag{38}$$

The output of this spiking neuron is yet another inhomogeneous Poisson spike train $S^{\mathrm{out}}(t)$ drawn from an ensemble $E^{\mathrm{out}}$ given a realization of the input spike-trains $S_i^{\mathrm{in}}$ such that

$$\langle S^{\mathrm{out}}(t) \rangle_{E^{\mathrm{out}}|\{S_i^{\mathrm{in}}\}} = m(t)\,. \tag{39}$$

It should be noted that not only the output spike train $S^{\mathrm{out}}(t)$ is stochastic in this model, but also the underlying output rate $m(t)$, which is a function of the stochastic variables $S_i^{\mathrm{in}}(t)$ and generally differs for each realization of the input. This is the reason why the input and output spike trains are not statistically independent. However, due to the linearity of the model neuron the output rate is still simply

$$\begin{aligned}
r^{\mathrm{out}}(t) \quad &:= \quad \langle S^{\mathrm{out}}(t) \rangle_{E^{\mathrm{out}},E^{\mathrm{in}}} && (40)\\
&\overset{(39,38,37)}{=} \quad r_0 + \sum_{i=1}^{n} w_i \, [\epsilon \circ r_i^{\mathrm{in}}](t) && (41)\\
&\overset{(36)}{=} \quad \underbrace{r_0 + \sum w_i c_i^{\mathrm{in}} \int_{-\infty}^{\infty} \epsilon(t)\,\mathrm{d}t}_{=:c^{\mathrm{out}}} + \sum_{i=1}^{n} w_i \, [\epsilon \circ a_i^{\mathrm{in}}](t) && (42)\\
&= \quad c^{\mathrm{out}} + \left[ \epsilon \circ \sum_{i=1}^{n} w_i \, a_i^{\mathrm{in}} \right](t) && (43)\\
&\overset{(1)}{=} \quad c^{\mathrm{out}} + [\epsilon \circ a^{\mathrm{out}}](t)\,, && (44)
\end{aligned}$$

and the joint firing rate is

$$\begin{aligned}
r_i^{\mathrm{in,out}}(t,t') \quad &:= \quad \langle S_i^{\mathrm{in}}(t) S^{\mathrm{out}}(t') \rangle_{E^{\mathrm{out}},E^{\mathrm{in}}} && (45)\\
&= \quad r_i^{\mathrm{in}}(t) r^{\mathrm{out}}(t') + w_i \epsilon(t'-t) r_i^{\mathrm{in}}(t) \quad \text{(see Kempter et al., 1999).} && (46)
\end{aligned}$$

The first term would result also from a rate model, while the second term captures the statistical dependencies between input and output spike-trains mediated by the synaptic weights $w_i$ and the EPSP $\epsilon$.

## 3.2 Spike-timing-dependent plasticity can perform SFA

In this section we will demonstrate that in an ensemble-averaged sense it is possible to generate the same weight distribution as in the continuous model by means of a spike-timing-dependent plasticity (STDP) rule with a specific learning window.

Synaptic plasticity that depends on the temporal order of pre- and postsynaptic spikes has been found in a number of neuronal systems (Debanne et al., 1994; Markram et al., 1997; Bi and Poo, 1998; Zhang et al., 1998; Feldman, 2000). Typically, synapses undergo long-term potentiation (LTP) if a presynaptic spike precedes a postsynaptic spike within a time scale of tens of milliseconds and long-term depression (LTD) for the opposite temporal order. Assuming that the change in synaptic efficacy occurs on a slower time scale than the typical interspike interval, the STDP weight dynamics can be modeled as

$$\Delta w_i = \gamma \sum_{\alpha}^{m_i^{\text{in}}} \sum_{\beta}^{m^{\text{out}}} W(t_{i\alpha}^{\text{in}} - t_{\beta}^{\text{out}}) \,. \tag{47}$$

Here $t_{i\alpha}^{\text{in}}$ denotes the spike times of the presynaptic spikes at synapse $i$ and $t_{\beta}^{\text{out}}$ the postsynaptic spike times. $W(t)$ is the learning window that determines if and to what extent the synapse is potentiated or depressed by a single spike pair. The convention is such that negative arguments $t$ in $W(t)$ correspond to the situation where the presynaptic spike precedes the postsynaptic spike. $m^{\text{in}}$ and $m^{\text{out}}$ are the numbers of pre- and postsynaptic spikes occurring in the time interval $[t_a, t_b]$ under consideration. $\gamma$ is a small positive learning rate. Note that due to the presence of this learning rate the absolute scale of the learning window $W$ is not important for our analysis.

We circumvent the well-known stability problem of STDP by applying an explicit weight normalization ($||\vec{w}|| =$const.) instead of weight-dependent learning rates as used elsewhere (Kistler and van Hemmen, 2000; Rubin et al., 2001; Gütig et al., 2003). Such a normalization procedure could be implemented by means of a homeostatic mechanism targeting the output firing rate, e.g. by synaptic scaling (for reviews see Turrigiano and Nelson, 2000; Abbott and Nelson, 2000).

Modeling the spike trains as sums of delta pulses, i.e. $S^{\text{in/out}} = \sum_j \delta(t - t_j^{\text{in/out}})$, the learning rule (47) can be rewritten as

$$\begin{aligned} \Delta w_i &= \gamma \int_{t_a}^{t_b} \int_{t_a}^{t_b} W(t-t') S_i^{\text{in}}(t) S^{\text{out}}(t') \, \mathrm{d}t \, \mathrm{d}t' && (48) \\ &\approx \gamma \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W(t-t') \hat{S}_i^{\text{in}}(t) \hat{S}^{\text{out}}(t') \, \mathrm{d}t \, \mathrm{d}t'. && (49) \end{aligned}$$

Taking the ensemble average allows us to retrieve the rates that underlie the spike trains and thus the signals $\hat{a}_i^{\text{in}}$ and $\hat{a}^{\text{out}}$ of the continuous model:

$$\begin{aligned} \langle \Delta w_i \rangle_{E^{\text{in}}, E^{\text{out}}} &\overset{(49)}{\approx} \gamma \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W(t-t') \langle \hat{S}_i^{\text{in}}(t) \hat{S}^{\text{out}}(t') \rangle_{E^{\text{in}}, E^{\text{out}}} \, \mathrm{d}t \, \mathrm{d}t' && (50) \\ &\overset{(46)}{=} \gamma \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W(t-t') \big( \hat{r}_i^{\text{in}}(t) \hat{r}^{\text{out}}(t') + w_i \epsilon(t'-t) \hat{r}_i^{\text{in}}(t) \big) \, \mathrm{d}t \, \mathrm{d}t' && (51) \\ &\overset{(36,44)}{\approx} \gamma \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W(t-t') [\hat{c}_i^{\text{in}} + \hat{a}_i^{\text{in}}](t) [\hat{c}^{\text{out}} + \epsilon \circ \hat{a}^{\text{out}}](t') \, \mathrm{d}t \, \mathrm{d}t' \\ &\quad + \gamma \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W(t-t') w_i \epsilon(t'-t) [\hat{c}_i^{\text{in}} + \hat{a}_i^{\text{in}}](t) \, \mathrm{d}t \, \mathrm{d}t' \,. && (52) \end{aligned}$$

Expanding the products in equation (52) gives rise to a number of terms, among which only one depends on both the input and the output signal $\hat{a}_i^{\text{in}}$ and $\hat{a}^{\text{out}}$. Because all signals have vanishing mean, terms containing just one of these signals lead to negligible contributions. The remaining terms depend only on the mean firing rates $c_i^{\text{in}}$ and $c^{\text{out}}$:

$$\begin{aligned} \langle \Delta w_i \rangle_{E^{\text{in}}, E^{\text{out}}} &\overset{(52)}{\approx} \gamma \int_{-\infty}^{\infty} W(t-t') \hat{a}_i^{\text{in}}(t) [\epsilon \circ \hat{a}^{\text{out}}](t') \, \mathrm{d}t \\ &\quad + \gamma w_i c_i^{\text{in}} T_{ab} \int_{-\infty}^{\infty} W(t) \epsilon(-t) \, \mathrm{d}t \\ &\quad + \gamma c^{\text{out}} c_i^{\text{in}} T_{ab} \int_{-\infty}^{\infty} W(t) \, \mathrm{d}t \,. \end{aligned} \tag{53}$$

The decisive term is the first one. The other two are rather unspecific in that they do not depend on the properties of the input and output signals $\hat{a}_i^{\text{in}}$ and $\hat{a}^{\text{out}}$.

The second term alone would generate a competition between the weights: Synapses that experience a higher mean input firing rate $c_i^{\mathrm{in}}$ grow more rapidly than those with smaller input firing rates. If we assume that the input neurons fire with the same mean firing rate, all weights grow with the same rate, i.e. the direction of the weight vector remains unchanged. Due to the explicit weight normalization this term has no effect on the weight dynamics and can be neglected.

If the integral over the learning window is positive, the third term in equation (53) favors a weight vector that is proportional to the vector of the mean firing rates of the input neurons. It thus stabilizes the homogeneous weight distribution and opposes the effect of the first term, which captures correlations in the input signals. Note that this is only true if the integral over the learning window is positive, otherwise this term introduces a competition between the weights (cf. Gütig et al., 2003, equation (8)). One possible interpretation is that the neuron has a 'default state' in which all synapses are equally strong and that correlations in the input need to surpass a certain threshold in order to be imprinted in the synaptic connections. Interestingly, this threshold is determined by the integral over the learning window, which implies that neurons that balance LTP and LTD should be more sensitive to input correlations.

An alternative possibility is that the neuron possesses a mechanism of canceling the effects of this term. From a computational perspective this would be sensible, as the mean firing rates $c_i^{\mathrm{in}}$ and $c^{\mathrm{out}}$ do not carry information about the input, neither in rate nor in a timing code. If we conceive neurons as information encoders aiming at adapting to the structure of their input, this term is thus more hindrance than help. Assuming that the neuron compensates for this term, the dynamics of the synaptic weights are governed exclusively by the correlations in the input signals as reflected by the first term. In the following we will restrict our considerations to this term and omit the others.

Rearranging the temporal integrations, we can rewrite the equation for the weight updates as

$$\langle \Delta w_i \rangle_{E^{\mathrm{in}},E^{\mathrm{out}}} \overset{(53)}{\approx} \gamma \int_{-\infty}^{\infty} [W \circ \epsilon](t) [\hat{a}^{\mathrm{out}} \star \hat{a}^{\mathrm{in}}](t) \, \mathrm{d}t \,. \tag{54}$$

The first conclusion we can draw from this reformulation is that for the dynamics of the learning process the convolution of the learning window with the EPSP and not the learning window alone is relevant. As discussed in section 3.4, this might have important consequences for functional interpretations of the shape of the learning window.

Second, by comparison with equation (29), it is obvious that in order to learn the same weight distribution as in the continuous model, the learning window has to fulfill the condition that

$$[W \circ \epsilon](t) \;=\; [f^{\mathrm{in}} \star f^{\mathrm{out}}](t) \;=:\; W_0(t) \tag{55}$$

$$\iff \quad \mathcal{F}_{W\circ\epsilon}(\nu) \;=\; \mathcal{F}_W(\nu)\mathcal{F}_\epsilon(\nu) \;=\; \mathcal{F}_{f^{\mathrm{in}}\star f^{\mathrm{out}}}(\nu) \;=\; \mathcal{P}_{f_{\mathrm{SFA}}}(\nu) \;=\; \mathcal{F}_{W_0}(\nu). \tag{56}$$

Here, $W_0$ is the convolution of $W$ with $\epsilon$ and is equal to the learning window in the limit of an infinitely short, $\delta$-shaped EPSP. As the power spectrum $\mathcal{P}_{f_{\mathrm{SFA}}}(\nu)$ is of course real, $W_0$ is symmetric in time. Note that the width of $W_0$ scales inversely with the width of the power spectrum $\mathcal{P}_{f_{\mathrm{SFA}}}$, which in turn is proportional to $\nu_{\mathrm{max}}$. Once the power spectrum $\mathcal{P}_{f_{\mathrm{SFA}}}$ and the EPSP is given, equation (56) uniquely determines the learning window $W$.

## 3.3 Learning windows

According to the last section, we require special learning windows in order to learn the slow directions in the input. This of course raises the question which window shapes are favorable and in particular if these are in agreement with physiological findings.

Given the shape of the EPSP and the power spectrum $\mathcal{P}_{f_{\mathrm{SFA}}}$, the learning window is uniquely determined by equation (56). Remember that the only parameter in the power spectrum $\mathcal{P}_{f_{\mathrm{SFA}}}$ is the frequency $\nu_{\mathrm{max}}$, above which the power spectrum of the input data was assumed to vanish. For simplicity, we model the EPSP as a single exponential with a time constant $\tau$:

$$\epsilon(t) = \theta(t)\, \mathrm{e}^{-\frac{t}{\tau}} \,. \tag{57}$$

For this particular EPSP shape, the learning window can be calculated analytically by inverting the Fourier transform in (56). The result can be written as

$$W(t) \quad = \quad \left[\frac{\mathrm{d}}{\mathrm{d}t} + \frac{1}{\tau}\right] W_0(t)\,. \tag{58}$$

$W_0$ is symmetric, so its derivative is antisymmetric. Thus, the learning window is a linear combination of a symmetric and an antisymmetric component. As the width of $W_0$ scales with the inverse of $\nu_{\max}$, its temporal derivative scales with $\nu_{\max}$. Accordingly, the symmetry of the learning window is governed by an interplay of the duration $\tau$ of the EPSP and the maximal input frequency $\nu_{\max}$. For $\tau \ll 1/\nu_{\max}$ the learning window is dominated by $W_0$ and thus symmetric whereas for $\tau \gg 1/\nu_{\max}$ the temporal derivative of $W_0$ is dominant, so the learning window is anti-symmetric.

We have assumed that the input signals have negligible power above the maximal input frequency $\nu_{\max}$. Thus, the temporal structure of the input signals can only provide a lower bound for $\nu_{\max}$. On the other hand, exceedingly high values for $\nu_{\max}$ lead to very narrow learning windows, thereby sharpening the coincidence detection and reducing the speed of learning. Moreover, it may be metabolically costly to implement physiological processes that are faster than necessary. Thus, it appears sensible to choose $\nu_{\max}$ such that $1/\nu_{\max}$ reflects the fastest time scale in the input signals. Accordingly, the symmetry of the learning window is governed by the relation between the length of the EPSP and the fastest time scale in the input data. If the EPSP is short enough to resolve the fastest input components, the learning window is symmetric. If the EPSP is too long to fully resolve the temporal structure of the input, i.e. it acts as a low-pass filter, the learning window will tend to be antisymmetric.

We choose a value of $\nu_{\max} = 1/(40\text{ms})$. The argument for this choice is that within a rate code, the cells that project to the neuron under consideration can hardly convey signals that vary on a faster time scale than the duration of their EPSP. It is thus reasonable to choose the time constant of the EPSP and the inverse of the cutoff frequency to have the same order of magnitude. Typical durations of cortical EPSPs are of the order of tens of milliseconds (see Koch et al. (1996) for further references and a critical discussion), so 40ms is a reasonable value.

Figure 3 illustrates the connection between $\mathcal{P}_{f_{\mathrm{SFA}}}$, $W_0$, the learning window and the EPSP. It also shows the learning windows for three different durations of the EPSP, while keeping $\nu_{\max} = 1/(40\text{ms})$. The oscillatory and slowly decaying tails of $W(t)$ are due to the sharp cutoff of the power spectrum $\mathcal{P}_{f_{\mathrm{SFA}}}$ at $|\nu| = \nu_{\max}$ and become less pronounced, if $\mathcal{P}_{f_{\mathrm{SFA}}}$ is smoothened out.

As negative time arguments in $W(t)$ correspond to the case, in which the presynaptic spike (and thus the onset of the resulting EPSP) precedes the postsynaptic spike, the shape of the theoretically derived learning window for physiologically plausible values of $\tau$ and $\nu_{\max}$ ($\tau = 1/\nu_{\max} = 40$ms, middle row in figure 3) predicts potentiation of the synapse when a postsynaptic spike is preceded by the onset of an EPSP and depression of the synapse when this temporal order in reversed. This behavior is in agreement with experimental data from neocortex and hippocampus in rats as well as from the optic tectum in Xenopus (Debanne et al., 1994; Bi and Poo, 1998; Feldman, 2000; Markram et al., 1997; Zhang et al., 1998). To further illustrate this agreement, Figure 4 compares the data as published by Bi and Poo (1998) with the learning window resulting from a smoothened power spectrum with the shape of a Cauchy function (35) instead of $\mathcal{P}_{f_{\mathrm{SFA}}}$. As demonstrated above, this corresponds to implementing the slowness principle in form of the 'trace rule'. Interestingly, the resulting learning window has the double-exponential shape that is regularly used in models of STDP (e.g. van Rossum et al., 2000; Song and Abbott, 2001; Gütig et al., 2003). As the absolute scale of the learning window is not determined in our analysis, it was adjusted to facilitate the comparison with the experimental data.

## 3.4 Interpretation of the learning windows

The last section leaves a central question open: why are these learning windows optimal for slowness learning and why does the EPSP play such an important role for the shape of the learning window?

Let us first discuss the case of the symmetric learning window, i.e. the situation in which the EPSP is shorter than the fastest time scale in the input signal. Then the convolution with the

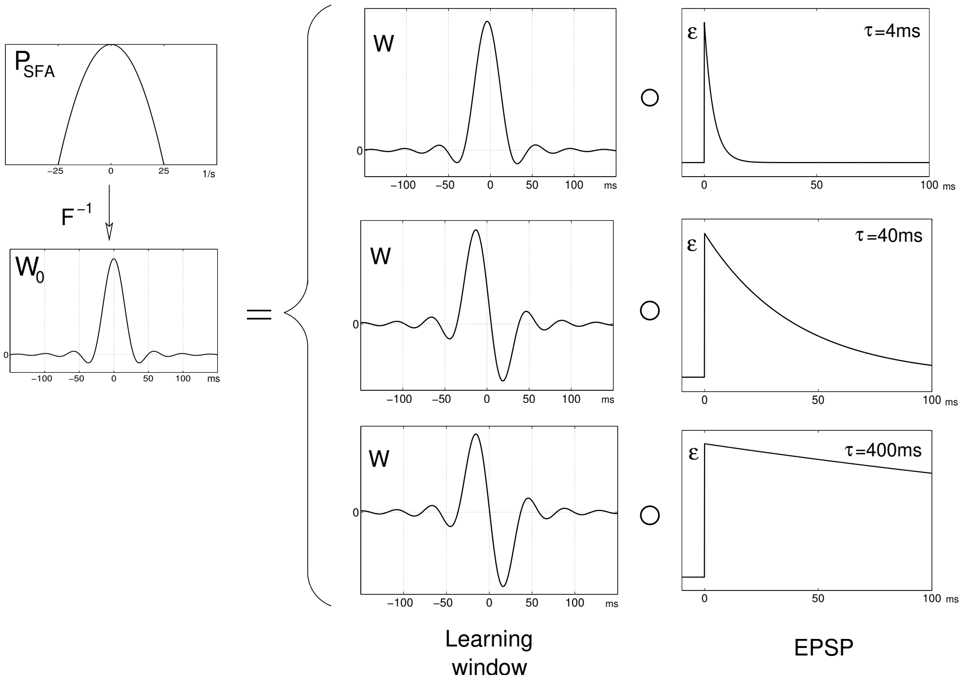


**Figure 3: Relation between the EPSP and the learning window**. The power spectrum $\mathcal{P}_{f_{\rm SFA}}$ is the Fourier transform of $W_0$, which in turn is the convolution of the learning window $W$ and the EPSP $\epsilon$. The figure shows the learning windows required for SFA for three different EPSP durations ($\tau = 4, 40, 400$ms). The maximal input frequency $\nu_{\max}$ was 1/(40ms) in all plots.

EPSP has practically no effect on the temporal structure of the signal and the output firing rate can be regarded as an instantaneous function of the input rates. We can thus neglect the EPSP altogether. The learning mechanism can then be understood as follows: Assume at a given time $t$ the postsynaptic firing rate $r^{\rm out}$ is high and causes a postsynaptic spike. Then the finite width of the learning window leads to potentiation not only of those synapses that participated in initiating the spike, but also of those which transmit a spike within a certain time window around the time of the postsynaptic spike. As this leads to an increase of the firing rate within this time window, the learning mechanism tends to equilibrate the firing rates for neighboring times and thus favors temporally slow output signals.

If the duration of the EPSP is longer than the fastest time scale in the input signal, the output firing rate is no longer an instantaneous function of the input signals, but generated by lowpass filtering the signal $a^{\rm out}$ with the EPSP. This is crucial for learning, because the objective of the continuous model is to optimize the slowness of $a^{\rm out}$, whose temporal structure is now "obscured" by the EPSP. In order to optimize the objective, the system thus has to develop a deconvolution mechanism to reconstruct $a^{\rm out}$. From this point of view, the learning window has to perform two tasks simultaneously. It has to first perform the deconvolution and then enforce slowness on the resulting signal. This is most easily illustrated by means of condition (55). The convolution of the learning window with the EPSP generates a function $W_0(t)$ that is independent of the EPSP and which coincides with the learning window for infinitely short EPSPs. Intuitively, we could solve this equation by choosing a learning window that consists of the "inverse" of the EPSP and the EPSP-free learning window $W_0$. An intuitive example is the limiting case of an infinitely long EPSP. The EPSP then corresponds to a Heaviside function and performs an integration, which can be inverted

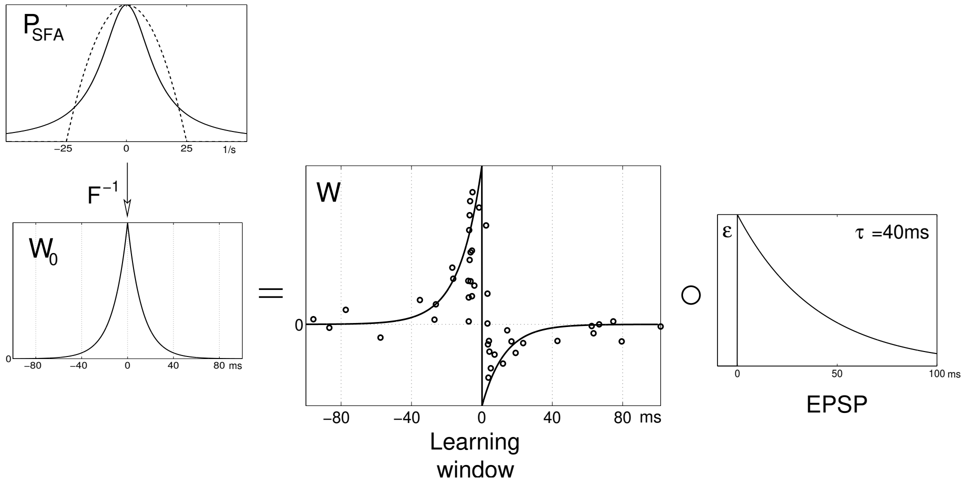


**Figure 4: Comparison of the learning window with experimental data**. The plot compares the theoretically predicted learning window with experimental data from hippocampal pyramidal cells as published by Bi and Poo (1998) (larger plot, middle). Instead of the ideal power spectrum $\mathcal{P}_{f_{\mathrm{SFA}}}$ with the abrupt cutoff at $\nu_{\max}$ as stated in equation (20), a Cauchy function with $\gamma$=1/(15ms) was used (top left, the dashed line is $\mathcal{P}_{f_{\mathrm{SFA}}}$ for $\nu_{\max}$=1/(40ms)). Again, the EPSP decay time was $\tau$ = 40ms. This learning window corresponds to an implementation of the 'trace rule' (Földiak, 1991; O'Reilly and Johnson, 1994; Wallis and Rolls, 1997) for a decay time of the exponential filter of 15ms.

by taking the derivative. Thus the learning window for long EPSPs is the temporal derivative of the learning window for short EPSPs. The dependence of the required learning window on the shape of the EPSP is thus caused by the need of the learning window to "invert" the EPSP.

These considerations shed a different light on the shape of physiologically measured learning windows. The antisymmetry of the learning window may not act as a physiological implementation of a causality detector after all but rather as a mechanism for compensating intrinsic lowpass filters in neuronal processing such as the EPSP. For functional interpretations of STDP, it may be more sensible to consider the convolution of the learning window with the EPSP than the learning window alone.

It should be noted that, according to our learning rule, the weights adapt in order to make a hypothetical instantaneous output signal $a^{\mathrm{out}}$ optimally slow. This does not necessarily imply that the output firing rate $r^{\mathrm{out}}$, which is generated by lowpass filtering $a^{\mathrm{out}}$ with the EPSP, is optimally slow. In principle, the system could generate more slowly varying signals by exploiting the temporal structure of the EPSP. However, the motivation for the slowness principle is the idea that the system learns to detect invariances in the *input* signal and that from this perspective the goal of creating a slowly varying output signal is not an end in itself, but a means to learn invariances.

# 4 Discussion

Neurons in the central nervous system display a wide range of invariances in their response behavior, examples of which are phase invariance in complex cells in the early visual system (Hubel and Wiesel, 1968), head direction invariance in hippocampal place cells (Muller et al., 1994), or more complex invariances in neurons associated with face recognition (Quiroga et al., 2005). If these invariances are learned, the associated learning rule must somehow reflect a heuristics as to which sensory stimuli are supposed to be categorized as being the same. Objects in our environment are unlikely to change completely from one moment to the next but rather undergo typical transformations. Intuitively, responses of neurons with invariances to these transformations should thus vary more slowly than others. The slowness principle uses this intuition and conjectures that neurons learn these invariances by enforcing their output signals to vary slowly without exploiting lowpass filtering.

Slow Feature Analysis (SFA,Wiskott and Sejnowski, 2002) is one implementation of the slowness principle in that it minimizes the mean square of the temporal derivative of the output signal for a given set of training data. SFA has been shown to successfully model a wide range of physiologically observed properties of complex cells in primary visual cortex (Berkes and Wiskott, 2005) as well as translation-, rotation-, and other invariances in the visual system (Wiskott and Sejnowski, 2002). In combination with a sparse coding objective, SFA has also been used to describe the self-organized formation of place cells in the hippocampal formation (Franzius et al., 2006).

As an algorithm SFA is highly efficient, but its formulation is rather technical and it has not yet been examined if it is feasible to implement SFA within the limitations of neuronal circuitry. In this paper, we approach this question analytically and demonstrate that such an implementation is possible in both continuous and spiking model neurons.

In the first part of the paper, we show that for linear continuous model neurons, the slowest direction in the input signal can be learned by means of Hebbian learning on lowpass filtered versions of the input and the output signal. The power spectrum of the lowpass filter required for implementing SFA can be derived from the learning objective and has the shape of an upside-down parabola.

The idea of using lowpass filtered signals for invariance learning is a feature that our model has in common with several others (Földiak, 1991; O'Reilly and Johnson, 1994; Wallis and Rolls, 1997). Section 2.5 discusses the relation of our model to these 'trace rules' and shows that they bear strong similarities.

The second part of the paper discusses the modifications that have to be made to adjust the learning rule for a Poisson neuron. We find that in an ensemble-averaged sense it is possible to reproduce the behavior of the continuous model neuron by means of spike-timing-dependent plasticity (STDP). Our study suggests that the outcome of STDP learning is not governed by the learning window alone but rather by the convolution of the learning window with the EPSP, which is of relevance for functional interpretations of STDP.

The learning window that realizes SFA can be calculated analytically. Its shape is determined by the interplay of the duration of the EPSP and the maximal input frequency $\nu_{\max}$, above which the input signals are assumed to have negligible power. If $\nu_{\max}$ is small, i.e. if the EPSP is sufficiently short to temporally resolve the most quickly varying components of the input data, the learning window is symmetric whereas for large $\nu_{\max}$ or long EPSPs, it is antisymmetric. Interestingly, physiologically plausible parameters lead to a learning window whose shape and width is in agreement with experimental findings. Based on this result we propose a new functional interpretation of the STDP learning window as an implementation of the slowness principle that compensates for neuronal lowpass filters such as the EPSP.

A different approach to unsupervised learning of invariances with a biologically realistic model neuron has been taken by Körding and König (2001). In their model, bursts of backpropagating spikes gate synaptic plasticity by providing sufficient amounts of dendritic depolarization. These bursts are assumed to be triggered by lateral connections that evoke calcium spikes in the apical dendrites of cortical pyramidal cells.

Of course the model presented here is not a complete implementation of SFA. We have only considered the central step of SFA, the extraction of the most slowly varying direction from a set of whitened input signals. To implement the full algorithm, additional steps are necessary: a nonlinear expansion of the input space, the whitening of the expanded input signals and a means of normalizing the weights. When traversing the dendritic arborizations of a postsynaptic neuron, axons often make more than one synaptic contact. As different input channels may be subjected to different nonlinearities in the dendritic tree (cf. London and Häusser, 2005) the postsynaptic neuron may have access to several nonlinearly transformed versions of the same presynaptic signals. Conceptually, this resembles a nonlinear expansion of the input signals. However, it is not obvious, how these signals could be whitened within the dendrite. On the network level, however, whitening could be achieved by adaptive recurrent inhibition between the neurons (Földiak, 1989). This mechanism may also be suitable for extracting several slow uncorrelated signals as required in the original formulation of SFA (Wiskott and Sejnowski, 2002) instead of just one. We assumed an explicit weight normalization in the description of our model. However, one could also use a modified learning rule that implicitly normalizes the weight vector as long as it extracts the signal with the largest variance. A possible biological mechanism is synaptic scaling (Turrigiano and Nelson, 2000), which is believed to multiplicatively rescale all synaptic weights according to the postsynaptic activity, similar to Oja's rule (Oja, 1982; Abbott and Nelson, 2000). Thus, it appears that most of the mechanisms necessary for an implementation of the full SFA algorithm are available but that it is not clear how to combine them in a biologically plausible way.

Another critical point in the analytical derivation for the spiking model is the replacement of the temporal by the ensemble average, as this allows to recover the rates that underlie the Poisson processes. The validity of the analytical results thus requires some kind of ergodicity in the training data, a condition, which of course needs to be justified for the specific input data at hand.

It is still open if the results presented here can be reproduced with more realistic model neurons. The spiking model neuron used here was highly simplified in that it had a linear relationship between input and output firing rate. In many real neurons highly nonlinear behavior was observed. We have also neglected nonlinearities in the learning rule such as the frequency- and the weight-dependence of STDP (Bi and Poo, 1998; Sjöström et al., 2001). Furthermore, modeling the spiking mechanism of a neuron by an inhomogeneous Poisson process is also a severe simplification that ignores basic phenomena of spike generation in biological neurons like refractoriness and thresholding. It is not clear how these characteristics would change the learning rule that leads to an implementation of the slowness principle. It seems to be a very difficult task to answer these questions analytically. Simulations will be necessary to verify the results derived here and to analyze which changes appear and which adaptations must be made in a more realistic model of neural information processing.

In summary, the analytical considerations presented here show that (i) slowness can be equivalently achieved by minimizing the variance of the time derivative signal or by maximizing the variance of the lowpass filtered signal, the latter of which can be achieved by standard Hebbian learning on the lowpass filtered input and output signals; (ii) the difference between SFA and the trace learning rule lies in the exact shape of the effective lowpass filter, for most practical purposes the results are probably equivalent; (iii) for a spiking Poisson model neuron with an STDP learning rule it is not the learning window that governs the weight dynamics but the convolution of the learning window with the EPSP; (iv) the STDP learning window that implements the slowness objective is in good agreement with learning windows found experimentally. With these results we have reduced the gap between slowness as an abstract learning principle and biologically plausiple STDP learning rules and we offer a completely new interpretation of the standard STDP learning window.

# 5 Funding

This work was generously supported by the Volkswagen Foundation.

# References

Abbott, L. F. and Nelson, S. B. (2000). Synaptic plasticity: Taming the beast. *Nature Neuroscience*, 3:1178–1183.

Becker, S. and Hinton, G. E. (1992). A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163.

Berkes, P. and Wiskott, L. (2005). Slow feature analysis yields a rich repertoire of complex cells. *Journal of Vision*, 5(6):579–602.

Bi, G.-q. and Poo, M.-m. (1998). Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.*, 18(24):10464–10472.

Debanne, D., Gähwiler, B. H., and Thomson, S. M. (1994). Asynchronous pre- and postsynaptic activity induces associative long-term depression in area CA1 of the rat hippocampus. *PNAS*, 91:1148–1152.

Feldman, D. E. (2000). Timing-based LTP and LTD at vertical input to layer II/III pyramidal cells in rat barrel cortex. *Neuron*, 27:45–56.

Földiak, P. (1989). Adaptive network for optimal linear feature extraction. In *Proceedings of the IEEE/INNS International Joint Conference on Neural Networks*, pages 401–405, New York. IEEE Press.

Földiak, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, 3:194–200.

Franzius, M., Sprekeler, H., and Wiskott, L. (2006). Slowness leads to place cells. In *Proceedings CNS 2006*.

Gütig, R., Aharonov, S., Rotter, S., and Sompolinsky, H. (2003). Learning input correlations through nonlinear temporally asymmetric Hebbian plasticity. *Journal of Neuroscience*, 23(9):3697–3714.

Hubel, D. and Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, 195:215–243.

Kempter, R., Gerstner, W., and van Hemmen, J. L. (1999). Hebbian learning and spiking neurons. *Physical Review E*, 59:4498–4514.

Kistler, W. M. and van Hemmen, J. L. (2000). Modeling synaptic plasticity in conjunction with the timing of pre- and postsynaptic action potentials. *Neural Computation*, 12:385.

Koch, C., Rapp, M., and Segev, I. (1996). A brief history of time (constants). *Cerebral Cortex*, 6:92–101.

Körding, K. P. and König, P. (2001). Neurons with two sites of synaptic integration learn invariant representations. *Neural Computation*, 13(12):2823–2849.

London, M. and Häusser, M. (2005). Dendritic computation. *Annu. Rev. Neurosci.*, 28:503–532.

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

Mitchison, G. (1991). Removing time variation with the anti-Hebbian differential synapse. *Neural Computation*, 3:312–320.

Muller, R., Bostock, E., Taube, J. S., and Kubie, J. L. (1994). On the directional firing properties of hippocampal place cells. *Journal of Neuroscience*, 14(12):7235–7251.

Oja, E. (1982). A simplified neuron as a principal component analyzer. *Journal of Mathematical Biology*, 15:267–273.

O'Reilly, R. C. and Johnson, M. H. (1994). Object recognition and sensitive periods: A computational analysis of visual imprinting. *Neural Computation*, 6(3):357–389.

Peng, H. C., Sha, L. F., Gan, Q., and Wei, Y. (1998). Energy function for learning invariance in multilayer perceptron. *Electronics Letters*, 34(3):292–294.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, 435:1102–1107.

Rubin, J., Lee, D. D., and Sompolinsky, H. (2001). Equilibrium properties of temporally asymmetric hebbian learning. *Phys. Rev. Lett.*, 86(2):364–367.

Sjöström, P. J., Turrigiano, G. G., and Nelson, S. B. (2001). Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. *Neuron*, 32(6):1149–1164.

Song, S. and Abbott, L. F. (2001). Cortical mapping and development through spike timing-dependent plasticity. *Neuron*, 32:339–350.

Stone, J. V. and Bray, A. (1995). A learning rule for extracting spatio-temporal invariances. *Network: Computation in Neural Systems*, 6:429–436.

Turrigiano, G. G. and Nelson, S. B. (2000). Hebb and homeostasis in neuronal plasticity. *Current Opinion in Neurobiology*, 10:358–364.

van Rossum, M. C. W., Bi, G. Q., and Turrigiano, G. G. (2000). Stable hebbian learning from spike timing-dependent plasticity. *Journal of Neuroscience*, 20(23):8812–8821.

Wallis, G. and Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology*, 51(2):167–194.

Wiskott, L. and Sejnowski, T. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770.

Zhang, L. I., Tao, H. W., Holt, C. E., Harris, W. A., and Poo, M.-m. (1998). A critical window for cooperation and competition among developing retinotectal synapses. *Nature*, 395(6697):37–44.