Diplomarbeit

Extracting Statistically Independent Components with a Generalized BCM Rule for Spiking Neurons

Stefan Klampfl

Institut für Grundlagen der Informationsverarbeitung
Technische Universität Graz
Vorstand: O. Univ.-Prof. Dipl.-Ing. Dr. rer.nat. Wolfgang Maass

Betreuer und Begutachter:
O. Univ.-Prof. Dipl.-Ing. Dr. rer.nat. Wolfgang Maass

Graz, im April 2006

Abstract

In this thesis an unsupervised learning rule is derived for spiking neurons that extracts statistically independent components from an ensemble of input spike trains. For that the approach of a recent result showing that maximizing information transmission for a single neuron yields a generalized Bienenstock-Cooper-Munro (BCM) rule for spiking neurons is extended in a way that a second neuron that receives the same input additionally minimizes the mutual information between its output and the output of the other neuron. The resulting synaptic plasticity rule gives an additional term which depends on the recent firing history of both neurons and which is sensitive to the momentary statistical dependence between the outputs. The learning rule is tested in a number of computer simulation experiments and found to be able to detect different correlation or rate modulation groups among the input. Finally, it is suggested how the rule can be made biologically more realistic by using (inhibitory) interneurons to make the information about the firing behavior of one neuron available at the site of the other neuron. This result can be viewed as a first step toward a (nonlinear) independent component analysis (ICA) method for spiking neurons.

Keywords: computational intelligence, unsupervised learning, neural networks, computational neuroscience, synaptic plasticity, BCM rule, information theory, independent component analysis

Kurzfassung

In dieser Diplomarbeit wird eine unüberwachte Lernregel für spikende Neuronen hergeleitet, die statistisch unabhängige Komponenten aus einer Menge von Input Spike Trains extrahiert. Dabei wurde der Ansatz eines kürzlichen Ergebnisses, das zeigt, dass die Maximierung der Informationsübertragung eines einzelnen Neurons eine allgemeine Bienenstock-Cooper-Munro (BCM) Regel für spikende Neuronen liefert, erweitert, sodass ein zweites Neuron, das denselben Input erhält, zusätzlich die Transinformation zwischen seinem Output und dem des anderen Neurons minimiert. Die resultierende Lernregel enthält einen zusätzlichen Term, der von der jüngsten Feuergeschichte beider Neurone abhängt und die momentane statistische Abhängigkeit zwischen den Outputs misst. Die Lernregel wird in einer Reihe von computersimulierten Experimenten getestet und ist etwa in der Lage, Gruppen verschiedener Korrelationen und verschieden modulierter Feuerraten innerhalb der Inputs zu erkennen. Abschließend wird noch vorgeschlagen, wie diese Lernregel biologisch realistischer gemacht werden könnte, indem man (inhibitorische) Interneuronen verwendet, die die Information über das Feuerverhalten eines Neurons für das andere verfügbar machen. Dieses Resultat kann als erster Schritt in die Richtung einer Methode für (nichtlineare) Independent Component Analysis (ICA) für spikende Neuronen angesehen werden.

Stichwörter: Maschinelle Intelligenz, Unüberwachtes Lernen, Neuronale Netze, Neuroinformatik, Synaptische Plastizität, BCM-Regel, Informationstheorie, Independent Component Analysis

I hereby certify that the work presented in this thesis is my own and that work performed by others is appropriately cited.

Ich versichere hiermit, diese Arbeit selbständig verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und mich auch sonst keiner unerlaubten Hilfsmittel bedient zu haben.

Acknowledgements

First, I would like to thank my advisors Prof. Wolfgang Maass and Robert Legenstein for their guidance and for giving me fruitful ideas and suggestions that have made this work possible. Furthermore, I wish to thank all other members of the Institute for Theoretical Computer Science (IGI) for including me in their group and their friendly (scientific and non-scientific) discussions.

Additionally, I'd like to thank Wulfram Gerstner from EPFL for his inspiring preceding work and for helping me on a particular problem.

Last but not least, I want to express my gratitude to my family and friends for all of their support.

Graz, April 2006 Stefan Klampfl

Contents

1. Introduction

Unsupervised learning is a very important concept in the theory of neural networks. In contrast to supervised learning, where each input pattern is paired with a target value and where the network's task is usually to infer a function from the data that predicts these labels for unseen patterns, the training data does not contain any target values at all and consists merely of the attribute vectors. This learning paradigm is commonly applied, for instance, to clustering or to reducing the dimensionality of data, as well as to learning associations between input patterns; in short, we often want to discover some underlying structure of the data (Duda et al., 2000; Bishop, 1995). This ability is also essential for biological neural networks in the brain, which are known to develop representations based on the statistical structure of the input. Since there is usually no teacher signal available that would allow supervised learning to take place in most parts of the brain, unsupervised learning plays an important role in the field of neuroscience dealing with plasticity and learning. A major area of research concerns the development of neuronal selectivity and the formation of cortical maps, for example (Dayan and Abbott, 2001; Gerstner and Kistler, 2002).

Most of the theory of unsupervised learning in neuroscience is based on the work of Hebb (Hebb, 1949) and his principle that the synaptic weight change is driven by correlated activity between pre- and postsynaptic neurons. This postulate has given rise to a large family of Hebbian learning rules, most prominently Oja's rule (Oja, 1982), that have been found to perform *principle component analysis* (PCA). This term refers to the process of finding the set of orthogonal directions that minimizes the error of the reconstructed data when projecting the input data onto these principal components. PCA is therefore often used for dimensionality reduction, e.g., for image compression.

While PCA has the nice property that the principle components of random variables are uncorrelated, it is often even more desirable to extract statistically independent components from some input data. This is a stronger condition because while independent random variables are also uncorrelated, uncorrelatedness on the other hand does not imply independence. In contrast to PCA, *independent component analysis* (ICA) does not seek a set of orthogonal components, but a set of independent components (Hyvärinen and Oja, 2000; Hyvärinen et al., 2001). It is closely related to blind source separation where the problem is to find the set of original sources from an observed mixture. By providing the independent components of the input ICA produces an efficient (sparse) coding scheme, therefore it is also a hot candidate for the theory of development of neural coding, where one of the main issues is the reduction of redundancy (Barlow, 1961; Barlow, 1989). In fact, it has been found that many cells in visual cortex develop receptive fields that can be reproduced by ICA (Hyvärinen et al., 2005).

However, learning rules for unsupervised learning of independent components have yet

been proposed mainly for pure rate models; a more general synaptic plasticity mechanism that extracts statistically independent components from a spiking network is still missing. In this thesis a learning rule is derived that tries to keep the output of two neurons that receive the same input at their synapses statistically independent by minimizing the mutual information between the output spike trains. This work is based on a recent approach that maximizes the information transmission of a spiking neuron (Toyoizumi et al., 2005a) where the authors find that maximizing the mutual information between an ensemble of input spike trains and the output spike train of a neuron yields a synaptic plasticity rule that exhibits the same features as the classical BCM rule. The Bienenstock-Cooper-Munro (BCM) model, originally developed as a pure rate model in the context of development of stimulus selectivity in the visual cortex (Bienenstock et al., 1982), has been one of the most influential concepts emerging from the spirit of Hebb's principle. It predicts regimes of both LTP and LTD depending on the postsynaptic activity, which are separated by a sliding threshold that is necessary for stability. By constructing a bridge between the BCM model and the concept of optimality in terms of information transmission by spike trains the classical BCM rule has been generalized to the case of spiking neurons with refractoriness (Toyoizumi et al., 2005a).

In this work this approach is extended in a way that a second neuron that receives the same input also maximizes its information transmission between input and output spike trains, but at the same time tries to minimize the mutual information between its output and the output of the other neuron. The resulting synaptic update rule is similar to the generalized BCM rule proposed in (Toyoizumi et al., 2005a); an additional term is included in the learning equation of neuron 2 that is sensitive to the momentary statistical dependence and that depends on the recent postsynaptic history of both neurons. However, this would require information about the firing behavior of neuron 1 to be non-locally available at the site of neuron 2, therefore also an attempt is made to provide this information via different synaptic connections. The proposed learning rule is tested in several computer simulation experiments.

This thesis is organized as follows. Chapter 2 gives a short overview of synaptic plasticity. First, Hebb's principle is introduced and some issues and drawbacks concerning Hebbian learning in general are discussed. Then several synaptic learning rules are presented that implement Hebb's idea in different ways, and finally it is explained how the BCM rule, as a special form of Hebbian plasticity, achieves stability and competition between synapses. Chapter 3 is then dedicated to information theory. First, it introduces information theoretic quantities such as entropy and mutual information, by which the information transmission of neurons can be quantified. Then the synaptic plasticity rule of (Toyoizumi et al., 2005a) is presented as a way how information transmission of a neuron can be maximized. Finally, it is shown how this learning rule relates to the BCM model. In chapter 4 the main results of this thesis are presented; a learning rule is derived that extends the one of the previous chapter in a way that a second neuron receiving the same input keeps its output statistically independent to the output of the other neuron. This is then verified in a number of computer simulation experiments, and it is also indicated how this approach might be extended to the case of more than two postsynaptic neurons. Chapter 5 deals with the question how this learning rule might

be made biologically more realistic by implementing the non-local term concerning the statistical dependence between the outputs of both neurons via synaptic connections. More precisely, a mechanism is proposed that modulates the gain of the second neuron according to the activity of both neurons, while the weights evolve according to the basic generalized BCM rule. Finally, chapter 6 concludes and gives further remarks.

2. Synaptic Plasticity

In biological neural networks, each synapse is characterized by a single parameter (often called the synaptic efficacy, or simply the "weight" of the synapse) that determines the amplitude of the postsynaptic response to an incoming action potential. However, the efficacy of synaptic connections between neurons in the brain is not fixed, but it varies depending on different factors, such as the pre- and postsynaptic firing frequencies or spike timings. The term *synaptic plasticity* refers to the variability of the strength of a synapse, i.e., the ease with which an action potential in one cell excites (or inhibits) a target cell. Persistent changes in the synaptic strength that last for a time of tens of minutes or longer up to days, months and years are called long-term potentiation (LTP) and long-term depression (LTD), depending on whether the weight has been increased or decreased[1]. Originally it has been proposed by Ramón y Cajal in 1894 that "memories might be formed by strengthening the connections between existing neurons to improve the effectiveness of their communication" (Ramón y Cajal, 1911; Squire and Kandel, 1999). Today it is widely believed that activity-dependent synaptic plasticity is the basic phenomenon underlying learning and memory, and it is also thought to play a crucial role in the development of neural circuits.

During the last decades, a large number of theoretical concepts and mathematical models have emerged that have helped to understand the functional consequences of synaptic modifications (Dayan and Abbott, 2001; Gerstner and Kistler, 2002; Cooper et al., 2004). In the formal theory of neural networks the weight w_{ij} of a synapse connecting neuron j to neuron i is considered as a parameter that can be adjusted in order to optimize the performance of a network for a given task. The process of parameter adaptation is called *learning* and the procedure for adjusting the weights is usually referred to as a *learning rule*. Experimentally inspired synaptic plasticity rules have been applied to a wide variety of tasks including pattern recognition and function approximation. One simple set of learning rules consider synaptic changes that are driven by correlated activity between pre- and postsynaptic neurons. This class of learning rules can be motivated by Hebb's principle (Hebb, 1949) and is therefore often called *Hebbian learning* (Fregnac, 2002; Erdi and Somogyvari, 2002; Brown and Chattarji, 1998). Apart from the work of Hebb, one of the most influential concepts has been the Bienenstock-Cooper-Munro (BCM) model originally developed to account for cortical organization and receptive field properties during development (Bienenstock et al., 1982).

In this chapter Hebb's original conjecture and the problems usually associated with

[1] More detailed experimental data suggest that synapses exhibit dynamics on a much shorter timescale than LTP and LTD, which also might be important for information processing in the brain. This gives rise to more complex dynamic synapse models (Markram et al., 1998; Maass and Markram, 2002) that depend on more parameters than just a single weight.

Hebbian learning rules such as competition and stability are introduced. Some synaptic plasticity rules emerging from Hebb's principle are then briefly discussed in section 2.2 and how they (or why they do not) overcome these difficulties. Section 2.3 is finally dedicated to the BCM rule, which augments standard Hebbian plasticity by the mechanism of a sliding threshold on the postsynaptic activity, which has some interesting properties concerning the stability of synaptic weights.

2.1. Hebb's Postulate

In 1949, Donald Hebb conjectured that a synapse is strengthened if both the pre- and postsynaptic neuron are simultaneously active. His postulate describes how the connection from presynaptic neuron A to a postsynaptic neuron B should be modified:

> "When an axon of cell A is near enough to excite cell B or repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased." (Hebb, 1949)

Hebb's original suggestion only concerned increases in synaptic strength, but it has been generalized to include decreases in strength as well, arising from repeated failure of neuron A to be involved in the activation of neuron B. That is, if one neuron is stimulating some other neuron and at the same time that receiving neuron is also firing the strength of the connection between the two neurons will be increased, and vice versa – if one neuron is active and the other one is not the connection strength is decreased. Today this famous statement is often rephrased in the sense that modifications in the synaptic strength are driven by correlations in the firing activity of pre- and postsynaptic neurons.

2.1.1. Locality, Stability, and Competition

Two aspects in Hebb's postulate are particularly important, namely *locality* and *cooperativity*. Locality means that the change in the synaptic weight can only depend on variables that are locally available at the synapse. These include pre- and postsynaptic firing rates and spike timings and also the current value of the synaptic efficacy, but not for instance the activity of other neurons or the weights of different synapses. The term cooperativity refers to the fact that both pre- and postsynaptic neuron have to be simultaneously active for a synaptic weight change to occur.

One problem with Hebb's rule is that the activity that increases the synaptic strength is reinforced by Hebbian plasticity, which leads to an even higher activity and further modification. Without appropriate adjustments or imposing certain constraints, this positive feedback process would produce an uncontrolled growth of weights. The easiest way to control synaptic strengthening is to impose an upper limit on the value of the synaptic efficacies, which is supported by experiments. It also makes sense to prevent weights from changing sign since excitatory synapses cannot change into inhibitory

synapse or vice versa. Thus, each weight w_{ij} is allowed to change only between 0 and a maximum value w_{max}, which is a constant.

Another problem associated with Hebbian modification is that synapses are modified independently. Competition between synapses is essential for any form of self-organization or pattern formation. For example, if all of the synaptic weights of a neuron are driven to their maximum value w_{max} the neuron completely loses its selectivity to different input patterns. Therefore, usually a regulation or competition mechanism is required so that some synapses are forced to weaken when others become strong. Thus, the basic Hebb rule often has to be augmented with terms that ensure stability of weights and competition between synapses (Abbott and Nelson, 2000; Chechik et al., 2002).

2.1.2. Synaptic Normalization

Competition between synapses can be introduced by imposing a global constraint or some regulation mechanism on the weights of all synapses of a neuron, e.g., by normalizing the sum of the squares of all weights (i.e, the norm of the weight vector) to a fixed value. Usually, it is distinguished between additive (subtractive) normalization, where the weight change is independent of the current value of the synaptic efficacy, and multiplicative normalization, where the amount of modification is proportional to the weight value. However, this often requires global information about the values of all efficacies to be available at each synapse and therefore violates the locality of synaptic plasticity, which is usually assumed. On the other hand, normalization can also be achieved with purely local learning rules, e.g., the multiplicative Oja rule (Oja, 1982).

2.1.3. Firing Rate vs. Spike Timing

In the original formulation of Hebbian learning the activity of each neuron is described by a single continuous variable, rather than a specific spike train. This paradigm is also used in the theory of artificial neural networks, where each unit maps real-valued inputs to a real-valued output. Normally, this variable represents the firing rate of the neuron; in this case it is restricted to nonnegative values. In order to allow negative values as well one could interpret the activity variable, for instance, as the difference between a firing rate and a fixed background rate. In these rate-based models the precise timings of individual spikes are considered unimportant.

However, recent experiments have shown that the amplitude and even the direction of weight changes critically depend on the relative timing of pre- and postsynaptic spikes. The temporal requirements for two neurons to be active together can then be formulated in a way that the change in the synaptic efficacy depends on the time differences between the spike times of the pre- and postsynaptic neuron ("spike-time dependent synaptic plasticity", or STDP, see section 2.2.3) on the time scale of milliseconds. More precisely, the synapse is strengthened if the presynaptic spike occurs shortly before the postsynaptic neuron fires, but it is weakened if the sequence of spikes is reversed. This observation is indeed in agreement with Hebb's postulate because presynaptic neurons that are active slightly before the postsynaptic neuron are those which "take part in

firing it" whereas those that fire later obviously did not contribute to the postsynaptic action potential (Gerstner and Kistler, 2002).

There has been an ongoing debate whether cortical neurons transmit information primarily in their average firing rates or the precise timing of their spikes. Recent results suggest that cortical plasticity jointly depend on the rate and relative timing of pre- and postsynaptic firing (Sjøstrøm et al., 2001; Nelson et al., 2002).

2.2. Hebbian Plasticity Rules

Rules for synaptic plasticity usually take the form of differential equations describing the temporal change of synaptic weights as a function of the pre- and postsynaptic activity and possibly other factors. Local forms of Hebbian rate-based models can therefore be written as (Gerstner and Kistler, 2002)

$$\frac{dw_{ij}}{dt} = F(w_{ij}, v_i, v_j), \tag{2.1}$$

where dw_{ij}/dt is the rate of change of the synaptic weight and F is some function of the pre- and postsynaptic activities v_i and v_j and of the current weight value w_{ij}. The dependence on the value of the synaptic efficacy is a natural consequence of the fact that w_{ij} is bounded, otherwise it could grow without limit.

2.2.1. Basic Hebb Rule

The simplest plasticity rule that follows from the spirit of Hebb's conjecture modifies the weight w_{ij} of a synapse connecting neuron j to neuron i by an amount proportional to the product of pre- and postsynaptic activities, i.e.,

$$\frac{dw_{ij}}{dt} = \alpha v_i u_j, \tag{2.2}$$

where α is a positive constant called the *learning rate*, which controls the speed of weight adaption[2]. It can be shown (Dayan and Abbott, 2001) that this basic Hebb rule is unstable, no matter if the activity variables are allowed to take on negative values or not, because the change of the length of the weight vector is always positive. In order to avoid unbounded growth one must impose an upper saturation constraint, but in a model where synaptic weights can only be strengthened all efficacies will finally saturate at their maximum level. The ability to induce LTD is therefore a necessary requirement for any useful learning rule.

One could, for example, introduce weight decay by adding a corresponding term to equation (2.2),

$$\frac{dw_{ij}}{dt} = \alpha v_i u_j - \gamma w_{ij}, \tag{2.3}$$

[2]Learning rules where the weight change is proportional to the negative product of pre- and postsynaptic activities, i.e., $\frac{dw_{ij}}{dt} = -\alpha v_i u_j$, are usually called *anti-Hebbian*.

with some paramteter $\gamma > 0$ such that the synaptic strength w_{ij} decays back to 0 in the absence of any activity. However, even with a weight decay term, the basic Hebb rule fails to induce competition between different synapses since all synapses are still modified independently.

2.2.2. Covariance Rules

Experiments suggest that synapses can depress also if presynaptic activity is accompanied by a low level of postsynaptic activity (and vice versa). This gives rise to a family of covariance or "gating" rules that compare the pre- or postsynaptic activity with a certain threshold. For example, one can define a learning rule

$$\frac{dw_{ij}}{dt} = \alpha(v_i - \theta_v)u_j, \tag{2.4}$$

where θ_v is a threshold that determines the level of postsynaptic activity above which LTD switches to LTP. We say that the weight changes are "gated" by the postsynaptic neuron. Alternatively, a similar threshold θ_u can be imposed on the presynaptic activity, i.e.,

$$\frac{dw_{ij}}{dt} = \alpha v_i(u_j - \theta_u). \tag{2.5}$$

The learning rule in equation (2.4) is also often called *postsynaptic gating*, that of equation (2.5) is usually named *presynaptic gating*. A convenient choice for the thresholds is the average value of the corresponding activity value. It is also possible to combine these two approaches by subtracting thresholds from both the pre- and postsynaptic activity, this then leads to the so-called *covariance rule*:

$$\frac{dw_{ij}}{dt} = \alpha(v_i - \langle v_i \rangle)(u_j - \langle u_j \rangle), \tag{2.6}$$

where the angular brackets $\langle\rangle$ denote the average over the training period. Note that this rule introduces the interesting (and maybe undesirable) effect that LTP is induced if both pre- and postsynaptic activities are low.

Although each of these covariance rules allow synaptic weights to decrease by introducing LTD to the learning equations, they are still unstable because of the same positive feedback argument given in the case of the basic Hebb rule. For example, even if the postsynaptic activity is gated by a constant threshold the weights would keep growing if the activity is above this threshold, thereby increasing the postsynaptic activity even further, and so on. Also, covariance rules are still non-competitive, unless the thresholds are allowed to slide as it is the case for the BCM-rule which is explained in detail in section 2.3.

2.2.3. Spike-Timing Dependent Plasticity

So far only pure rate models have been considered where the temporal change of the synaptic efficacy depends on scalar activity variables, however, experiments show that

the precise timing of spikes is also important. In fact, spike-timing dependent plasticity (STDP) has emerged in the recent years as the experimentally most studied form of synaptic plasticity (Abbott and Nelson, 2000; Song et al., 2000; Bi and Poo, 2001; Froemke and Dan, 2002; Legenstein et al., 2005). It is a formulaton of Hebb's principle for spike-based models where the relative timing of spikes has a decisive influence on the magnitude and direction of the weight change, i.e., a synaptic weight is increased if the presynaptic spike occurs shortly before the postsynaptic neuron fires, but it is weakened if a presynaptic action potential is preceded by a postsynaptic event[3].

The weight change is given by the equation

$$\frac{dw_{ij}(t)}{dt} = \int_0^\infty W(s) S_j(t) S_i(t-s) ds, \tag{2.7}$$

where $S_j(t) = \sum_f \delta(t - t_j^{(f)})$ and $S_i(t) = \sum_f \delta(t - t_i^{(f)})$ are the pre- and postsynaptic spike trains[4], respectively, and $W(s)$ is the so-called *learning window* that specifies the amount of change in synaptic strength in dependence of the time-difference s between pre- and postsynaptic action potentials. A learning window is usually given by

$$W(s) = \begin{cases} W_+ \exp\left(\frac{s}{\tau_+}\right) & \text{if } s < 0, \\ -W_- \exp\left(-\frac{s}{\tau_-}\right) & \text{if } s > 0, \end{cases} \tag{2.8}$$

with postive parameters W_+, W_- and time constants τ_+, τ_- specifying the amount of LTP and LTD. A typical learning window is shown in figure 2.1, which illustrates that the direction of weight changes depends on the temporal order of pre- and postsynaptic spikes.

With this version of STDP the weights will still saturate at the minimum or maximum efficacy allowed, however, temporal competition between synapses is introduced (Song et al., 2000). This is because different synapses control the timing of postsynaptic spikes. Synapses that are able to evoke postsynaptic action potentials get strengthened, whereas those synapses that are less effective in controlling postsynaptic spiking are weakened.

2.3. The Bienenstock-Cooper-Munro Rule

The covariance based rules in section 2.2.2 have the ability to produce LTD, but are still non-competitive and unstable. An alternative plasticity rule was suggested by Bienenstock, Cooper, and Munro originally in the context of development of stimulus selectivity in the visual cortex (Bienenstock et al., 1982), where the change in the synaptic efficacy not only depends on the instantaneous pre- and postsynaptic instantaneous activities, but also on a slowly varying time-averaged value of the postsynaptic activity. In other words, in contrast to the postsynaptic gating rule in equation (2.4) where LTD and LTP

[3] In the case of anti-Hebbian plasticity, the temporal order of spikes is reversed, i.e., the synaptic weight is strengthened if the presynaptic spike occurs shortly *after* the postsynaptic action potential.

[4] Here, spike trains are represented as sums of Dirac-δ functions located at the pre- or postsynaptic spike times $t_j^{(f)}$ or $t_i^{(f)}$, where the sum runs over all pre- or postsynaptic spikes f.

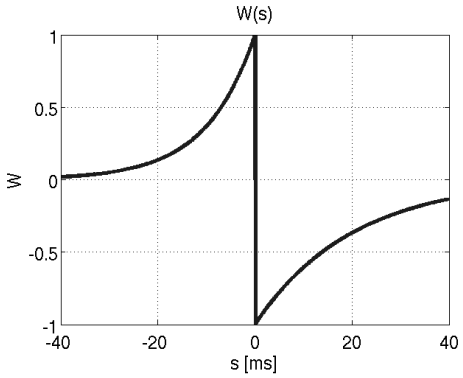


Figure 2.1.: Typical two-phase learning window of STDP (2.8) as a function of the time difference $s = t_j^{(f)} - t_i^{(f)}$ between pre- and postsynaptic spike. Here, $W_+ = W_- = 1$, $\tau_+ = 10$ms and $\tau_- = 20$ms (Gerstner and Kistler, 2002).

are separated by a fixed threshold on the postsynaptic activity, the threshold is allowed to vary in a way that depends nonlinearly on a running average of the postsynaptic rate. This rule has been called the BCM rule (see also e.g., (Intrator and Cooper, 1998; Cooper et al., 2004)) according to the initial letters of its inventors. It takes the form

$$\frac{dw_{ij}(t)}{dt} = \alpha\phi(v_i, \bar{v}_i)u_j - \gamma w_{ij}, \tag{2.9}$$

where $\bar{v}_i$ is a running average of the activity of the postsynaptic neuron, v_i, and γw_{ij} is an (optional) weight decay term. ϕ is a nonlinear function of v_i and $\bar{v}_i$ that is negative if the postsynaptic rate is below a certain threshold (thereby introducing LTD) and positive above this threshold (accounting for LTP), i.e.,

$$\phi(v, \bar{v}) \begin{cases} > 0 & \text{if } v > \theta(\bar{v}), \\ < 0 & \text{if } v < \theta(\bar{v}). \end{cases} \tag{2.10}$$

The threshold θ itself, where the function ϕ changes sign, is a non-linear function of the average postsynaptic activity $\bar{v}$, usually the following is chosen:

$$\theta(\bar{v}) = \left(\frac{\bar{v}}{c_0}\right)^p \bar{v}, \tag{2.11}$$

with some positive constants c_0 and p.

An example for the function $\phi(v_i, \bar{v}_i)$ is shown in figure 2.2 that illustrates the regimes of LTD and LTP separated by a threshold. Sometimes a second threshold is introduced below which no synaptic modification occurs at all. Figure 2.3 shows the curve for two different values of the threshold, where the bottom plot corresponds to lower postsynaptic activity. It can be seen that in this case the amount of LTP is higher, which is likely

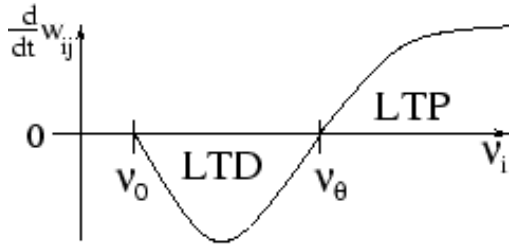


Figure 2.2.: Weight change of the BCM rule as a function of the postsynaptic rate. Synaptic plasticity is characterized by a sliding threshold v_θ that depends on the running average of the postsynaptic activity separating regimes of LTD and LTP: below v_θ synapses are depressed, above v_θ potentiation is observed. Sometimes a second threshold v_0 is introduced below which no synaptic modification occurs (Gerstner and Kistler, 2002).

to result in an increasing postsynaptic firing rate. On the other hand, in the case of higher recent postsynaptic activity (top plot in figure 2.3) the amount of LTD dominates, thereby decreasing the output rate of the postsynaptic neuron. In this way the learning rule is stabilized. To see this, consider some synapses whose efficacies are growing. This results in an increase of postsynaptic activity, which itself results in an increase of the running average of the postsynaptic rate. Due to the properties of the function $\phi(v_i, \bar{v}_i)$ LTD is then introduced at a higher level of postsynaptic activity, so the synaptic strengths stop growing and are stabilized or decrease again. Thus, with the BCM rule the synaptic weights cannot grow without limit as it is the case for the other standard Hebbian learning rules presented in this chapter.

Summarizing, the BCM rule exhibits two basic properties:

- regimes of both LTP and LTD, depending on the activity of the postsynaptic neuron, and
- a sliding threshold that separates these regimes and that depends on a running average of the postsynaptic activity.

It is necessary for stability that the threshold separating LTD and LTP is an adaptive variable; with a fixed threshold the rule would still be unstable (like the covariance or gating rules in section 2.2.2). With a sliding threshold the BCM rule implements competition between synapses because strengthening some synapses increases the postsynaptic firing rate, which raises the threshold and makes it more difficult for other synapses to increase or even maintain their current efficacies.

Figure 2.4 compares the BCM rule with some other Hebbian learning rules presented in this chapter by schematically depicting how synaptic weights are modified as a function of postsynaptic activity. The basic Hebb rule (2.2) can only increase weights and lacks an ability for synaptic depression. The covariance rule (2.6) introduces LTD by using a

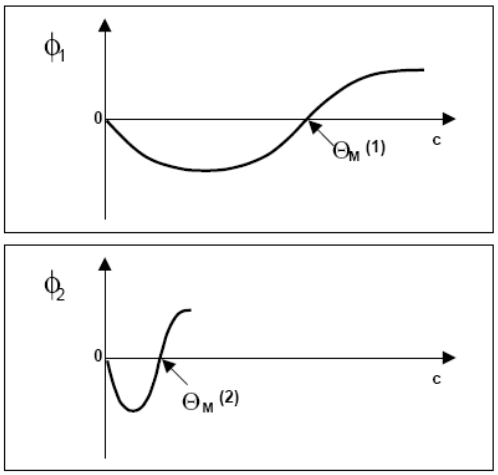


Figure 2.3.: The BCM synaptic modification function ϕ plotted as a function of the output activity of the postsynaptic cell (denoted here as c) for two different values of the threshold θ. The bottom plot corresponds to lower postsynaptic activity (Intrator and Cooper, 1998; Cooper et al., 2004).

fixed threshold for the pre- and/or postsynaptic activity, but still remains unstable and non-competitive. Competition and stability is only introduced with the BCM rule where the regimes of LTP and LTD are separated by a sliding threshold.

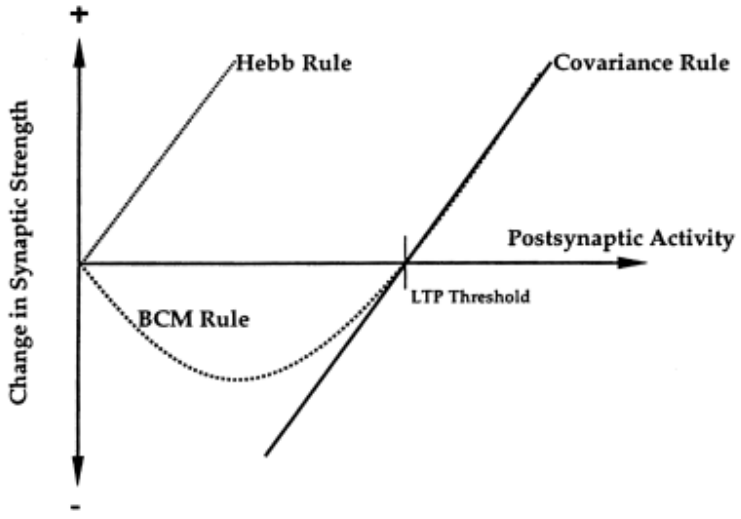


Figure 2.4.: Schematic drawing of the change in synaptic strength as a function of postsynaptic activity for different Hebbian learning rules: basic Hebb rule, covariance rule and BCM rule. Both the covariance rule and the BCM rule postulate a threshold above which there is LTP and below which there is LTD. Taken from (Brown and Chattarji, 1998).

3. Information Theory

In neural coding (Rieke et al., 1997; Dayan and Abbott, 2001) we usually want to know how sensory stimuli are encoded in the response of neurons, i.e., we are interested in the question "What does the response tell us about the stimulus?" In trying to quantify the information that sensory neurons convey about the outside world we instead ask "*How much* does the response tell us about the stimulus?" The techniques for answering this question are provided by Shannon's information theory (Shannon and Weaver, 1949; Cover and Thomas, 1991). Furthermore, it allows us to introduce the concept of optimality in terms of maximal information transmission between the input and output of a neuron.

Information theory was invented by Claude E. Shannon (Shannon and Weaver, 1949) as a general framework for quantifying the ability of a coding scheme or a communications channel to transmit information. In a communications system as proposed by Shannon the transmitter chooses a particular message X out of a set of possible messages. The message is then encoded into a signal and sent over a communications channel before the signal is converted back into a message Y by the receiver. We are interested in cases where we observe some "output" Y and are trying to gain information about the "input" X. Since the channel is usually assumed to be stochastic or noisy it has limited capabilities to convey information, thus the amount of "information" that Y tells us about X is also limited. These quantities are described by the measures of entropy and mutual information, which depend on the probabilities with which each of these messages and combinations of them occur.

This chapter first gives a short overview of the information theoretic measures of entropy and mutual information and how they can be used to quantify the information transmission capability of a neuron. Section 3.2 summarizes a recent approach to maximize the mutual information between the input and output of a spiking neuron model (Toyoizumi et al., 2005a) and explains how the resulting synaptic learning rule relates to the BCM model presented in chapter 2 and how it extends the BCM rule to the case of spiking neurons with refractoriness.

3.1. Quantifying Information Transmission

In neuroscience applications we are interested in the capability of neurons to transmit information, that is, we ask how much the output of a neuron tells us about the input. The neuron itself is considered as a noisy channel and the messages transmitted and received are neural stimuli and responses, i.e., continuous functions of time or spike trains.

In this section, the information theoretic quantities are introduced in a more abstract way using discrete random variables[1]. Bold letters are used to distinguish random variables ($\mathbf{X}$) from their specific realizations (X). Both the probability that the random variable $\mathbf{X}$ takes value X and the probability distribution of $\mathbf{X}$ are denoted as $P(X)$.

3.1.1. Entropy

The entropy $H(\mathbf{X})$ of a random variable $\mathbf{X}$ is a quantity that describes how much information is *available* in the distribution $P(X)$. For example, if $\mathbf{X}$ can take on only one single value then no information (in the colloquial sense) can be transmitted because the message is always the same. Thus, it is necessary to quantify the variability allowed by the distribution $P(X)$. Furthermore, the entropy of the joint distribution of independent random variables should be equal to the sum of entropies of the individual distributions, i.e.

$$H(\mathbf{X}_1, \mathbf{X}_2) = H(\mathbf{X}_1) + H(\mathbf{X}_2), \qquad \text{if } P(X_1, X_2) = P(X_1)P(X_2) \text{ for all } X_1, X_2, \quad (3.1)$$

according to our intuition that the available information coming from independent distributions is simply added. Shannon argues (Shannon and Weaver, 1949) that the negative logarithm is the only function satisfying these two conditions. By averaging over the distribution $P(X)$ he defines the entropy of $\mathbf{X}$ as

$$H(\mathbf{X}) = -\sum_X P(X) \log_2 P(X), \quad (3.2)$$

where the sum runs over all possible values X of the random variable $\mathbf{X}$. The term *entropy* comes from physics where an analogous quantity is defined in thermodynamics and statistical mechanics. The base 2 logarithm used in (3.2) indicates that the entropy is measured in "bits".

Intuitively, the entropy measures the surprise or unpredictability associated with a random variable. For example, if the random variable takes on a certain value with probability 1, then the entropy is 0. On the other hand, the entropy is maximal if all possible values occur with equal probability, in this case the distribution contains more randomness and is therefore less predictable. If $\mathbf{X}$ takes on K possible values each with probability $1/K$, then the entropy is equal to $H(\mathbf{X}) = \log_2 K$.

3.1.2. Mutual Information

In order to characterize the amount of information that an output Y carries about the input X, we have to compare the total output distribution $P(Y)$ with the conditional

[1] The entropy is also defined for continuous variables using the probability density function, however, its definition requires some care. Since we could transmit an infinite amount of information using the endless sequence of decimal digits of a single continuous variable, it is necessary to include some limit on the measurement accuracy. Otherwise, each continuous variable would have infinite entropy. On the other hand, entropy differences (such as the mutual information) are well defined (Rieke et al., 1997; Dayan and Abbott, 2001). In this thesis the focus is on discrete variables only.

distribution given the input $P(Y|X)$, averaged over all X. That is, we subtract the average conditional response entropy $H(\mathbf{Y}|\mathbf{X})$ from the total response entropy $H(\mathbf{Y})$; this difference is then called the *mutual information* between the input $\mathbf{X}$ and the output $\mathbf{Y}$:

$$\begin{aligned} I(\mathbf{X},\mathbf{Y}) &= H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}) \\ &= -\sum_{Y} P(Y)\log_2 P(Y) + \sum_{X,Y} P(X)P(Y|X)\log_2 P(Y|X) \\ &= \sum_{X,Y} P(X,Y)\log_2 \frac{P(Y|X)}{P(Y)}, \end{aligned} \tag{3.3}$$

where $P(X,Y)$ is the joint probability that input X and output Y occurs.

The entropy of the probability distribution $P(Y|X)$ is lower (or equal) than the entropy of the distribution $P(Y)$ because the knowledge of the input X cannot increase the unpredictability of the output Y. Usually, the same input always produces similar responses, therefore the distribution $P(Y|X)$ is often sharply peaked and thus has lower entropy than $P(Y)$. If in the ideal case, each input X has a distinct output Y, i.e., if the output is a deterministic function of the input, the conditional entropy is 0 and the mutual information equals the entropy of the output (or input) distribution. Thus, the entropies of the input and output distributions pose upper limits on the mutual information that can be transmitted. On the other hand, if the output is completely unaffected by the input, i.e., if $P(Y|X) = P(Y)$ for all X and Y, it follows immediately from (3.3) that $I(\mathbf{X},\mathbf{Y}) = 0$. This means that the mutual information between two independent random variables is zero, which can be seen more easily if the last line of (3.3) is rewritten as

$$I(\mathbf{X},\mathbf{Y}) = \sum_{X,Y} P(X,Y)\log_2 \frac{P(X,Y)}{P(X)P(Y)}, \tag{3.4}$$

which compares the joint distribution $P(X,Y)$ with the independent distribution $P(X)\cdot P(Y)$. Thus another way of thinking about the mutual information is how much the entropy of the whole system $H(\mathbf{X},\mathbf{Y})$ is less than the sum of the entropies $H(\mathbf{X})$ and $H(\mathbf{Y})$ (which would be the entropy of the system if $\mathbf{X}$ and $\mathbf{Y}$ were independent), i.e.

$$I(\mathbf{X},\mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y}) - H(\mathbf{X},\mathbf{Y}). \tag{3.5}$$

3.1.3. Information Rate

When quantifying the entropy of or the mutual information between signals (e.g., spike trains) we find that these values are usually proportional to the length of the signal, i.e., the time over which it is observed. Obviously, longer spike trains have a higher entropy, and one can encode an arbitrary amount of information with a sufficiently long spike train. To measure information transmission independent of the signal duration it is thus natural to define the *entropy rate*,

$$H'(\mathbf{X}^T) = \frac{H(\mathbf{X}^T)}{T}, \tag{3.6}$$

as the entropy divided by the time T. Here, $\mathbf{X}^T$ indicates that this random variable describes some signal of length T, e.g., $\mathbf{X}^T$ might characterize all spike trains of duration T. Analogously, the *information rate* can be defined as

$$I'(\mathbf{X}^T, \mathbf{Y}^T) = \frac{I(\mathbf{X}^T, \mathbf{Y}^T)}{T}. \tag{3.7}$$

For $T \to 0$ (3.7) defines the momentary information rate at a given point in time, for $T \to \infty$ it refers to the "true" information rate provided by the signals described by $\mathbf{X}$ and $\mathbf{Y}$.

3.1.4. Kullback-Leibler Divergence

Another measure commonly used in statistics which is also related to information theory is the *Kullback-Leibler (KL) divergence*. It is a similarity measure for probability distributions and defines a "distance" between two distributions $P(X)$ and $Q(X)$:

$$D(P(X)||Q(X)) = \sum_X P(X) \log_2 \frac{P(X)}{Q(X)}. \tag{3.8}$$

Equation (3.8) has the property of a distance measure that $D(P(X)||Q(X)) \geq 0$ with equality if and only if $P(X) = Q(X)$ for all X. However, unlike a distance measure it is not symmetric, i.e. in general $D(P(X)||Q(X)) \neq D(Q(X)||P(X))$[2].

Comparing (3.8) with (3.4) it can be seen that the mutual information between $\mathbf{X}$ and $\mathbf{Y}$ is actually a Kullback-Leibler divergence between the distributions $P(X, Y)$ and $Q(X, Y) = P(X)P(Y)$. In this way, the mutual information (3.4) measures the "distance" between the joint distribution and the independent distributions, or how far the variables $\mathbf{X}$ and $\mathbf{Y}$ are away from being independent.

3.2. Maximizing Mutual Information

Entropy and mutual information are useful quantities for characterizing the efficiency of neural coding and selectivity. It is then natural to ask under what conditions a neuron transmits as much information as possible, i.e., we introduce the concept of optimality in terms of information transmission. However, each optimization has to be performed under some constraints. For instance, it would be possible to encode an infinite amount of information in the output of a single neuron if the postsynaptic firing rate could take on arbitrarily high values, which is not realistic from a biological point of view. Thus, it is essential to include some constraint that limits the firing rate to a realistic range, e.g., by holding the average firing rate fixed.

Information theoretic concepts have so far been used in the context of neuroscience mainly because they allow to compare the performance of neural systems with theoretical limits, but synaptic update rules for optimal information transmission have yet been

[2]Note that the mutual information is symmetric in its arguments (i.e., $I(\mathbf{X}, \mathbf{Y}) = I(\mathbf{Y}, \mathbf{X})$), but the KL divergence is not.

analyzed mainly for pure rate models (Linsker, 1989; Nadal and Parga, 1997; Bell and Sejnowski, 1995). Recent work has also been done on interpreting spike-timing dependent plasticity (Bell and Parra, 2005; Bohte and Mozer, 2005; Chechik, 2003; Pfister et al., 2005; Toyoizumi et al., 2005c) and intrinsic plasticity (Triesch, 2005) in this context.

In this section, the results of a different approach (Toyoizumi et al., 2005a) are presented. In this paper the authors derive a synaptic update rule for a spiking-neuron model with refractoriness that maximizes the mutual information between an ensemble of presynaptic spike trains and the output of the postsynaptic neuron. Instead of looking at a specific implementation of synaptic plasticity and analyzing it in the context of information transmission they ask what the optimal synaptic update rule would be that guarantees to transmit as much information as possible. Mutual information is maximized under the constraint that the postsynaptic firing rate stays as close as possible to a constant target firing rate. This idea is consistent with the widespread findings of homeostatic processes that try to push the neuron back into its preferred target firing state (Turrigiano and Nelson, 2004). The resulting learning rule exhibits the basic properties of the BCM rule (section 2.3), i.e., regimes of LTP and LTD separated by a sliding threshold, and is thus a natural extension of the classical BCM rule to the case of spiking neurons.

3.2.1. Stochastically Spiking Neuron Model

The learning rule presented in (Toyoizumi et al., 2005a) extends the BCM model, which was originally designed for a pure rate model of neuronal activity, to the case of spiking neurons with refractoriness. Several spiking neuron models like the Integrate-and-fire neurons account for a broad range of neuronal firing behavior (Gerstner and Kistler, 2002). Most of these models are deterministic, e.g., an action potential is fired whenever the membrane potential reaches a certain threshold from below. However, it turns out that a stochastically spiking neuron model is better suited for an information theoretic analysis, where it is necessary to define a probabilistic relationship between input and output spike trains. In the model presented in (Toyoizumi et al., 2005a) a spike is generated at each time with a probability that depends on the current membrane potential and the time since the last output spike. The use of such a stochastic model allows an easier formulation of the probabilistic relationship between spike trains as it is needed for quantifying the mutual information and makes this value differentiable with respect to the neuron's weights, thereby allowing us to formulate a gradient learning rule.

It is convenient to formulate the model in discrete time with step size Δt, where t^k denotes the k-th time step, i.e., $t^k = k\Delta t$. The postsynaptic neuron receives input through N synapses. A presynaptic spike train at synapse j ($j = 1, \ldots, N$) is described as a sequence x_j^k ($k = 1, \ldots, K$) of zeros (no spike) and ones (spike). The upper index k denotes time bin k. Thus, $x_j^k = 1$ indicates that a presynaptic spike arrived at synapse j at a time t_j with $t^{k-1} \leq t_j \leq t^k$. Each presynaptic spike evokes a PSP with exponential time course $\epsilon(t - t_j^{(f)})$ with time constant $\tau_m = 10$ms. The membrane potential at time

step t^k is calculated as the total PSP

$$u(t^k) = u_r + \sum_{j=1}^{N} \sum_{n=1}^{k} w_j \epsilon(t^k - t^n) x_j^n, \tag{3.9}$$

where $u_r = -70\text{mV}$ is the resting potential and w_j is the weight of synapse j.

The probability ρ^k of firing in time step k is a function of the membrane potential u and the refractory state R of the neuron,

$$\rho^k = 1 - \exp[-g(u(t^k))R(t^k)\Delta t] \approx g(u(t^k))R(t^k)\Delta t. \tag{3.10}$$

g is a smooth increasing function of the membrane potential; thus, the larger the membrane potential, the higher the probability of emitting a spike. For $\Delta t \to 0$, we may think of $g(u)R(t)$ as the instantaneous firing rate. In (Toyoizumi et al., 2005a), they choose

$$g(u) = r_0 \log \left\{ 1 + \exp \left[\frac{u - u_0}{\Delta u} \right] \right\}, \tag{3.11}$$

which implements a stochastic threshold around u_0. Below u_0 the firing probability goes to 0, above u_0 it increases linearly with the membrane potential (with slope $r_0/\Delta u$; see figure 3.1(a)). $R(t)$ is the refractory state of the neuron ($R(t) \in [0, 1]$) which depends only on the time of the last postsynaptic spike $\hat{t}$,

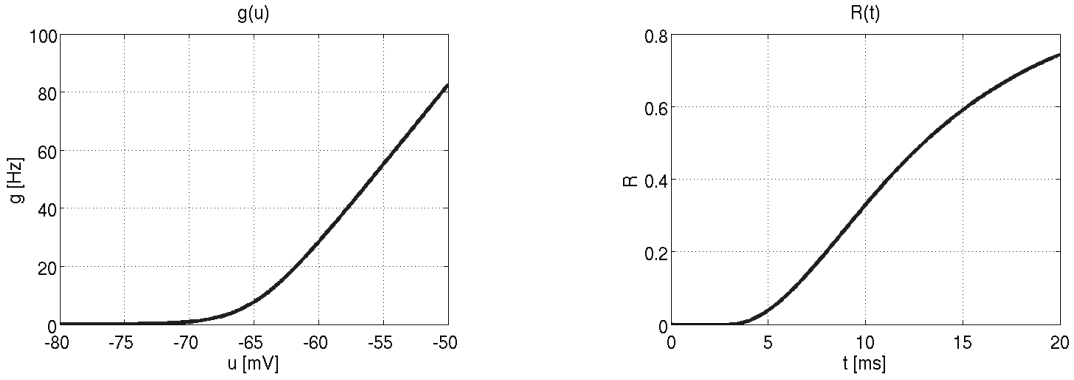
$$R(t) = \frac{(t - \hat{t} - \tau_{abs})^2}{\tau_{refr}^2 + (t - \hat{t} - \tau_{abs})^2} \Theta(t - \hat{t} - \tau_{abs}), \tag{3.12}$$

where τ_{abs} is the absolute refractory time, i.e., no spike can occur before τ_{abs} after the last postsynaptic event. τ_{refr} models the relative refractory time, during which it is hard, but not impossible, to emit a spike; this parameter specifies how fast $R(t)$ in (3.12) goes back to 1 (see figure 3.1(b)). The Heaviside function $\Theta(x)$ takes a value of 1 for positive arguments and 0 otherwise. With a function $R(t)$ as in (3.12) the neuron model has the properties of a renewal process, where the state of the system (and hence the probability of generating the next event) depends only on the "age" of the system, i.e., the time $t - \hat{t}$ since the last event (Gerstner and Kistler, 2002). However, the model can easily be generalized to include the dependence on earlier spikes as well.

The output of the postsynaptic neuron at time step k is denoted as a variable $y_i^k = 1$ if a postsynaptic spike occurred and 0 otherwise. A specific spike train up to time step k is denoted with an uppercase letter, $Y^k = (y_i^1, y_i^2, \ldots, y_i^k)$. Since spikes are generated by a random process, it is important to distinguish the random variable $\mathbf{Y}^k$ from a specific realization Y^k. The same holds for the input, $\mathbf{X}^k$ is the random variable characterizing the inputs at all synapses $1 \leq j \leq N$ up to time step k, X^k is a specific realization of all input spike trains up to time step k and $X_j^k = (x_j^1, x_j^2, \ldots, x_j^k)$ is a specific spike train at synapse j.

For given input spike trains X^k and postsynaptic spike history Y^{k-1} we can write the probability of emitting a postsynaptic spike at time step k using the firing probability ρ^k (3.10) as

$$P(y^k | Y^{k-1}, X^k) = (\rho^k)^{y^k} (1 - \rho^k)^{(1-y^k)}, \tag{3.13}$$



(a) (b)

Figure 3.1.: (a) Gain function g (3.11) as a function of the membrane potential u. $u_0 = -65\text{mV}$, $\Delta u = 2\text{mV}$, $r_0 = 11\text{Hz}$. (b) Refractory state R (3.12) as a function of the time since the last postsynaptic spike $t - \hat{t}$ ($\hat{t} = 0$). $\tau_{abs} = 3\text{ms}$, $\tau_{refr} = 10\text{ms}$.

which is a binary distribution because it evaluates to either ρ^k or $1 - \rho^k$, depending on $y^k \in \{0, 1\}$. The marginal probability, given only the postsynaptic history, is found as

$$P(y^k|Y^{k-1}) = (\bar{\rho}^k)^{y^k}(1 - \bar{\rho}^k)^{(1-y^k)}, \tag{3.14}$$

where $\bar{\rho}^k = \left\langle \rho^k \right\rangle_{\mathbf{X}^k|Y^{k-1}} = \sum_{X^k} \rho^k P(X^k|Y^{k-1})$ is the average firing probability in time step k.

Since the spiking probabilities for each time step are independent given the postsynaptic history, we obtain the probability of an entire output spike train Y^K given the input X^K by taking the product over all binwise probabilities,

$$P(Y^K|X^K) = \prod_{k=1}^{K} P(y^k|Y^{k-1}, X^k) = \prod_{k=1}^{K} (\rho^k)^{y^k}(1 - \rho^k)^{(1-y^k)}, \tag{3.15}$$

and analogously, for the probability of an output spike train,

$$P(Y^K) = \prod_{k=1}^{K} P(y^k|Y^{k-1}) = \prod_{k=1}^{K} (\bar{\rho}^k)^{y^k}(1 - \bar{\rho}^k)^{(1-y^k)}. \tag{3.16}$$

With equations (3.13) to (3.16) a probabilistic relationship between an output spike train and an ensemble of input spike trains has been established, which is needed for quantifying the information transmission of the neuron.

3.2.2. Maximizing Information Transmission

The information transmitted between an ensemble of input spike trains $\mathbf{X}^K$ and an output spike train $\mathbf{Y}^K$ of total duration $K\Delta t$ can be quantified by the mutual information (3.3)

$$I(\mathbf{Y}^K, \mathbf{X}^K) = \sum_{Y^K, X^K} P(Y^K, X^K) \log \frac{P(Y^K|X^K)}{P(Y^K)}. \tag{3.17}$$

It is easier to transmit a large amout of information if the neuron keeps firing at a high rate, in this way more information about the input spike trains can be encoded in the output spike train. However, this is costly from the point of view of energy consumption and also difficult to implement by the biophysical machinery. Therefore, information transmission is optimized under the condition that the firing statistics $P(Y^K)$ stays as close as possible to a target distribution $\tilde{P}(Y^K)$, which is chosen to be that of a constant instantaneous rate $\tilde{g}$ (30Hz) modulated by the refractory variable $R(t)$. This "distance" can be expressed by the Kullback-Leibler divergence (3.8), and the quantity to maximize is thus

$$L = I(\mathbf{Y}^K, \mathbf{X}^K) - \gamma D(P(Y^K)||\tilde{P}(Y^K)), \tag{3.18}$$

with some positive constant γ that controls the influence of how far away the current firing behavior is from the target firing rate.

Assuming that the weights w_j can change between some bounds $0 \leq w_j \leq w_{max}$ a gradient ascent rule is derived on L (3.18). Using equations (3.13) to (3.16), equation (3.18) can be rewritten as $L = \sum_{k=1}^{K} \Delta L^k$, with

$$\Delta L^k = \left\langle \log \frac{P(y^k|Y^{k-1}, X^k)}{P(y^k|Y^{k-1})} - \gamma \log \frac{P(y^k|Y^{k-1})}{\tilde{P}(y^k|Y^{k-1})} \right\rangle_{\mathbf{X}^k, \mathbf{Y}^k}, \tag{3.19}$$

where $\langle\cdot\rangle_{\mathbf{X}^k, \mathbf{Y}^k}$ denotes the average over the joint distribution $P(X^k, Y^k)$, i.e., the L term in (3.18) is decomposed into separate contributions for each time bin. Applying gradient ascent to the weight w_j, it is changed in each step by

$$\Delta w_j^k = \alpha \frac{\partial \Delta L^k}{\partial w_j}, \tag{3.20}$$

with some learning rate α[3]. Evaluation of the gradient (see (Toyoizumi et al., 2005a; Toyoizumi et al., 2005b) for details) yields

$$\Delta w_j^k = \alpha \left\langle C_j^k (F^k - \gamma G^k) \right\rangle_{\mathbf{X}^k, \mathbf{Y}^k}, \tag{3.21}$$

[3]Note that in equations (3.17) and (3.19) the base 2 of the logarithm, which is usually used in information theory, has for simplicity been omitted. All logarithms are equal up to a constant factor which can be accounted for in the learning rate α.

with terms

$$C_j^k = \sum_{l=k-k_a}^{k} \sum_{n=1}^{l} \epsilon(t^l - t^n) x_j^n \frac{\partial \rho^l}{\partial u} \left[\frac{y^l}{\rho^l} - \frac{1-y^l}{1-\rho^l} \right], \tag{3.22}$$

$$F^k = \log \frac{P(y^k|Y^{k-1}, X^k)}{P(y^k|Y^{k-1})} = y^k \log \frac{\rho^k}{\bar{\rho}^k} + (1-y^k) \log \frac{1-\rho^k}{1-\bar{\rho}^k}, \tag{3.23}$$

$$G^k = \log \frac{P(y^k|Y^{k-1})}{\tilde{P}(y^k|Y^{k-1})} = y^k \log \frac{\bar{\rho}^k}{\tilde{\rho}^k} + (1-y^k) \log \frac{1-\bar{\rho}^k}{1-\tilde{\rho}^k}, \tag{3.24}$$

where C_j^k is a coincidence measure between postsynaptic spikes and PSPs generated by presynaptic spikes at synapse j. The time span k_a of the coincidence window is given by the width of the autocorrelation of the postsynaptic spike train (Toyoizumi et al., 2005b). The term F^k compares the instantaneous firing probability ρ^k at time step k with the average probability $\bar{\rho}^k$, and analogously, G^k compares the average firing probability $\bar{\rho}^k$ with the target value $\tilde{\rho}^k = \tilde{g} R(t^k) \Delta t$.

Under the assumption of a small learning rate α the expectations in (3.21) can be approximated by averaging over a single long trial, and taking the limit $\Delta t \to 0$ one can define an online rule

$$\frac{dw_j(t)}{dt} = \alpha C_j(t) B^{post}(t-\delta), \tag{3.25}$$

with a postsynaptic factor

$$\begin{aligned} B^{post}(t) = \delta(t - \hat{t} - \delta) \log \left[\frac{g(u(t))}{\bar{g}(t)} \left(\frac{\tilde{g}}{\bar{g}(t)} \right)^{\gamma} \right] \\ - R(t)[g(u(t)) - (1+\gamma)\bar{g}(t) + \gamma \tilde{g}], \end{aligned} \tag{3.26}$$

where $\hat{t}$ is the firing time of the last postsynaptic spike and δ inside the Dirac-δ function and in (3.25) is a small delay. The rate $\bar{g}(t) = \langle g(u(t)) \rangle_{\mathbf{X}|Y}$ denotes an expectation over the input distribution given the postsynaptic firing history and can be estimated by a running average of $g(t)$ with a large exponential time window (with a time constant of 10s). The term $B^{post}(t)$ can be decomposed into two terms: the first one compares the instantaneous firing intensity $g(u)$ with its running average $\bar{g}(t)$, thereby measuring the momentary significance of the postsynaptic state; the second term compares the running average with the target rate $\tilde{g}$, which accounts for homeostatic processes. During postsynaptic action potentials the postsynaptic term B^{post} has marked peaks (see figure 3.2). Their amplitude and sign depend on the membrane potential at the moment of action potential firing.

The term $C_j(t)$ in (3.25) is sensitive to correlations between the postsynaptic neuron and its presynaptic input at synapse j and is given by the differential equation

$$\frac{dC_j(t)}{dt} = -\frac{C_j(t-\delta)}{\tau_C} + \sum_f \epsilon(t - t_j^{(f)}) S(t) [\delta(t - \hat{t} - \delta) - g(u(t)) R(t)], \tag{3.27}$$

with a time constant $\tau_C = 1$s. Here $g(u(t))R(t)$ is the instantaneous firing rate of the neuron modulated by the refractory function $R(t)$, and $S(t) = g'(u(t))/g(u(t))$ is the

sensitivity (the prime denoting the derivate with respect to u) of the neuron to a change of its membrane potential. The term with the Dirac-δ function induces a positive jump of C_j immediately (with short delay δ) after each postsynaptic spike. Between postsynaptic spikes, C_j evolves continuously. Significant changes of C_j are conditioned on the presence of a PSP $\epsilon(t - t_j^{(f)})$ caused by spike arrival at synapse j. In the absence of presynaptic input, the correlation estimate decays with time constant τ_C back to 0.

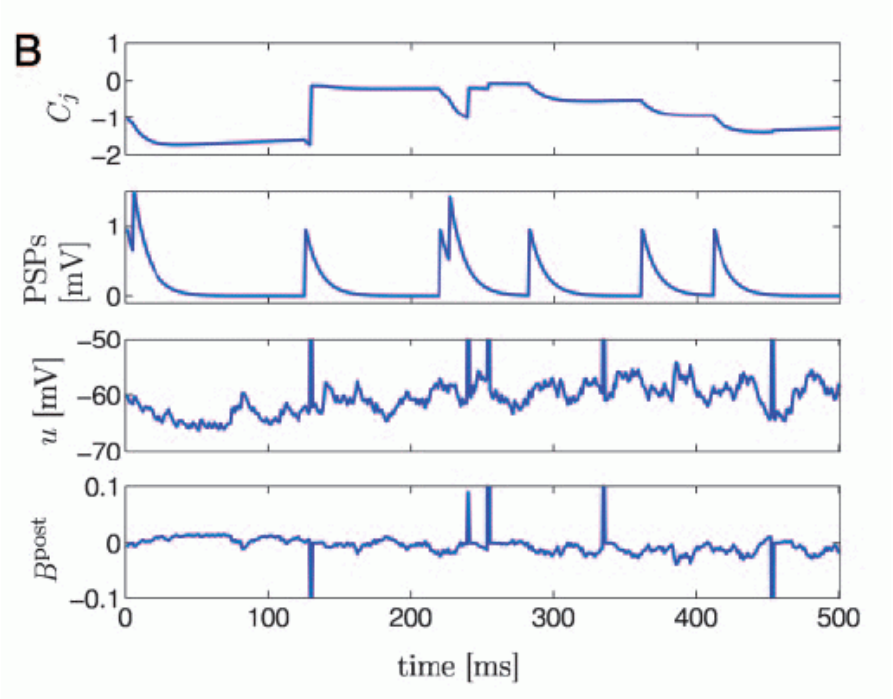


Figure 3.2.: Visualization of the terms C_j and B^{post} during 500ms. From top to bottom: the measure C_j that is sensitive to correlations between the state of the postsynaptic neuron and presynaptic spike arrival at synapse j, the PSPs caused by spike arrivals at the same synapse j, the membrane potential u, and the postsynaptic factor B^{post} as a function of time. Taken from (Toyoizumi et al., 2005a).

Both the correlation term C_j and the postsynaptic factor B^{post} can be estimated online (see figure 3.2) and use only information that could be available at the site of the synapse, thus (3.25) is a local learning rule. The direction of change in the value of a synaptic efficacy is determined by a subtle interplay between these terms, which can both be negative or positive. This learning rule derived from the principle of information maximization for a spiking neuron has some interesting properties, e.g., it drives neurons to spontaneously detect and specialize for groups of coherent inputs, and it is also sensitive to weak spike-spike correlations in the input (Toyoizumi et al., 2005a).

3.2.3. Relation to the BCM Rule

Maximizing the mutual information between an ensemble of input spike trains and the postsynaptic output spike train of a neuron yields a synaptic update rule (3.25) that depends on correlations between pre- and postsynatpic activity measured by the term C_j and a variable B^{post} that characterizes the postsynaptic state. Remember from chapter 2 that in the standard formulation of Hebbian learning the changes of synaptic efficacies are driven by correlations between the activities of pre- and postsynaptic neurons, similar to the function $C_j(t)$. However, in the learning rule (3.25) these correlations are augmented by a postsynaptic factor B^{post} that can change the direction of synaptic weight change depending on the firing behavior of the postsynaptic neuron compared to its recent history and to the desired target activity.

To explore the balance between potentiation and depression of synapses, a simplified neuron model without refractoriness is considered (Toyoizumi et al., 2005a). In this special case it reduces to a pure rate model with Poisson firing statistics and the synaptic update rule (3.25) can be written as

$$\frac{dw_j(t)}{dt} = \alpha v_j \phi(v^{post}, \theta), \tag{3.28}$$

where v_j is the presynaptic firing rate of synapse j. $\phi(v^{post}, \theta)$ is a function that depends on the instantaneous postsynaptic firing rate v^{post} and a parameter θ that denotes the transition from the regime of potentiation to that of depression. It depends on the recent firing history of the neuron and is given by

$$\theta(t) = \bar{v}^{post}(t) \left(\frac{\bar{v}^{post}(t)}{\tilde{g}} \right)^{\gamma}, \tag{3.29}$$

where $\tilde{g}$ denotes a target value for the postsynaptic firing rate implemented by homeostatic processes and $\bar{v}^{post}(t)$ is a running average of the postsynaptic rate.

Note the similarity of equations (3.28) and (3.29) to equations (2.9) and (2.11) on page 11, respectively. The function $\phi(v^{post}, \theta)$ shown in figure 3.3 is characteristic for the BCM learning rule presented in section 2.3: it has regimes of both potentiation and depression, separated by a sliding threshold that depends in a highly nonlinear way on a running average of the postsynaptic activity. Thus, information maximization under the constraint of a fixed target firing rate automatically yields a learning rule with properties similar to that postulated in (Bienenstock et al., 1982), and thereby extends the classical BCM rule to the case of spiking neurons with refractoriness.

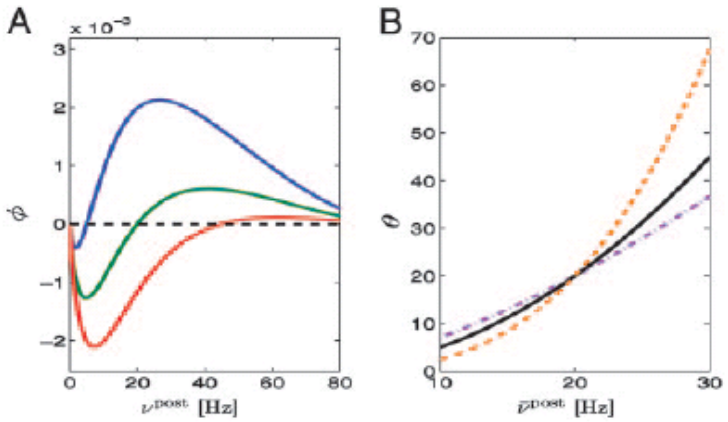


Figure 3.3.: Relation of the update rule (3.25) to the classical BCM rule in the case of Poisson firing statistics. (A) The function $\phi(v^{post}, \theta)$ has regimes of both LTD and LTP separated by a sliding threshold θ that depends on the average postsynaptic firing rate $\bar{v}^{post}$. It is shown as a function of v^{post} for 3 different values of $\bar{v}^{post}$ (from top to bottom: 10Hz, 20Hz, 30Hz; $\tilde{g} = 20$Hz). (B) The threshold θ as a function of $\bar{v}^{post}$ for different choices of γ, i.e., $\gamma = 0.5$ (dashed), $\gamma = 1$ (solid), $\gamma = 2$ (dashdot). Taken from (Toyoizumi et al., 2005a).

4. Extracting Independent Components

In this chapter the main results of this thesis are presented. Based on the generalized Bienenstock-Cooper-Munro learning rule for spiking neurons given in section 3.2, which maximizes the mutual information between the input spike trains and the output spike train of a neuron, this rule is extended in a way that a second neuron which receives the same input also maximizes the mutual information between the input and its output, but at the same time tries to keep the mutual information between its output and the output of the other neuron as low as possible. In this way we try to extract statistically independent components from the inputs.

First, in section 4.1 the methods and models defined in the previous section are repeated briefly since they are reused here again. In section 4.2 we try to find an expression for the mutual information between the output spike trains of two neurons that receive the same input at their synapses. This is then needed in section 4.3 to derive a learning rule that minimizes this quantity. Finally, the results of some simulation experiments of this synaptic update rule are presented in section 4.4.

4.1. Methods and Models

The same neuron model and spike train representations are used as in section 3.2.1, which are repeated here in brevity for convenience.

We use discrete time with step size Δt, where t^k denotes the k-th time step, i.e., $t^k = k\Delta t$. Two postsynaptic neurons receive the same input at N synapses each. A presynaptic spike train at synapse j ($j = 1, \ldots, N$) is described as a sequence x_j^k ($k = 1, \ldots, K$) of zeros (no spike) and ones (spike). Each presynaptic spike evokes a PSP with exponential time course $\epsilon(t - t_j^{(f)})$ with time constant $\tau_m = 10$ms. The membrane potential of neuron i ($i = 1, 2$) at time step t^k is calculated as

$$u_i(t^k) = u_r + \sum_{j=1}^{N} \sum_{n=1}^{k} w_{ij} \epsilon(t^k - t^n) x_j^n, \tag{4.1}$$

where $u_r = -70$mV is the resting potential and w_{ij} is the weight of synapse j of neuron i.

The probability ρ_i^k of neuron i to fire in time step k is a function of the membrane potential and the refractory state of the neuron,

$$\rho_i^k = 1 - \exp[-g(u_i(t^k))R_i(t^k)\Delta t] \approx g(u_i(t^k))R_i(t^k)\Delta t, \tag{4.2}$$

where g and R are the same functions as in equations (3.11) and (3.12) on page 20.

The output of postsynaptic neuron i at time step k is denoted as a variable $y_i^k = 1$ if a postsynaptic spike occurred and 0 otherwise. A specific spike train up to time step k is denoted with an uppercase letter, $Y_i^k = (y_i^1, y_i^2, \ldots, y_i^k)$. The random variable $\mathbf{Y}_i^k$ describes the ensemble of possible output spike trains of neuron i; a specific realization is denoted as Y_i^k. $\mathbf{X}^k$ is the random variable characterizing the inputs at all synapses $1 \leq j \leq N$ up to time step k, X^k is a specific realization of all input spike trains up to time step k and X_j^k is a specific spike train at synapse j.

For given input spike trains X^k and postsynaptic spike history Y_i^{k-1} we can write the probability of neuron i to emit a spike at time step k as

$$P(y_i^k | Y_i^{k-1}, X^k) = (\rho_i^k)^{y_i^k} (1 - \rho_i^k)^{(1-y_i^k)}, \tag{4.3}$$

and the marginal probability, given only the postsynaptic history, as

$$P(y_i^k | Y_i^{k-1}) = (\bar{\rho}_i^k)^{y_i^k} (1 - \bar{\rho}_i^k)^{(1-y_i^k)}, \tag{4.4}$$

with $\bar{\rho}_i^k = \left\langle \rho_i^k \right\rangle_{\mathbf{X}^k | Y_i^{k-1}} = \sum_{X^k} \rho_i^k P(X^k | Y_i^{k-1})$. The probability of an entire output spike train Y_i^K given the input X^K is obtained by taking the product over all binwise probabilities,

$$P(Y_i^K | X^K) = \prod_{k=1}^{K} P(y_i^k | Y_i^{k-1}, X^k) = \prod_{k=1}^{K} (\rho_i^k)^{y_i^k} (1 - \rho_i^k)^{(1-y_i^k)}, \tag{4.5}$$

and analogously, for the probability of an output spike train,

$$P(Y_i^K) = \prod_{k=1}^{K} P(y_i^k | Y_i^{k-1}) = \prod_{k=1}^{K} (\bar{\rho}_i^k)^{y_i^k} (1 - \bar{\rho}_i^k)^{(1-y_i^k)}, \tag{4.6}$$

cf. equations (3.13) to (3.16).

4.2. Mutual Information Between Output Spike Trains

The mutual information (see section 3.1.2) between the output spike trains of both postsynaptic neurons is given as

$$\begin{aligned} I(\mathbf{Y}_1^K, \mathbf{Y}_2^K) &= \sum_{Y_1^K, Y_2^K} P(Y_1^K, Y_2^K) \log \frac{P(Y_1^K | Y_2^K)}{P(Y_1^K)} \\ &= \sum_{Y_1^K, Y_2^K} P(Y_1^K, Y_2^K) \log \frac{P(Y_1^K, Y_2^K)}{P(Y_1^K) P(Y_2^K)}. \end{aligned} \tag{4.7}$$

This measure tells how much information about Y_1^K is conveyed by observing a specific spike train Y_2^K (or vice versa). Note that the mutual information is maximal if there is a one-to-one mapping between both spike trains, e.g., if both spike trains are always equal.

In this case the mutual information equals the entropy of the output distribution. The mutual information is minimal (i.e., zero) if no information is gained about one spike train by knowing the other output, that is, if both random processes are independent, i.e., $P(Y_1^K, Y_2^K) = P(Y_1^K)P(Y_2^K)$ for all Y_1^K, Y_2^K. Note that although both neurons process the input independently, the mutual information is not zero because they receive the same input. Each output implicitly conveys information about the input, thus about the other output. However, if specific input spike trains are held fixed, the output distributions given these inputs are independent.

To evaluate the mutual information between output spike trains Y_1^K and Y_2^K, it is necessary to have an expression for the joint probability $P(Y_1^K, Y_2^K)$ or the conditional probability $P(Y_1^K|Y_2^K)$ (or $P(Y_2^K|Y_1^K)$, respectively). According to the Total Probability Theorem we can write

$$P(Y_1^K|Y_2^K) = \sum_{X^K} P(Y_1^K|X^K, Y_2^K)P(X^K|Y_2^K). \tag{4.8}$$

With Bayes' Theorem we have

$$P(X^K|Y_2^K) = \frac{P(X^K)P(Y_2^K|X^K)}{P(Y_2^K)}, \tag{4.9}$$

and, since for given input X^K, Y_1^K is independent of Y_2^K,

$$P(Y_1^K|X^K, Y_2^K) = P(Y_1^K|X^K). \tag{4.10}$$

Inserting (4.9) and (4.10) back into (4.8), we get

$$P(Y_1^K|Y_2^K) = \frac{1}{P(Y_2^K)} \sum_{X^K} P(X^K)P(Y_1^K|X^K)P(Y_2^K|X^K), \tag{4.11}$$

and since $P(Y_1^K, Y_2^K) = P(Y_1^K|Y_2^K)P(Y_2^K)$,

$$P(Y_1^K, Y_2^K) = \sum_{X^K} P(X^K)P(Y_1^K|X^K)P(Y_2^K|X^K). \tag{4.12}$$

Together with equation (4.5), we can finally write for the joint probability

$$P(Y_1^K, Y_2^K) = \left\langle \prod_{k=1}^{K} (\rho_1^k)^{y_1^k}(1-\rho_1^k)^{(1-y_1^k)}(\rho_2^k)^{y_2^k}(1-\rho_2^k)^{(1-y_2^k)} \right\rangle_{\mathbf{X}^K}. \tag{4.13}$$

The expression in (4.13), however, is difficult to evaluate because the expectation is over the product of the binwise probabilities. It would be preferable if we could formulate the joint probability as a product of binwise probabilities, thereby making the derivative of the mutual information easier. Therefore we turn to a binwise analysis of the output joint probability.

The binwise joint probability of output spikes at time step k, given the postsynaptic histories and the input, can be written as

$$P(y_1^k, y_2^k | Y_1^{k-1}, Y_2^{k-1}, X^k) = P(y_1^k | Y_1^{k-1}, X^k) P(y_2^k | Y_2^{k-1}, X^k), \tag{4.14}$$

since for a given input spike train X^k, the output spike trains Y_1^k and Y_2^k are independent. The marginal joint probability given only the postsynaptic histories is then found as (using (4.14) and (4.3))

$$\begin{aligned} P(y_1^k, y_2^k | Y_1^{k-1}, Y_2^{k-1}) &= \sum_{X^k} P(y_1^k, y_2^k | Y_1^{k-1}, Y_2^{k-1}, X^k) P(X^k | Y_1^{k-1}, Y_2^{k-2}) \\ &= \left\langle (\rho_1^k)^{y_1^k} (1-\rho_1^k)^{(1-y_1^k)} (\rho_2^k)^{y_2^k} (1-\rho_2^k)^{(1-y_2^k)} \right\rangle_{\mathbf{X}^k | Y_1^{k-1}, Y_2^{k-2}}. \end{aligned} \tag{4.15}$$

Analogously to (4.5) and (4.6) the output joint probability is calculated as a product of the joint probabilities of each time bin,

$$P(Y_1^K, Y_2^K) = \prod_{k=1}^{K} P(y_1^k, y_2^k | Y_1^{k-1}, Y_2^{k-1}). \tag{4.16}$$

The expression inside the angular brackets in the last line of (4.15) can be reformulated in the following way:

$$(\rho_1^k \rho_2^k)^{y_1^k y_2^k} (\rho_1^k (1-\rho_2^k))^{y_1^k (1-y_2^k)} ((1-\rho_1^k) \rho_2^k)^{(1-y_1^k) y_2^k} ((1-\rho_1^k)(1-\rho_2^k))^{(1-y_1^k)(1-y_2^k)}; \tag{4.17}$$

in this way, depending on the values of y_1^k and y_2^k, this expression is equal to exactly one of the four terms in equation (4.17) since the exponent of that term is 1, the other ones are 0. In this case we can pull the expectation to the inner terms and write

$$\begin{aligned} P(Y_1^K, Y_2^K) = \prod_{k=1}^{K} &(\bar{\bar{\rho}}_{12}^k)^{y_1^k y_2^k} (\bar{\bar{\rho}}_1^k - \bar{\bar{\rho}}_{12}^k)^{y_1^k (1-y_2^k)} \cdot \\ &\cdot (\bar{\bar{\rho}}_2^k - \bar{\bar{\rho}}_{12}^k)^{(1-y_1^k) y_2^k} (1 - \bar{\bar{\rho}}_1^k - \bar{\bar{\rho}}_2^k + \bar{\bar{\rho}}_{12}^k)^{(1-y_1^k)(1-y_2^k)}, \end{aligned} \tag{4.18}$$

with $\rho_{12}^k = \rho_1^k \rho_2^k$ and $\bar{\bar{\cdot}} = \langle \cdot \rangle_{\mathbf{X}^K | Y_1^{k-1}, Y_2^{k-1}}$.

The mutual information between output spike trains is zero if and only if this joint probability of (4.18) is equal to the independent distribution of the output spike trains, which is repeated here for convenience:

$$\begin{aligned} P(Y_1^K) P(Y_2^K) &= \prod_{k=1}^{K} (\bar{\rho}_1^k)^{y_1^k} (1-\bar{\rho}_1^k)^{(1-y_1^k)} (\bar{\rho}_2^k)^{y_2^k} (1-\bar{\rho}_2^k)^{(1-y_2^k)} \\ &= \prod_{k=1}^{K} (\bar{\rho}_1^k \bar{\rho}_2^k)^{y_1^k y_2^k} (\bar{\rho}_1^k - \bar{\rho}_1^k \bar{\rho}_2^k)^{y_1^k (1-y_2^k)} \cdot \\ &\quad \cdot (\bar{\rho}_2^k - \bar{\rho}_1^k \bar{\rho}_2^k)^{(1-y_1^k) y_2^k} (1 - \bar{\rho}_1^k - \bar{\rho}_2^k + \bar{\rho}_1^k \bar{\rho}_2^k)^{(1-y_1^k)(1-y_2^k)}, \end{aligned} \tag{4.19}$$

with $\bar{\rho}_i^k = \langle \rho_i^k \rangle_{\mathbf{X}^K | Y_i^{k-1}}$.

4.3. Learning Rule

We try to minimize the mutual information between the output of two neurons by deriving a learning rule for neuron 2 while letting the weights of neuron 1 evolve according to the generalized BCM rule that maximizes information transmission (cf. section 3.2). A trivial way to minimize the mutual information between output spike trains is to let the neuron emit the same spike train for all inputs, e.g., emit no spikes at all by setting all weights to 0. In this case the output conveys no information about the input, thus about the other spike train, and the mutual information is 0. Hence, to get a more reasonable result we also maximize the information transmission of neuron 2 and perform the optimization under the additional constraint that the output distribution, $P(Y_2^K)$, stays close to a desired target distribution $\tilde{P}(Y_2^K)$, and we choose that of a neuron with constant instantaneous rate $\tilde{g}$ (e.g., 30Hz), as in section 3.2 and in (Toyoizumi et al., 2005a).

Instead of using the mutual information between the ouput spike trains in the derivation of the learning rule, it turns out that it is more convenient from a mathematical point of view to minimize the information rate (which is the mutual information per time) between the outputs (cf. section 3.1.3). The quantity to maximize is therefore

$$L = I(\mathbf{X}^K, \mathbf{Y}_2^K) - \gamma_1 I'(\mathbf{Y}_1^K, \mathbf{Y}_2^K) - \gamma_2 D(P(Y_2^K)||\tilde{P}(Y_2^K)), \tag{4.20}$$

where $I(\mathbf{X}^K, \mathbf{Y}_2^K)$ is the mutual information between the input spike trains and the output spike train of neuron 2, $I'(\mathbf{Y}_1^K, \mathbf{Y}_2^K) = I(\mathbf{Y}_1^K, \mathbf{Y}_2^K)/(K \cdot \Delta t)$ is the information rate between the output of both neurons, and

$$D(P(Y_2^K)||\tilde{P}(Y_2^K)) = \sum_{Y_2^K} P(Y_2^K) \log \frac{P(Y_2^K)}{\tilde{P}(Y_2^K)} \tag{4.21}$$

is the Kullback-Leibler divergence (see section 3.1.4) between the output distribution and the desired output distribution. γ_1 and γ_2 are constants specifying how hard we want to enforce these additional constraints. Note that since the information rate has dimension bit/s γ_1 has dimension seconds and γ_2 is dimensionless.

We can write equation (4.20) as $L = \sum_{k=1}^{K} \Delta L^k$ with

$$\Delta L^k = \left\langle \log \frac{P(y_2^k|Y_2^{k-1}, X^k)}{P(y_2^k|Y_2^{k-1})} - \frac{\gamma_1}{\Delta t} \log \frac{P(y_1^k, y_2^k|Y_1^{k-1}, Y_2^{k-1})}{P(y_1^k|Y_1^{k-1})P(y_2^k|Y_2^{k-1})} - \gamma_2 \log \frac{P(y_2^k|Y_2^{k-1})}{\tilde{P}(y_2^k|Y_2^{k-1})} \right\rangle_{\mathbf{X}^k, \mathbf{Y}_1^k, \mathbf{Y}_2^k}. \tag{4.22}$$

Assuming slow changes of synaptic weights, a gradient descent algorithm is applied to maximize the objective function L (4.20) and the weight w_{2j} of neuron 2 is changed at each time step by

$$\Delta w_{2j}^k = \alpha \frac{\partial \Delta L^k}{\partial w_{2j}}, \tag{4.23}$$

with an appropriate learning rate α.

After evaluation of the gradient (see appendix A for a detailed derivation) we arrive at a learning rule

$$\Delta w_{2j}^k = \alpha \left\langle C_{2j}^k (F_2^k - \frac{\gamma_1}{\Delta t} F_{12}^k - \gamma_2 G_2^k) \right\rangle_{\mathbf{X}^k, \mathbf{Y}_1^k, \mathbf{Y}_2^k}, \tag{4.24}$$

where

$$C_{2j}^k = \sum_{l=1}^{k} \left[\frac{y_2^l}{\rho_2^l} - \frac{1-y_2^l}{1-\rho_2^l} \right] \frac{\partial \rho_2^l}{\partial u_2} \sum_{n=1}^{l} \epsilon(t^l - t^n) x_2^n \tag{4.25}$$

$$F_2^k = y_2^k \log \frac{\rho_2^k}{\bar{\rho}_2^k} + (1-y_2^k) \log \frac{1-\rho_2^k}{1-\bar{\rho}_2^k} \tag{4.26}$$

$$\begin{aligned} F_{12}^k = {} & y_1^k y_2^k \log \frac{\bar{\bar{\rho}}_{12}^k}{\bar{\rho}_1^k \bar{\rho}_2^k} + y_1^k (1-y_2^k) \log \frac{\bar{\bar{\rho}}_1^k - \bar{\bar{\rho}}_{12}^k}{\bar{\rho}_1^k - \bar{\rho}_1^k \bar{\rho}_2^k} + \\ & + (1-y_1^k) y_2^k \log \frac{\bar{\bar{\rho}}_2^k - \bar{\bar{\rho}}_{12}^k}{\bar{\rho}_2^k - \bar{\rho}_1^k \bar{\rho}_2^k} + \\ & + (1-y_1^k)(1-y_2^k) \log \frac{1 - \bar{\bar{\rho}}_1^k - \bar{\bar{\rho}}_2^k + \bar{\bar{\rho}}_{12}^k}{1 - \bar{\rho}_1^k - \bar{\rho}_1^k + \bar{\rho}_1^k \bar{\rho}_2^k} \end{aligned} \tag{4.27}$$

$$G_2^k = y_2^k \log \frac{\bar{\rho}_2^k}{\tilde{\rho}_2^k} + (1-y_2^k) \log \frac{1-\bar{\rho}_2^k}{1-\tilde{\rho}_2^k}. \tag{4.28}$$

The term C_{2j}^k is sensitive to correlations between the input spike train at synapse j and the output spike train of neuron 2. It counts the coincidences between postsynaptic spikes ($y_2^l = 1$) and the time course of PSPs generated by presynaptic spikes ($x_j^n = 1$), normalized to an expected value $\langle C_{2j}^k \rangle_{\mathbf{Y}^k | X^k} = 0$. The quantity G_2^k compares the average firing probability $\bar{\rho}_2^k$ of neuron 2 at time step k with the desired target firing probability $\tilde{\rho}_2^k = \tilde{g} R(t^k) \Delta t$, thereby trying to maintain the postsynaptic target firing rate $\tilde{g}$, and analogously the term F_2^k compares the instantaneous firing probability ρ_2^k with the average probability $\bar{\rho}_2^k$. Note that each of these three terms also occurs equally in the rule presented in section 3.2 (cf. equations (3.22) to (3.24)).

Aditionally, the value F_{12}^k accounts for the statistical independence between the firing probabilities of both neurons, given the postsynaptic histories. It basically compares the product of average firing probabilities, $\bar{\rho}_1^k \bar{\rho}_2^k$ of neuron 1 and 2 at time step k with the average product of firing probabilites $\bar{\bar{\rho}}_{12}^k = \overline{\overline{\rho_1^k \rho_2^k}}$ (where $\bar{\cdot} = \langle \cdot \rangle_{\mathbf{X}^k | Y_i^{k-1}}$ and $\bar{\bar{\cdot}} = \langle \cdot \rangle_{\mathbf{X}^k | Y_1^{k-1}, Y_2^{k-1}}$). From now on, we assume that $\bar{\rho}_i^k = \bar{\bar{\rho}}_i^k$.

We note that the terms F_2^k, F_{12}^k and G_2^k depend on postsynaptic variables only; we therefore introduce a postsynaptic factor $B_{12}^k = (F_2^k - \frac{\gamma_1}{\Delta t} F_{12}^k - \gamma_2 G_2^k)/\Delta t$ and take the limit $\Delta t \to 0$ (see appendix A for details). Under the assumption of a small learning rate α the expectations $\langle \cdot \rangle_{\mathbf{X}^k, \mathbf{Y}_1^k, \mathbf{Y}_2^k}$ from equation (4.24) can be approximated by averaging over a single long trial that allows us to define an online rule:

$$\frac{dw_{2j}(t)}{dt} = \alpha C_{2j}(t) B_{12}(t - \delta). \tag{4.29}$$

According to (Toyoizumi et al., 2005a; Toyoizumi et al., 2005b) the correlation term C_2^k can be transformed into the differential equation

$$\frac{dC_{2j}(t)}{dt} = -\frac{C_{2j}(t)}{\tau_C} + \sum_f \epsilon(t - t_j^{(f)}) S_2(t)[\delta(t - \hat{t}_2 - \delta) - g(u_2(t))R_2(t)], \tag{4.30}$$

with a time constant $\tau_C = 1$s. Here $g(u_2(t))R_2(t)$ is the instantaneous firing rate of neuron 2 modulated by the refractory function $R_2(t)$, and $S_2(t) = g_2'(u_2(t))/g_2(u_2(t))$ is the sensitivity (the prime denoting the derivate with respect to u) of neuron 2 to a change of its membrane potential. This correlation term is analagous to that in equation (3.27).

The postsynaptic factor $B_{12}(t)$ is composed of two terms

$$B_{12}(t) = B_2^{post}(t) - \gamma_1 B_{12}^{post}(t), \tag{4.31}$$

with

$$\begin{aligned} B_2^{post}(t) = \delta(t - \hat{t}_2 - \delta) \log \left[\frac{g(u_2(t))}{\bar{g}_2(t)} \left(\frac{\tilde{g}}{\bar{g}_2(t)} \right)^{\gamma_2} \right] - \\ - R_2(t) \left[g(u_2(t)) - (1 + \gamma_2)\bar{g}_2(t) + \gamma_2 \tilde{g} \right] \end{aligned} \tag{4.32}$$

and

$$\begin{aligned} B_{12}^{post}(t) = \delta(t - \hat{t}_2 - \delta) \left\{ \delta(t - \hat{t}_1 - \delta) \log \frac{\bar{g}_{12}(t)}{\bar{g}_1(t)\bar{g}_2(t)} - R_1(t) \left[\frac{\bar{g}_{12}(t)}{\bar{g}_2(t)} - \bar{g}_1(t) \right] \right\} - \\ - R_2(t) \left\{ \delta(t - \hat{t}_1 - \delta) \left[\frac{\bar{g}_{12}(t)}{\bar{g}_1(t)} - \bar{g}_2(t) \right] - R_1(t) \left[\bar{g}_{12}(t) - \bar{g}_1(t)\bar{g}_2(t) \right] \right\}, \end{aligned} \tag{4.33}$$

where $\hat{t}_1$ and $\hat{t}_2$ are the last postsynaptic spike times of neuron 1 and neuron 2, respectively. The rate $\bar{g}_i(t) = \langle g(u_i(t)) \rangle_{\mathbf{X}|Y_i}$ denotes an expectation of the instantaneous rate of neuron i over the input distribution given the recent firing history of the postsynaptic neuron and is estimated in a numerical implementation by keeping a running average of the firing rate with a sufficiently large exponential time window (In the implementation a time constant of 10s is used). Similarly, the average product of firing rates $\bar{g}_{12}(t) = \langle g(u_1(t))g(u_2(t)) \rangle_{\mathbf{X}|Y_1,Y_2}$ is calculated by a running average over the product of postsynaptic firing rates with the same time constant. In this way, the quantity $\bar{\rho}_{12}^k = \bar{g}_{12}(t^k)R_1(t^k)R_2(t^k)(\Delta t)^2$ measures the joint probability for the two neurons to spike simultaneously within a time step of size Δt and $\bar{\rho}_1^k \bar{\rho}_2^k = \bar{g}_1(t^k)\bar{g}_2(t^k)R_1(t^k)R_2(t^k)(\Delta t)^2$ estimates the independent distribution of postsynaptic spiking. The term $B_{12}^{post}(t)$ tries to drive the joint distribution close towards the independent distribution, thereby minimizing the mutual information between the output neurons. It compares the average product of postsynaptic rates $\bar{g}_{12}(t)$ with the product of the averages of the postsynaptic rates $\bar{g}_1(t)\bar{g}_2(t)$, for all different postsynaptic states.

Additionally, $B_2^{post}(t)$ compares the average rate of neuron 2 $\bar{g}_2(t)$ with the target rate $\tilde{g}$, thereby accounting for homeostatic processes that tend to push the neuron back into its preferred firing behaviour, and the instantaneous rate $g(u_2(t))$ with the average rate $\bar{g}_2(t)$, reflecting the momentary significance of the postsynaptic rate. This term is analogous to that in equation (3.26).

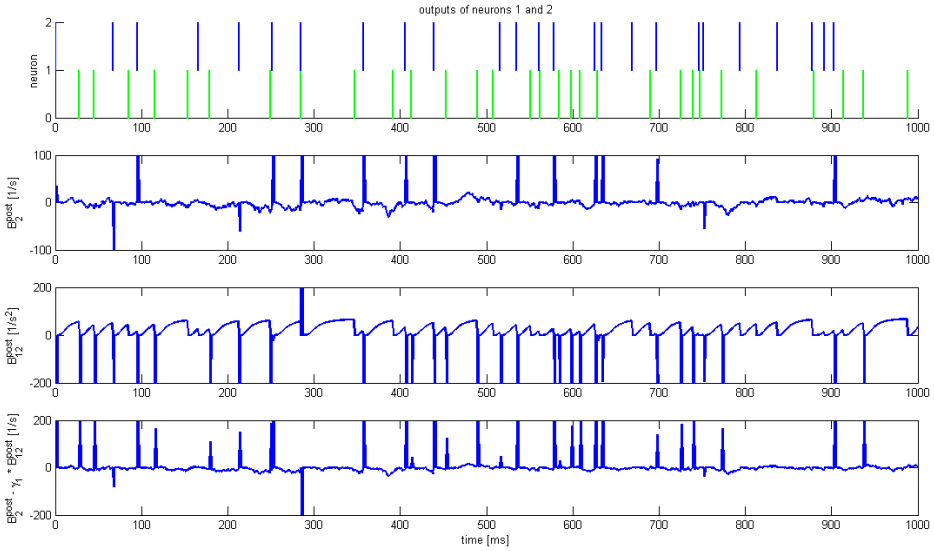


Figure 4.1.: Visualization of the term B_{12}^{post} during 1 second. From top to bottom: the output spike trains of neuron 1 (green) and neuron 2 (blue), the postsynaptic term B_2^{post}, the term sensitive to the momentary statistical dependence between the outputs B_{12}^{post}, and the combined postsynaptic term $B_2^{post} - \gamma_1 B_{12}^{post}$ as a function of time ($\gamma_1 = 0.1$).

Thus, the learning rule (4.29) is basically the same as derived in section 3.2, however, the postsynaptic term B_2^{post} is augmented by an expression B_{12}^{post} that is sensitive to the momentary statistical dependence between the outputs of two neurons. Figure 4.1 shows these terms, as well as the combined postsynaptic term $B_2^{post} - \gamma_1 B_{12}^{post}$, as a function of time for a sample of output spike trains of both neurons during 1 second. One sees that the standard postsynaptic term B_2^{post} has peaks during action potentials of neuron 2 and their sign and amplitude depends on the recent firing history, as already mentioned in section 3.2 (cf. figure 3.2). The term B_{12}^{post} (4.33) usually has negative peaks when one of the two postsynaptic neurons is firing, according to the terms $-\delta(t - \hat{t}_2 - \delta)R_1(t)\left[\frac{\bar{g}_{12}(t)}{\bar{g}_2(t)} - \bar{g}_1(t)\right]$ and $-\delta(t - \hat{t}_1 - \delta)R_2(t)\left[\frac{\bar{g}_{12}(t)}{\bar{g}_1(t)} - \bar{g}_2(t)\right]$, respectively, i.e., the amplitude of these peaks depend on the recent firing history and the refractory state of the other neuron. Additionally, there may be positive peaks when both neurons fire simulta-

neously in the same time step (according to $\delta(t-\hat{t}_2-\delta)\delta(t-\hat{t}_1-\delta)\log\frac{\bar{g}_{12}(t)}{\bar{g}_1(t)\bar{g}_2(t)}$). In times when no neuron is firing, the term evolves according to $R_1(t)R_2(t)\left[\bar{g}_{12}(t)-\bar{g}_1(t)\bar{g}_2(t)\right]$, i.e., it is reminiscent of the time course of the refractory variables (cf. figure 3.1(b)) since the gain averages are approximately constant during one second. The combined postsynaptic term $B_{12}=B_2^{post}-\gamma_1 B_{12}^{post}$ now has additional peaks compared to the original B_2^{post} due to the effect of B_{12}^{post}. It is also possible that some peaks are weakened or strengthened, or that their direction even gets reversed. Note, however, that the actual weight change of neuron 2 still depends on both the combined postsynaptic term and the correlation term C_{2j}.

4.4. Results

Using a setup with two postsynaptic neurons that receive the same input at 100 synapses a learning rule has been derived that minimizes the mutual information between both output spike trains under the constraint that each neuron by itself maximizes information transmission and that the output firing rates stay close to a desired target firing rate. In the following experiments, we let the weights of neuron 1 evolve according to the learning rule of (Toyoizumi et al., 2005a) presented in section 3.2, which maximizes the mutual information between input and output spike trains, and apply the update rule presented in this chapter to the second neuron. More precisely, the learning rule

$$\frac{dw_{1j}}{dt}=\alpha_1 C_{1j}(t)B_1^{post}(t-\delta), \tag{4.34}$$

with $C_{1j}(t)$ and $B_1^{post}(t-\delta)$ defined in (3.27) and (3.26)[1], respectively, is applied to neuron 1 ($\alpha_1=10^{-5}$, $\gamma=1$), and the weights of neuron 2 are changed according to

$$\frac{dw_{2j}(t)}{dt}=\alpha_2 C_{2j}(t)\left[B_2^{post}(t-\delta)-\gamma_1 B_{12}^{post}(t-\delta)\right], \tag{4.35}$$

with the additional term B_{12}^{post} given in (4.33).

4.4.1. Correlation Experiment

In a first experiment the 100 inputs consist of Poisson spike trains at 20Hz each, however, correlation among the inputs is established in the following way: The spike trains at the first 80 synapses are divided into two groups (group 1: spike trains 1 to 40, group 2: spike trains 41 to 80). Within each group a correlation coefficient of 0.5 is established, but spike trains from different groups are uncorrelated. The remaining 20 synapses receive uncorrelated Poisson input (group 3). Weights are initialized randomly between 0.10 and 0.12 for both neurons and can change between the bounds 0 and 1. The results can be seen in figures 4.2 to 4.4.

Figure 4.2 (top) shows the evolution of weights for both neurons during 30 minutes of learning. Weights close to the maximal efficacy $w_{max}=1$ are developed for one

[1]The additional index 1 indicates that the terms are applied to neuron 1.

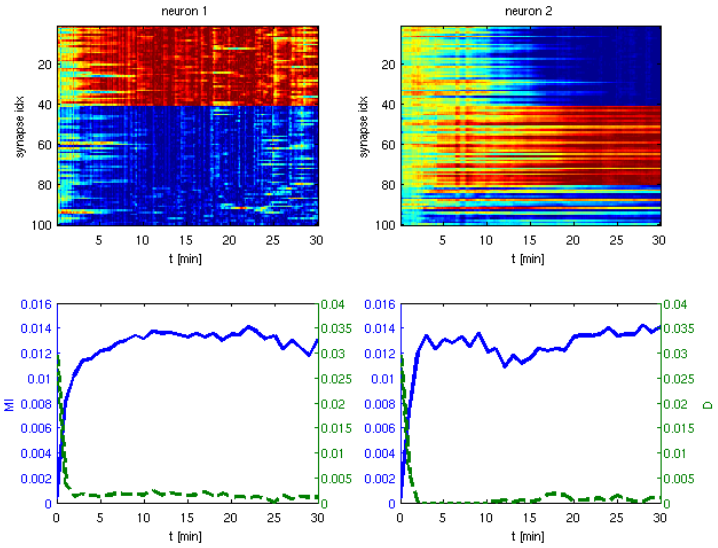


Figure 4.2.: Correlation experiment. (Top) Evolution of weights during 30 minutes of learning for both postsynaptic neurons receiving Poisson input at 20Hz from 100 synapses. Inputs to synapses 1 to 40 and 41 to 80 are both correlated with a coefficient of 0.5, but any two inputs belonging to different grups are uncorrelated. Synapses 81 to 100 receive uncorrelated Poisson input. (red: strong synapses, $w_j \approx 1$, blue: depressed synapses, $w_j \approx 0$.) Weights were initialized randomly between 0.10 and 0.12, $\alpha_2 = 10^{-6}$, $\gamma_1 = 0.1$, $\gamma_2 = 10$. (Bottom) Evolution of the average mutual information per time bin between input and output (blue solid line, left scale) and the Kullback-Leibler divergence per time bin for both neurons (green dashed line, right scale) as a function of time. Averages are calculated over segments of 1 minute.

of the groups of synapses that receives correlated input (group 1 in this case) whereas those for the other correlated group (group 2) as well as those for the uncorrelated group (group 3) stay low (cf. similar results in (Toyoizumi et al., 2005a)). Neuron 2 develops strong weights to the other correlated group of synapses (group 2) whereas the efficacies of the second correlated group (group 1) remain low. The uncorrelated synapses of group 3 develop efficacies close to 0 as well, however, some of these synapses still increase their weight. Since uncorrelated inputs are likely to produce uncorrelated output spike trains and uncorrelated Poisson spike trains have a mutual information of 0, the tendency to produce mutual information independent output develops stronger weights to group 1 and group 3. However, the update rule also tries to maximise the mutual information between input and output, therefore the weights of the uncorrelated group 3 are weakened, and strong weights are developed solely for the correlated group 1.

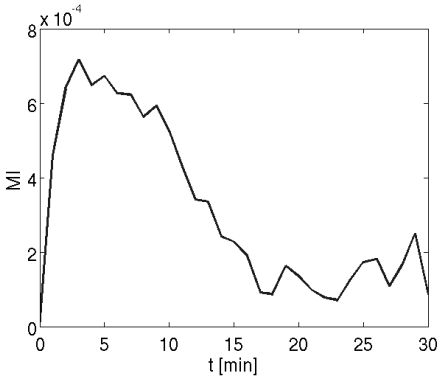


Figure 4.3.: Correlation experiment. Evolution of the average mutual information per time bin between both output spike trains as a function of time. Averages are calculated over segments of 1 minute.

Figure 4.2 (bottom) shows the average mutual information per time bin between input and output spike trains for both neurons, as well as the average Kullback-Leibler divergence per time bin. It can be seen that for both neurons the mutual information is maximized and the target output distribution of a constant firing rate of 30Hz is approached well. Figure 4.3 shows the evolution of the average mutual information per time bin between the output spike trains of the postsynaptic neurons 1 and 2. After an initial increase where the weights of both neurons start to grow simultaneously, the amount of information drops as both neurons develop strong efficacies to their specific parts of the input.

Figure 4.4 shows the final weight distribution for both postsynaptic neurons after 30 minutes of learning in 9 consecutive trials. Note that in four trials (numbers 2, 4, 5, and 8) the weights of neuron 1 are driven to the first and in the other five runs (numbers 1, 3, 6, 7, and 9) to the second correlated group of inputs, whereas the efficacies of the

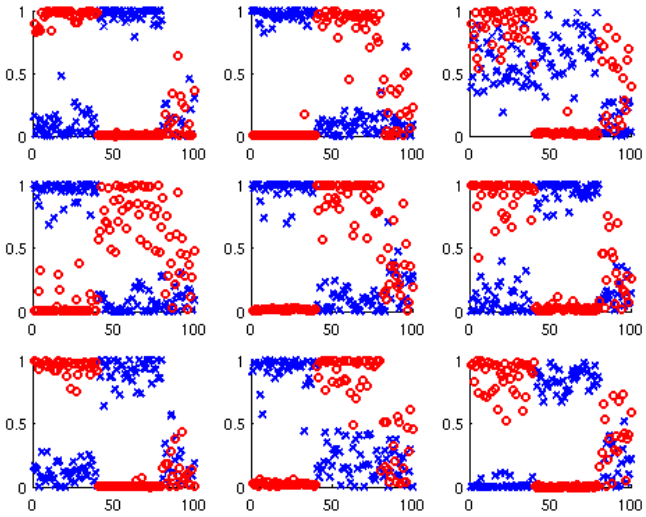


Figure 4.4.: Correlation experiment. Final weight distributions of both postsynaptic neurons (blue crosses: neuron 1, red circles: neuron 2) in 9 consecutive trials after 30 minutes of learning.

other correlated group remain low. In all cases the update rule of neuron 2 drives the weights to the other correlated group as to develop a mutual information independent output.

4.4.2. Time-Varying Correlations

In a second experiment again correlated input spike trains are considered, however, this time the correlations change over time. Again, 100 synapses receive Poisson input at the same rate of 20Hz and two correlation groups of 50 synapses each are established among the inputs such that spike trains from the same group are correlated, but spike trains from different groups are uncorrelated. More precisely, the 100 synapses are separated into 4 groups of 25 inputs each (group A, $1 \leq j \leq 25$, group B, $26 \leq j \leq 50$, group C, $51 \leq j \leq 75$, and group D, $76 \leq j \leq 100$). All groups are correlated with a coefficient of 0.5, but first, inputs to group A and B are uncorrelated with respect to group C and D. After 15 minutes correlations change such that A becomes correlated with C and B becomes correlated to D. After 35 minutes, finally A and D as well as B and C become correlated. Weights were initialized randomly between 0.10 and 0.12 for both neurons and could change between the bounds 0 and 1.

Figure 4.5 shows the evolution of weights and the time course of the mutual information and the Kullback-Leibler divergence for both neurons. It can be seen that neuron 1 always develops strong weights to one of the correlation groups with 50 synapses even

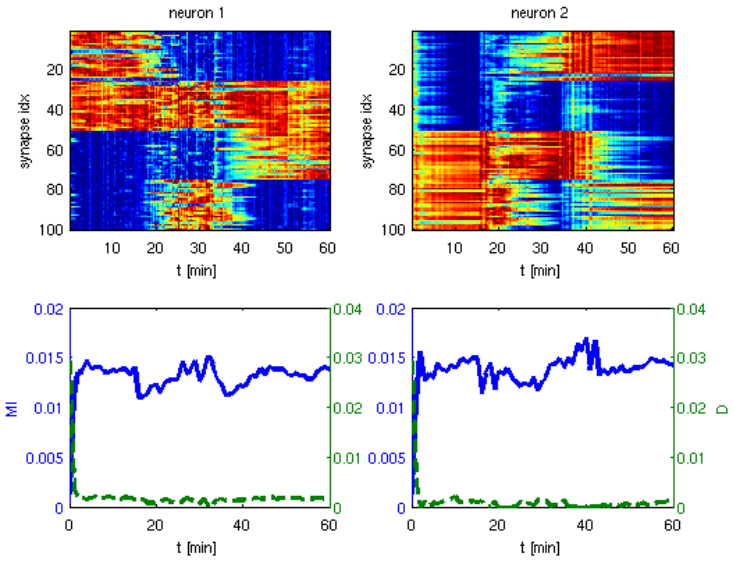


Figure 4.5.: Time-varying correlations. (Top) Evolution of weights during 60 minutes of learning for both postsynaptic neurons receiving Poisson input at 20Hz from 100 synapses. The 100 synapses are separated into 4 groups of 25 inputs each (group A, $1 \leq j \leq 25$, group B, $26 \leq j \leq 50$, group C, $51 \leq j \leq 75$, and group D, $76 \leq j \leq 100$). All groups are correlated with a coefficient of 0.5, but first, inputs to group A and B are uncorrelated with respect to group C and D. After 15 minutes correlations change such that A becomes correlated with C and B becomes correlated to D. After 35 minutes, finally A and D as well as B and C become correlated. (red: strong synapses, $w_j \approx 1$, blue: depressed synapses, $w_j \approx 0$.) Weights were initialized randomly between 0.10 and 0.12, $\alpha_2 = 10^{-6}$, $\gamma_1 = 0.1$, $\gamma_2 = 10$. (Bottom) Evolution of the average mutual information per time bin between input and output (blue solid line, left scale) and the Kullback-Leibler divergence per time bin for both neurons (green dashed line, right scale) as a function of time. Averages are calculated over segments of 1 minute.

when correlations change. Initially, these are groups A and B, after 15 minutes of learning, efficacies grow for groups B and D, and finally after 35 minutes, groups B and C. Note that the groups to which neuron 1 specializes may vary randomly from trial to trial (e.g., due to different initial weights), however, at each time 50 correlated synapses have high weights. Neuron 2, on the other hand, develops strong efficacies to the complementary group of inputs, whose synaptic weights for neuron 1 remained low, i.e., initially to groups C and D, after 15 minutes to A and C and after 35 minutes to A and D.

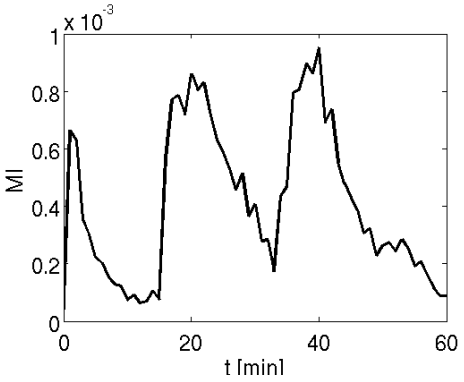


Figure 4.6.: Time-varying correlations. Evolution of the average mutual information per time bin between both output spike trains as a function of time. Averages are calculated over segments of 1 minute.

Figure 4.6 shows the evolution of the average mutual information per time bin between both output spike trains as a function of time. Whenever the correlations among the inputs change the statistical dependence between the outputs increases for a short time during which the neurons adapt to the new situation, but after a while the mutual information goes to zero again. This means that the learning rule not only drives the neurons' weights to different correlation groups, but also manages to keep the outputs independent if the correlations between the inputs change.

4.4.3. Rate Modulation Experiment

In another experiment, again Poisson input is presented to 100 synapses, however, the rate of the inputs to synapses 1 to 40 is modulated periodically with $r_0 + A\sin(2\pi t/T)$ ($r_0 = 20$Hz, $A = 10$Hz, $T = 100$ms); the rate of inputs to synapses 41 to 80 is modulated in the same way, but phase-shifted by 180 degrees ($r_0 + A\sin(2\pi t/T + \pi)$). Synapses 81 to 100 receive Poisson input at a constant rate of $r_0 = 20$Hz. Figure 4.7 (top) shows the evolution of weights for both neurons during 60 minutes of learning. Neuron 1 develops strong weights to one of the two groups with rate modulation. The synapses of neuron 2 develop high efficacies to the second rate modulated group of synapses. Figure 4.7 (bottom) shows the average mutual information per time bin between input and output

spike trains for both neurons, as well as the average Kullback-Leibler divergence per time bin; figure 4.8 shows the evolution of the average mutual information per time bin between the output spike trains of the postsynaptic neurons 1 and 2.

In this rate modulation paradigm the statistical depencence between the outputs is much smaller than in the experiments before, where correlations have been established among the inputs (cf. fig. 4.3 and 4.8). Even in this more difficult case the neurons' weights are driven to separate groups in the input.

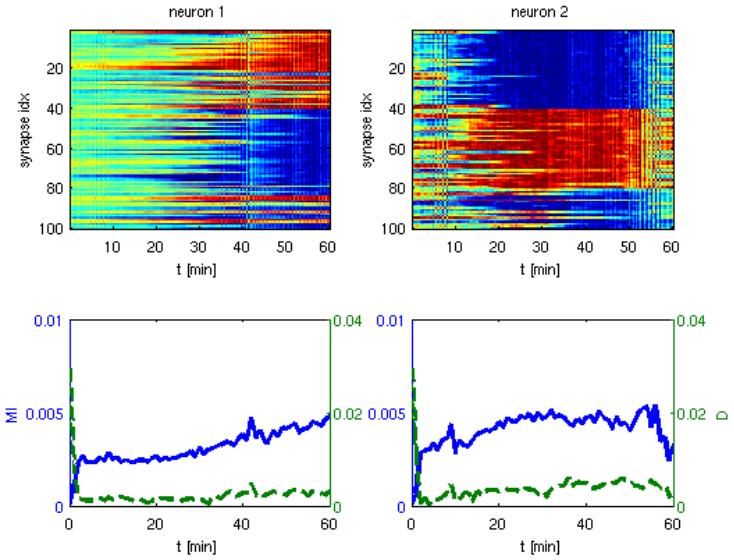


Figure 4.7.: Rate modulation experiment. (Top) Evolution of weights during 60 minutes of learning for both postsynaptic neurons receiving Poisson input from 100 synapses. The rate of inputs to synapses 1 to 40 was modulated periodically with $r_0 + A\sin(2\pi t/T)$ ($r_0 = 20$Hz, $A = 10$Hz, $T = 100$ms). The rate of inputs to synapses 41 to 80 was modulated in the same way, but phase-shifted by 180 degrees ($r_0 + A\sin(2\pi t/T + \pi)$). Synapses 81 to 100 received Poisson input at a constant rate of $r_0 = 20$Hz. (red: strong synapses, $w_j \approx 1$, blue: depressed synapses, $w_j \approx 0$.) Weights were initialized randomly between 0.10 and 0.12. $\alpha_2 = 10^{-5}$, $\gamma_1 = 0.1$, $\gamma_2 = 1$. (Bottom) Evolution of the average mutual information per time bin between input and output (blue solid line, left scale) and the Kullback-Leibler divergence per time bin for both neurons (green dashed line, right scale) as a function of time. Averages are calculated over segments of 1 minute.

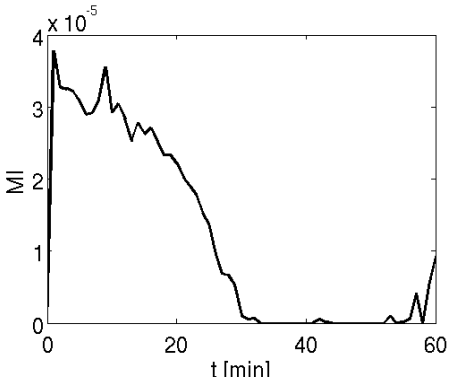


Figure 4.8.: Rate modulation experiment. Evolution of the average mutual information per time bin between input and output and the Kullback-Leibler divergence per time bin for both neurons as a function of time. Averages are calculated over segments of 1 min.

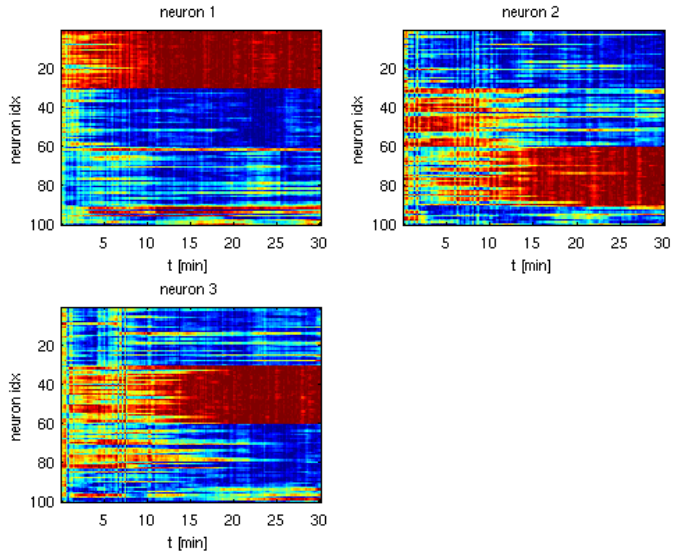


Figure 4.9.: More than two neurons. Evolution of weights during 30 minutes of learning for three postsynaptic neurons receiving the same Poisson input at 20Hz from 100 synapses. Inputs to synapses 1 to 30, 31 to 60 and 61 to 90 are each correlated with a coefficient of 0.5, but any two inputs belonging to different grups are uncorrelated. Synapses 91 to 100 receive uncorrelated Poisson input. (red: strong synapses, $w_j \approx 1$, blue: depressed synapses, $w_j \approx 0$.) Weights were initialized randomly between 0.10 and 0.12, $\alpha_i = 5 \cdot 10^{-6}$, $\gamma_1 = 0.03$, $\gamma_2 = 10$.

4.4.4. More Than Two Neurons

In the experiments so far we considered only two neurons that tried to keep their outputs statistically independent. It is natural to ask whether and how our learning rule can be extended to the case of more than two neurons. Looking at equations (4.34) and (4.35) we find that the weights w_{ij} of each neuron i are modified by terms C_{ij} and B_i^{post}; additionally, a term B_{ik}^{post} has to be considered for each pair of neurons i and k accounting for the mutual information between the outputs of these neurons. There are basically two possibilities, either we consider in the learning rule of each neuron the statistical dependencies on previous neurons, i.e.,

$$\frac{dw_{ij}(t)}{dt} = \alpha_i C_{ij}(t) \left[B_i^{post}(t-\delta) - \gamma_1 \sum_{k=1}^{i-1} B_{ki}^{post}(t-\delta) \right], \tag{4.36}$$

or on all other neurons, i.e.,

$$\frac{dw_{ij}(t)}{dt} = \alpha_i C_{ij}(t) \left[B_i^{post}(t-\delta) - \gamma_1 \sum_{k \neq i} B_{ki}^{post}(t-\delta) \right]. \tag{4.37}$$

The second version (4.37) looks more appealing because of symmetry reasons, however, experiments show that the first version (4.36) is more robust in the sense that the outputs of the neurons are more reliably statistically independent. This may be because the first neuron can maximize information transmission without influence from the firing behavior of the other neurons, the second neuron is influenced only by the the first neuron, and so on, whereas in (4.37) each neuron is influenced by all other neurons. Therefore the neurons sometimes may not be able to find a solution at all, especially if the number of neurons is large. On the other hand, version (4.36) needs different learning rates for every neuron.

In this experiment the latter version (4.37) is used, where each neuron receives information about the mutual information between its output and the output of all other neurons. Three postsynaptic neurons receive 100 inputs consisting of correlated Poisson spike trains at 20Hz each. This time the spike trains at the first 90 synapses were divided into three groups (group 1: spike trains 1 to 30, group 2: spike trains 31 to 60, and group 3: spike trains 61 to 90). Within each group a correlation coefficient of 0.5 was established, but spike trains from different groups were uncorrelated. The remaining 10 synapses received uncorrelated Poisson input (group 4). Weights were initialized randomly between 0.10 and 0.12 for all three neurons and could change between the bounds 0 and 1.

Figure 4.9 shows the evolution of weights for all three postsynaptic neurons. Each neuron develops strong weights to one of the correlated groups: neuron 1 to group 1, neuron 2 to group 3, and neuron 3 to group 2. One sees that although the synaptic efficacies of neuron 2 initially grow for the second group of inputs, they later change to group 3 as neuron 3 starts to drive its weights toward group 2. This demonstrates how the weight change of each neuron is influenced by the mutual information between its

output and the output of the other neurons. The strength of the uncorrelated group of synapses remained low for all three neurons, although some sporadically increased their efficacies. This is due to the fact that uncorrelated input does not induce statistical dependence between the outputs and that a constant target firing rate of 30Hz has to be maintained.

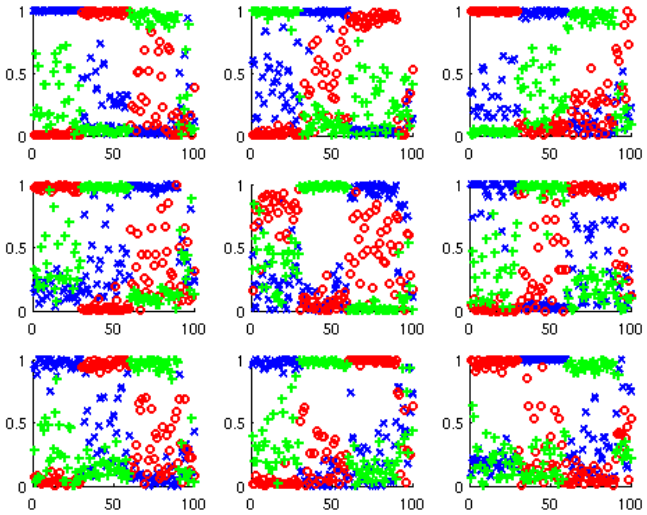


Figure 4.10.: More than two neurons. Final weight distributions of all three postsynaptic neurons (blue crosses: neuron 1, red circles: neuron 2, green plusses: neuron 3) in 9 consecutive trials after 30 minutes of learning.

Figure 4.10 shows the final weight distribution for all three postsynaptic neurons after 30 minutes of learning in 9 consecutive trials. Each neuron develops strong synaptic efficacies to a different correlation group in the inputs, even though the group number may vary from trial to trial. This means that via the synaptic update rule (4.37) for more than two neurons the three postsynaptic neurons in this case extract three independent components in the input.

5. Using Interneurons

In the previous chapter a learning rule was presented that optimizes the information transmission of two neurons, but at the same time keeps their outputs statistically independent. However, the rule for updating the weights of neuron 2 (4.29) is not local since it requires access to the firing behavior of neuron 1 (cf. (4.33)). In a more realistic setup neuron 2 would receive this information via synaptic connections from neuron 1.

In general, such connections are formed in the brain by (inhibitory) interneurons. These interneurons exhibit a large functional diversity, i.e., they may not only influence the membrane potential of the target neuron in the traditional additive way, but they are also able to affect the gain of the summated synaptic potentials, for example, or completely disable spiking of the postsynaptic neuron for a short period of time. Additionally there may be different time constants due to different neurotransmitters (e.g. $\mathrm{GABA_A}$ or $\mathrm{GABA_B}$) or different forms of inhibition like shunting inhibition, where due to strategic position of inhibitory synapses on the target neuron the postsynaptic membrane voltage can be clamped to the resting potential by "shunting" the excitatory input from large parts of the dendritic tree.

When using inhibitory interneurons there is a trade-off between the performance of the learning rule (or approximation of (4.29), respectively) and the complexity of the connection between the two neurons. In the simplest case we can assume that there is a single interneuron that synapses on neuron 2 and emits a spike whenever neuron 1 fires an action potential. This is a strong simplification since usually a single neuron is not sufficient to make an interneuron fire. Any form of inhibition between neuron 1 and neuron 2 has to implement or to approximate the effect of the term B_{12}^{post} (4.33), while the weights of both neurons evolve according to the learning rule (3.25) presented in section 3.2, which maximizes the mutual information between input and output (cf. figure 4.1).

First, a brief overview of types of interneurons in the brain and their functional properties is given. In section 5.2 a gain modulation mechanism is introduced by which the gain of neuron 2 is changed in order to implement the term sensitive to the statistical dependence between the outputs implicity in the postsynaptic term of the BCM rule. Then, section 5.3 presents some results of computer simulation experiments and in general discusses the use of interneurons for this task.

5.1. Interneurons of the Neocortex

The majority of neocortical neurons (about 70-80%) are excitatory pyramidal neurons, which have relatively stereotyped anatomical, physiological and molecular properties. The remaining 20-30% consist of interneurons, most of which are inhibitory. Their

characteristics are highly diverse (see (Toledo-Rodriguez et al., 2002; Markram et al., 2004) for reviews); and this daunting variety of inhibitory interneurons is currently one of the largest obstacle in the quest to fully understand the principles of neural circuits in the brain. However, interneurons also share some common features, most of which distinguish them from pyramidal neurons. Interneurons use GABA (γ-aminobutyric acid) as their neurotransmitter, and they can receive both excitatory and inhibitory synapses onto their somata.

Interneurons are usually classified by the domain of target cells on which their synapses are placed[1]. This selective innervation allows each type of interneuron to effect its target cell in a different specific way (Miles et al., 1996; Buhl et al., 1994). Typical target domains are

- the *(peri-)somatic region* of the target cells; in this case inhibitory interneurons affect the gain of the summated synaptic potentials and thereby the action potential discharge of target cells (Wang et al., 2002; Miles et al., 1996; Buhl et al., 1995). According to their appearance, these cells are usually named *basket cells*. These neurons are often involved in phasing and synchronizing neuronal activity (Cobb et al., 1995).

- the *dendrites* of the target cells; this type of inhibition influences dendritic processing and integration of synaptic inputs to influence synaptic plasticity. It also accounts for the "classical" form of inhibition where the membrane potential of the postsynaptic neuron is additively influenced. Interneurons targeting dendrites include bitufted cells, bipolar cells, double bouquet cells and neurogliaform cells, as well as Martinotti cells and neurons exclusive to layer I, such as the Cajal-Retzius cells.

- the *axon initial segment* of the target cells; this targeting places these interneurons in a powerful position to override all the complex dendritic integration and somatic gain settings by "editing" the neuron's action potential output. Due to their candlestick-like axonal terminals these neurons are called *chandelier cells* and affect the generation and timing of action potentials (Zhu et al., 2004).

5.2. Gain Modulation

Gain modulation has emerged in recent years as a general neural computational principle by which cortical neurons combine and process information (Salinas and Thier, 2000; Chance et al., 2002). The term refers to a change in the response amplitude of a neuron that is independent of its selectivity or receptive field characteristics. It is a nonlinear way to combine or integrate information from different sources of input, thereby giving rise to multiplicative interactions between neurons. Gain modulation has been found

[1] Different types of interneurons are also identified from their electrophysiological properties (i.e., according to their characteristic onset and steady-state response to a step current injection into the soma) and their molecular properties (Toledo-Rodriguez et al., 2002; Markram et al., 2004).

to be a useful for certain computations such as coordinate transformations or invariant object recognition (Salinas and Thier, 2000). However, the exact mechanisms by which it occurs are yet unknown; it has been found to be caused by background synaptic inputs (Chance et al., 2002) or by activation of GABA receptors, both tonically (Semyanov et al., 2004) or by inhibitory synapses close to the soma, usually via basket cells (Miles et al., 1996; Wang et al., 2002).

Gain modulation is not equivalent to the modification of neuronal responses by traditional additive excitation or inhibition. This distinction can be illustrated by looking at the firing rate of a neuron in resonse to an injected current or, equivalently, to the membrane potential (cf. figure 3.1(a)). Classical excitation (inhibition) would merely shift this curve to the left (right), whereas gain modulation leads to a change in the slope of the firing-rate curve, thereby corresponding to a multiplicative or divisive scaling, distinct from these additive or subractive shifts.

In our case we have to model the mutual information or the statistical dependence between the outputs of two neurons receiving the same input by modulating the gain $g(u_2(t))$ of neuron 2. Remember that the postsynaptic term of neuron 2 can be written as

$$B_2(t) = B_2^{post}(t) - \gamma_1 B_{12}^{post}(t), \tag{5.1}$$

where

$$\begin{aligned} B_2^{post}(t) = \delta(t - \hat{t}_2 - \delta) \log \left[\frac{g(u_2(t))}{\bar{g}_2(t)} \left(\frac{\tilde{g}}{\bar{g}_2(t)} \right)^{\gamma_2} \right] - \\ - R_2(t) \left[g(u_2(t)) - (1+\gamma_2)\bar{g}_2(t) + \gamma_2 \tilde{g} \right] \end{aligned} \tag{5.2}$$

and

$$\begin{aligned} B_{12}^{post}(t) = \delta(t - \hat{t}_2 - \delta) \left\{ \delta(t - \hat{t}_1 - \delta) \log \frac{\bar{g}_{12}(t)}{\bar{g}_1(t)\bar{g}_2(t)} - R_1(t) \left[\frac{\bar{g}_{12}(t)}{\bar{g}_2(t)} - \bar{g}_1(t) \right] \right\} - \\ - R_2(t) \left\{ \delta(t - \hat{t}_1 - \delta) \left[\frac{\bar{g}_{12}(t)}{\bar{g}_1(t)} - \bar{g}_2(t) \right] - R_1(t) \left[\bar{g}_{12}(t) - \bar{g}_1(t)\bar{g}_2(t) \right] \right\}. \end{aligned} \tag{5.3}$$

For a discrete time implementation with step size Δt, we can distinguish between 4 postsynaptic states for both neurons in each time step k: one where both are spiking, one where neither of them emits a spike and two cases where only one of them fires. For these cases (denoted as two binary variables y_1^k, y_2^k) the postsynaptic term evaluates to

- $y_1^k = y_2^k = 1$:

$$B_2(t) = \frac{1}{\Delta t} \log \left[\frac{g(u_2(t))}{\bar{g}_2(t)} \left(\frac{\tilde{g}}{\bar{g}_2(t)} \right)^{\gamma_2} \right] - \frac{1}{(\Delta t)^2} \gamma_1 \log \frac{\bar{g}_{12}(t)}{\bar{g}_1(t)\bar{g}_2(t)}, \tag{5.4}$$

- $y_1^k = 0$, $y_2^k = 1$:

$$B_2(t) = \frac{1}{\Delta t} \log \left[\frac{g(u_2(t))}{\bar{g}_2(t)} \left(\frac{\tilde{g}}{\bar{g}_2(t)} \right)^{\gamma_2} \right] + \frac{1}{\Delta t} \gamma_1 R_1(t) \left[\frac{\bar{g}_{12}(t)}{\bar{g}_2(t)} - \bar{g}_1(t) \right], \tag{5.5}$$

- $y_1^k = 1$, $y_2^k = 0$:

$$B_2(t) = -R_2(t)\left[g(u_2(t)) - (1+\gamma_2)\bar{g}_2(t) + \gamma_2\tilde{g}\right] + \frac{1}{\Delta t}\gamma_1 R_2(t)\left[\frac{\bar{g}_{12}(t)}{\bar{g}_1(t)} - \bar{g}_2(t)\right], \tag{5.6}$$

- $y_1^k = y_2^k = 0$:

$$B_2(t) = -R_2(t)\left[g(u_2(t)) - (1+\gamma_2)\bar{g}_2(t) + \gamma_2\tilde{g}\right] - \gamma_1 R_1(t)R_2(t)\left[\bar{g}_{12}(t) - \bar{g}_1(t)\bar{g}_2(t)\right], \tag{5.7}$$

where for simplicity t stands for $t^k = k\Delta t$.

We want to model the contribution of the term expressing the mutual information between the outputs (5.3) by modulating the gain $g(u_2(t))$. That is, we again apply the simple postsynaptic BCM-term

$$\begin{aligned} B_2^{post}(t) = \delta(t - \hat{t}_2 - \delta)\log\left[\frac{g_2'(t)}{\bar{g}_2(t)}\left(\frac{\tilde{g}}{\bar{g}_2(t)}\right)^{\gamma_2}\right] - \\ - R_2(t)\left[g_2'(t) - (1+\gamma_2)\bar{g}_2(t) + \gamma_2\tilde{g}\right] \end{aligned} \tag{5.8}$$

to neuron 2, but try to encapsulate the effect of making the output independent of neuron 1 in changing the gain $g(u_2(t))$ into $g_2'(t)$.

First, we try to find arithmetic expressions for $g_2'(t)$ by comparing formula (5.8) with equations (5.4) to (5.7). Then we get

- $y_1^{k-1} = y_2^{k-1} = 1$:

$$g_2'(t) = g(u_2(t))\left(\frac{\bar{g}_1(t)\bar{g}_2(t)}{\bar{g}_{12}(t)}\right)^{G_{inh}}, \tag{5.9}$$

- $y_1^{k-1} = 0$, $y_2^{k-1} = 1$:

$$g_2'(t) = g(u_2(t))\exp\left[R_1(t)\gamma_1\left(\frac{\bar{g}_{12}(t)}{\bar{g}_2(t)} - \bar{g}_1(t)\right)\right], \tag{5.10}$$

- $y_1^{k-1} = 1$, $y_2^{k-1} = 0$:

$$g_2'(t) = g(u_2(t)) - G_{inh}\left[\frac{\bar{g}_{12}(t)}{\bar{g}_1(t)} - \bar{g}_2(t)\right], \tag{5.11}$$

- $y_1^{k-1} = y_2^{k-1} = 0$:

$$g_2'(t) = g(u_2(t)) + R_1(t)\gamma_1\left[\bar{g}_{12}(t) - \bar{g}_1(t)\bar{g}_2(t)\right], \tag{5.12}$$

where G_{inh} is a dimensionless constant proportional to γ_1.

The gain of neuron 2 is changed multiplicatively if it itself has spiked in the previous time step (denoted by $y_2^{k-1} = 1$) and additively otherwise ($y_2^{k-1} = 0$). Remember that the running average $\bar{g}_{12}(t)$ estimates the joint postsynaptic spiking probability

$P(y_1^k, y_2^k | Y_1^{k-1}, Y_2^{k-1})$, whereas the product $\bar{g}_1(t)\bar{g}_2(t)$ is a measure for the independent spiking probability $P(y_1^k | Y_1^{k-1}) P(y_2^k | Y_2^{k-1})$. Usually, $\bar{g}_{12}(t)$ is larger than $\bar{g}_1(t)\bar{g}_2(t)$ and if the outputs are statistically independent, these values would be approximately equal. Since $\bar{g}_1(t)$ and $\bar{g}_2(t)$ will approach the target firing rate $\tilde{g}$ due to the BCM-rule, the update rule will try to push $\bar{g}_{12}(t)$ towards the product $\bar{g}_1(t)\bar{g}_2(t)$.

In the first case (5.9) the gain is in general multiplicatively decrased by a factor measuring how far the current postsynaptic activity is away from statistical independence. This means that simultaneous activity of both neurons is punished if the neurons' activity has recently been statistically dependent (e.g., if they were both highly active before). When on the other hand only neuron 1 has spiked in the previous time step (5.11) the gain is inhibited by an additive amount. The cases where neuron 1 is not spiking, i.e., (5.10) and (5.12), would result in an multiplicative or additive increase of the gain. Note that these values depend not only on the current firing rates, but also on the refractory state of neuron 1, that is, it differs between whether neuron 1 did not spike because of a low firing rate or because of refractoriness.

However, figure 4.1 on page 34 suggests that significant effects are encountered only when one of the two neurons is firing; also the influence of simultaneous action potentials within the same time step can be neglected as Δt gets small. Therefore we focus only on cases (5.10) and (5.11) representing a multiplicative increase or an additive decrease of the gain. Incorporating an exponential decay with time constant τ_g, we define two functions

$$k_{g_2}(t) = \prod_f \left[1 + \left(G_0(t_2^{(f)}) - 1 \right) \exp\left(\frac{t - t_2^{(f)}}{\tau_g} \right) \right], \tag{5.13}$$

$$d_{g_2}(t) = -\sum_f G_1(t_1^{(f)}) \exp\left(\frac{t - t_1^{(f)}}{\tau_g} \right), \tag{5.14}$$

with

$$G_0(t) = \exp\left[\gamma_1 R_1(t) \left(\frac{\bar{g}_{12}(t)}{\bar{g}_2(t)} - \bar{g}_1(t) \right) \right], \tag{5.15}$$

$$G_1(t) = \min\left\{ g(u_2(t)), G_{inh}\left[\frac{\bar{g}_{12}(t)}{\bar{g}_1(t)} - \bar{g}_2(t) \right] \right\}, \tag{5.16}$$

such that the gain is modulated according to

$$g_2'(t) = g(u_2(t)) \cdot k_{g_2}(t) + d_{g_2}(t). \tag{5.17}$$

The term $k_{g_2}(t)$ models the multiplicative gain increase which occurs at spike times of neuron 2 ($t_2^{(f)}$); in the absence of action potentials it decays back to 1. The term $d_{g_2}(t)$ accounts for the additive decrease of the gain at postsynaptic spike times of neuron 1. Note that the gain cannot take negative values, therefore it is decreased at most by its current value $g(u_2(t))$. Between spikes this inhibition term decays back to 0. Figure 5.1 shows both of these terms as a function of time during 1 second for a sample of

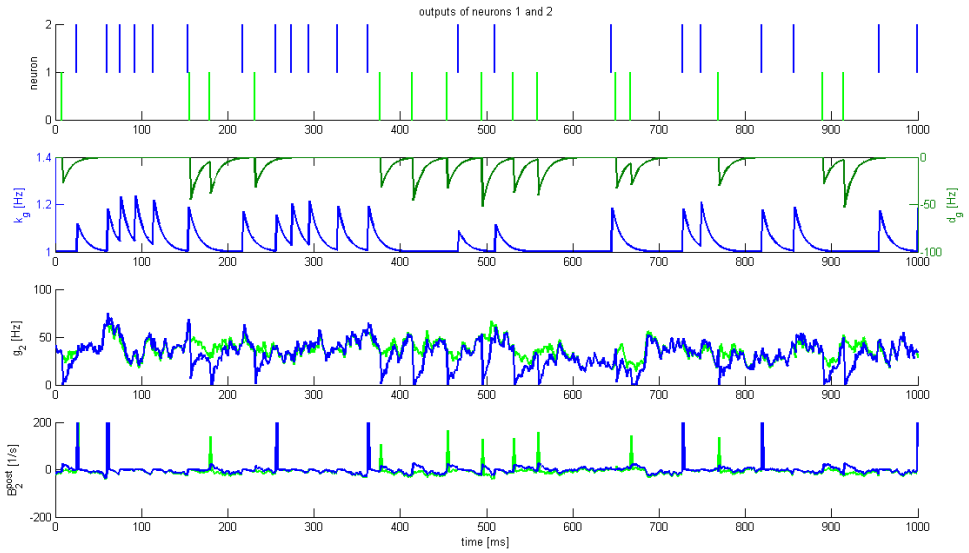


Figure 5.1.: Visualization of the gain modulation mechanism during 1 second. From top to bottom: the output spike trains of neuron 1 (green) and neuron 2 (blue), the terms k_{g_2} (blue, left scale) and d_{g_2} (green, right scale), the original gain $g(u_2)$ (green) and the modulated gain g_2' (blue), and the postsynaptic term B_2^{post} without (green) and with (blue) gain modulation as a function of time ($\tau_g = 10$ms).

postsynaptic spike trains of both neurons, as well as the gain g_2 and the postsynaptic term B_2^{post} both with and without gain modulation.

One sees that whenever neuron 1 emits a spike the gain of neuron 2 is decreased; the firing probability is almost 0 for a short period of time. On the other hand, if neuron 2 fires an action potential itself its gain is slightly increased. Thus repetitive firing of neuron 2 is more enforced if these spikes are not accompanied by spikes from neuron 1. Considering the time course of the postsynaptic term one finds that the peaks caused by firings of neuron 2 are approximated well due to the multiplicative gain modulation, whereas peaks caused by action potentials of neuron 1 cannot be fully modelled since it is not possible to make the gain negative. However, this effect is still accounted for by the fact that the postsynaptic term is larger for a short period of time afterwards which is given by the time constant of the additive gain decrease.

5.3. Results

The effect of the term B_{12}^{post} measuring the statistical dependence between both outputs has now been implemented by a gain modulation mechanism, such that the synaptic weights of both neurons simply evolve according to the generalized BCM rule for spiking neurons. First, it has to be said that due to the time course of this term (cf. figure 4.1), which has sharp peaks at postsynaptic spike times, it is rather hard to model with synaptic connections that have certain time constants, etc. Furthermore, it is necessary to distinguish between spikes of neuron 1 and spikes of neuron 2 (or between spikes of the target neuron and those of other neurons, respectively), that is, probably more than one interneuron is needed. Third, inhibition alone, in the sense that firing of the target neuron is made more difficult, is not enough, it is also necessary to enforce the emission of action potentials.

Despite these difficulties I have derived a gain modulation mechanism from purely theoretical considerations how the gain has to be changed in order to achieve the desired effect. The result suggests that the gain of neuron 2 (i.e., the firing propability as a function of the membrane potential) should be additively decreased in the case of spikes of neuron 1, and multiplicatively increased when neuron 2 fires itself. However, it has yet to be discussed how this approach may in detail be implemented by one or more interneurons. Gain changes can usually be implemented by basket cells, which synapse at regions on or near the soma. The inhibitory effect usually results in a firing probabilty close to 0 for a short period of time and therefore might also be realized by fully prohibiting action potential generation during that time, e.g., by chandelier cells and by innervation of the axon initial segment.

Gain changes have to occur during a short period of time since they presumably carry information about spike times. Therefore smaller time constants ($GABA_A$) are favored; in the simulations a time constant of 10ms has been used. Longer time constants of about 100ms ($GABA_B$) have proven to be not so suitable. Attempts to change the membrane potential, either in the traditional additive way, or by shunting inhibition, have not been fruitful either.

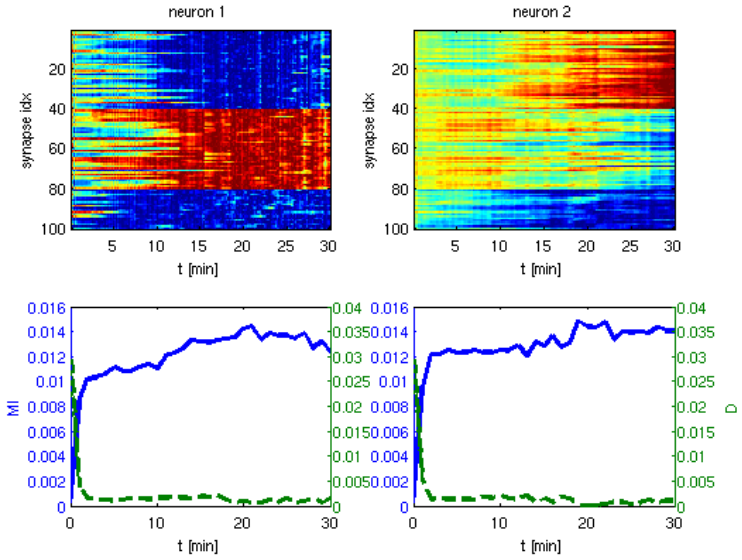


Figure 5.2.: Correlation experiment using gain modulation. (Top) Evolution of weights during 30 minutes of learning for both postsynaptic neurons receiving Poisson input at 20Hz from 100 synapses. Inputs to synapses 1 to 40 and 41 to 80 are both correlated with a coefficient of 0.5, but any two inputs belonging to different grups are uncorrelated. Synapses 81 to 100 receive uncorrelated Poisson input. (red: strong synapses, $w_j \approx 1$, blue: depressed synapses, $w_j \approx 0$.) Weights were initialized randomly between 0.10 and 0.12, $\alpha_2 = 3 \cdot 10^{-6}$, $\gamma_1 = 0.01$, $\gamma_2 = 10$. (Bottom) Evolution of the average mutual information per time bin between input and output (blue solid line, left scale) and the Kullback-Leibler divergence per time bin for both neurons (green dashed line, right scale) as a function of time. Averages are calculated over segments of 1 minute.

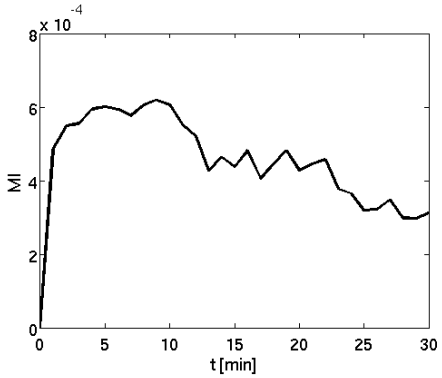


Figure 5.3.: Correlation experiment using gain modulation. Evolution of the average mutual information per time bin between both output spike trains as a function of time. Averages are calculated over segments of 1 minute.

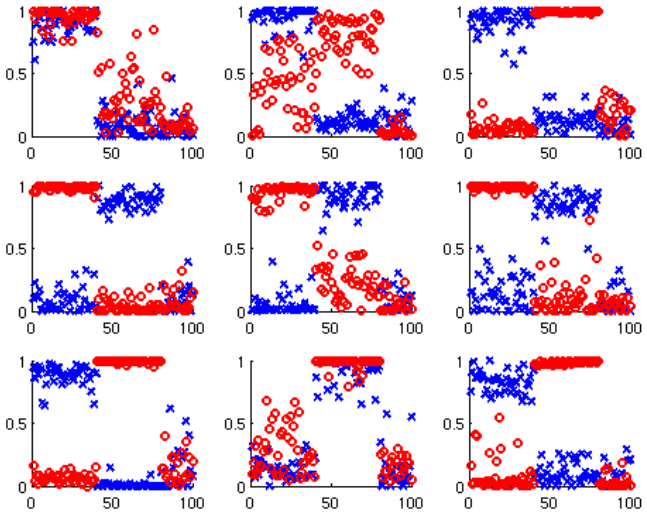


Figure 5.4.: Correlation experiment using gain modulation. Final weight distributions of both postsynaptic neurons (blue crosses: neuron 1, red circles: neuron 2) in 9 consecutive trials after 30 minutes of learning.

This approach is tested by the correlation experiment of section 4.4.1, where the input consisted of two correlated groups, but spike trains from different groups were uncorrelated (cf. figures 4.2, 4.3, and 4.4). Figures 5.2 and 5.3 show the evolution of weights and information transmission for both neurons, and the evolution of mutual information between the outputs, respectively. Even though the weights of neuron 2 start growing for the same correlation group in the input as to which neuron 1 specializes, it finally develops strong efficacies to the other group. However, in 2 of 9 consecutive trials (trial 1 and 8 in figure 5.4) the neurons failed to specialize for different correlation groups.

6. Conclusion

In this thesis a learning rule for spiking neurons is derived that extracts statistically independent components from an ensemble of input spike trains. The approach is based on a recent result proposing a generalized BCM rule for spiking neurons that maximizes information transmission between the input and the output of a stochastically spiking neuron model (Toyoizumi et al., 2005a). There a synaptic update rule has been derived which maximizes the mutual information between input and output spike trains under the constraint of a constant target firing rate and which exhibits all the basic features of the BCM model (Bienenstock et al., 1982), namely regimes of LTP and LTD separated by a sliding threshold on the postsynaptic activity. In this thesis this idea is extended in a way that a second neuron, which receives the same input, also maximizes information transmission, but at the same time tries to keep its output statistically independent to the output of the first neuron.

Optimization under these constraints yields a plasticity rule similar to the generalized BCM rule for spiking neurons proposed in (Toyoizumi et al., 2005a), however, an additional term is included that is sensitive to the momentary statistical dependence between the outputs of both neurons. This term depends on the recent firing history of both neurons; more precisely, it compares the average product of firing rates of both neurons (accounting for the output joint probability) with the product of their average firing rates (estimating the independent output distribution). Thereby, it minimizes the mutual information between the output spike trains. The learning rule is tested in several computer simulation experiments, and it is also suggested how it may be extended to the case of more neurons.

However, this requires information about the firing behavior of one neuron to be non-locally available at the site of the other neuron. In a biologically more realistic setup, this neuron would receive this information via (usually inhibitory) synaptic connections. Information about the firing behavior of both neurons could at least in principle be available for an interneuron connecting them. For simplicity, it is assumed that there is an interneuron that fires whenever its presynaptic neuron fires. In reality, one neuron alone is not sufficient to make an interneuron fire, therefore one could, for instance, replace the neurons by populations of clones of neurons.

Any interneuron has to implement the effect of the term B_{12}^{post} sensitive to the momentary statistical dependence between the outputs. It turns out that it is necessary to distinguish between spikes of both neurons, i.e., spikes of neuron 1 have a different effect than spikes of neuron 2. Furthermore, firing of the target neuron has to be both weakened and encouraged, that is, both inhibition and excitation are needed. Here, the approximation is achieved by a gain modulation mechanism derived from theoretical considerations how the gain function should be changed to have the desired effect. How-

ever, it has yet to be discussed how this approach may in detail be implemented by one or more interneurons. It should be also possible to extend this approach to the case of more than two neurons (e.g., analogously to the result suggested in section 4.4.4), but there has been no time to investigate this idea.

The proposed learning rule is not perfect; for example, there are a lot of parameters (α_i, γ_1, γ_2) that require a relatively fine tuning in order to get some reasonable results. For example, if the learning rate α is too high the weights do not converge to a stable solution, but start to oscillate. The same applies for the case if γ_1 is too large; if it is too low, however, the statistical dependence between the outputs has no effect. Especially in the experiments of chapter 5 I have invested a lot of time searching for good parameters. Furthermore, the learning rule has yet to be tested in more complex experiments. Simulations have been performed in MATLAB on a standard personal computer.

The learning rule presented in this thesis provides an idea for a first step towards an application of (nonlinear) independent component analysis (ICA) (Hyvärinen and Oja, 2000) to the case of spiking neurons. It seems to be an important mechanism of neural systems in the brain to extract statistically independent features for several reasons. First, it can be shown that ICA and related approaches can reproduce many properties of cells in the brain, especially in visual cortex (Hyvärinen et al., 2005). Second, ICA produces an efficient coding scheme in the sense that it finds sparse representations of the input data, supporting the idea that one of the main aims of stimulus processing is the reduction of redundancy (Barlow, 1961; Barlow, 1989). Finally, ICA provides a generative model and tries to explain how the observed signals are constructed. While I do not propose that independent component analysis in the brain works via the suggested learning rule, it nevertheless shows some interesting properties and it might provide some ideas for future work.

A. Derivation of the Learning Rule

The quantity we want to maximize is

$$L = I(\mathbf{X}^K, \mathbf{Y}_2^K) - \gamma_1 I'(\mathbf{Y}_1^K, \mathbf{Y}_2^K) - \gamma_2 D(P(Y_2^K)||\tilde{P}(Y_2^K)), \tag{A.1}$$

where the random variables $\mathbf{X}^K$, $\mathbf{Y}_1^K$ and $\mathbf{Y}_2^K$ describe the input spike trains and the output spike trains of neuron 1 and 2, respectively, of length $K \cdot \Delta t$. $I(\mathbf{X}^K, \mathbf{Y}_2^K)$ is the mutual information between the input spike trains and the output spike train of neuron 2, and $I'(\mathbf{Y}_1^K, \mathbf{Y}_2^K) = I(\mathbf{Y}_1^K, \mathbf{Y}_2^K)/(K \cdot \Delta t)$ is the information rate between the output spike trains of both neurons. The Kullback-Leibler divergence $D(P(Y_2^K)||\tilde{P}(Y_2^K))$ measures the distance between the current distribution $P(Y_2^K)$ over all possible output spike trains from some target distribution $\tilde{P}(Y_2^K)$. γ_1 and γ_2 are weighting constants, where γ_2 is dimensionless and γ_1 has dimension s.

The mutual information $I(\mathbf{X}, \mathbf{Y})$ between two random variables $\mathbf{X}$ and $\mathbf{Y}$ is defined by

$$\begin{aligned} I(\mathbf{X}, \mathbf{Y}) &= \sum_{X,Y} P(X,Y) \log \frac{P(Y|X)}{P(Y)} \\ &= \sum_{X,Y} P(X,Y) \log \frac{P(X,Y)}{P(X)P(Y)}, \end{aligned} \tag{A.2}$$

where the sum runs over all instances X and Y of the random variables $\mathbf{X}$ and $\mathbf{Y}$. The Kullback-Leibler divergence $D(P(Y)||\tilde{P}(Y))$ between two distributions $P(Y)$ and $\tilde{P}(Y)$ of the same random variable $\mathbf{Y}$ is given by the expression

$$D(P(Y)||\tilde{P}(Y)) = \sum_{Y} P(Y) \log \frac{P(Y)}{\tilde{P}(Y)}, \tag{A.3}$$

where the sum runs over all instances Y of $\mathbf{Y}$.

Note that the expressions for mutual information and Kullback-Leibler divergence are both expectation values, therefore we can write

$$L = \left\langle \log \frac{P(Y_2^K|X^K)}{P(Y_2^K)} - \frac{\gamma_1}{K\Delta t} \log \frac{P(Y_1^K, Y_2^K)}{P(Y_1^K)P(Y_2^K)} - \gamma_2 \log \frac{P(Y_2^K)}{\tilde{P}(Y_2^K)} \right\rangle_{\mathbf{X}^K, \mathbf{Y}_1^K, \mathbf{Y}_2^K}. \tag{A.4}$$

Remember from chapter 4 how the following probabilities are defined (for $i = 1, 2$):

$$P(Y_i^K|X^K) = \prod_{k=1}^{K} P(y_i^k|Y_i^{k-1}, X^k),$$
$$P(Y_i^K) = \prod_{k=1}^{K} P(y_i^k|Y_i^{k-1}),$$
$$P(Y_1^K, Y_2^K) = \prod_{k=1}^{K} P(y_1^k, y_2^k|Y_1^{k-1}, Y_2^{k-1}),$$
$$\tilde{P}(Y_i^K) = \prod_{k=1}^{K} \tilde{P}(y_i^k|Y_i^{k-1}).$$

With these expressions we can write L as $L = \sum_{k=1}^{K} \Delta L^k$ with

$$\Delta L^k = \left\langle \log \frac{P(y_2^k|Y_2^{k-1}, X^k)}{P(y_2^k|Y_2^{k-1})} - \frac{\gamma_1}{\Delta t} \log \frac{P(y_1^k, y_2^k|Y_1^{k-1}, Y_2^{k-1})}{P(y_1^k|Y_1^{k-1})P(y_2^k|Y_2^{k-1})} - \gamma_2 \log \frac{P(y_2^k|Y_2^{k-1})}{\tilde{P}(y_2^k|Y_2^{k-1})} \right\rangle_{\mathbf{X}^k, \mathbf{Y}_1^k, \mathbf{Y}_2^k}. \quad \text{(A.5)}$$

Assuming slow changes of synaptic weights, a gradient ascent algorithm is applied to maximize the objective function (A.1) and the weight w_{2j} of neuron 2 is changed at each time step by

$$\Delta w_{2j}^k = \alpha \frac{\partial \Delta L^k}{\partial w_{2j}}, \quad \text{(A.6)}$$

with an appropriate learning rate $\alpha > 0$.

A.1. Evaluation of the Gradient

To evaluate the gradient we have to calculate the partial derivative of (A.5) with respect to w_{2j}. This expression contains several terms which are functions of the input spike trains X^k and the output spike trains Y_1^k and Y_2^k.

The average of an arbitrary function f_w with arguments x, y_1 and y_2 is by definition

$$\langle f_w(x, y_1, y_2) \rangle_{\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2} = \sum_{x, y_1, y_2} p_w(x, y_1, y_2) f_w(x, y_1, y_2), \quad \text{(A.7)}$$

where $p_w(x, y_1, y_2)$ denotes the joint probability of the triple (x, y_1, y_2) to occur. The sum runs over all configurations of x, y_1 and y_2 and the subscript w indicates that both the probability distribution p_w and the function f_w may depend on an additional parameter w.

We have $p_w(x, y_1, y_2) = p(x)p(y_1|x)p_w(y_2|x)$[1], where $p(x)$ is a given input distribution and $p(y_1|x)$ and $p_w(y_2|x)$ the conditional probabilities of generating outputs y_1 and y_2 given the input x. Note that since we take the derivative with respect to weights of neuron 2 only, $p_w(y_2|x)$ depends on the additional parameter w whereas $p(y_1|x)$ does not. Hence, (A.7) can be transformed into:

$$\begin{aligned}\langle f_w(x, y_1, y_2)\rangle_{\mathbf{x},\mathbf{y}_1,\mathbf{y}_2} &= \sum_{x,y_1,y_2} p(x)p(y_1|x)p_w(y_2|x)f_w(x, y_1, y_2) \\ &= \sum_{x,y_1} p(x, y_1) \sum_{y_2} p_w(y_2|x)f_w(x, y_1, y_2) \\ &= \left\langle \sum_{y_2} p_w(y_2|x)f_w(x, y_1, y_2) \right\rangle_{\mathbf{x},\mathbf{y}_1} .\end{aligned} \tag{A.8}$$

Taking the derivative with respect to w, the product rule yields two terms,

$$\begin{aligned}\frac{\partial}{\partial w} \langle f_w(x, y_1, y_2)\rangle_{\mathbf{x},\mathbf{y}_1,\mathbf{y}_2} = &\left\langle \sum_{y_2} p_w(y_2|x) \frac{\partial}{\partial w} f_w(x, y_1, y_2) \right\rangle_{\mathbf{x},\mathbf{y}_1} \\ &+ \left\langle \sum_{y_2} \frac{\partial}{\partial w} p_w(y_2|x)f_w(x, y_1, y_2) \right\rangle_{\mathbf{x},\mathbf{y}_1} ,\end{aligned} \tag{A.9}$$

where the first term contains the derivative of the function f_w and the second term contains the derivative of the conditional probability p_w. Since

$$\frac{\partial}{\partial w} p_w(y_2|x) = p_w(y_2|x) \frac{\partial}{\partial w} \log p_w(y_2|x), \tag{A.10}$$

the right-hand side of (A.9) evaluates to

$$\left\langle \frac{\partial}{\partial w} f_w(x, y_1, y_2) \right\rangle_{\mathbf{x},\mathbf{y}_1,\mathbf{y}_2} + \left\langle \left[\frac{\partial}{\partial w} \log p_w(y_2|x) \right] f_w(x, y_1, y_2) \right\rangle_{\mathbf{x},\mathbf{y}_1,\mathbf{y}_2} , \tag{A.11}$$

i.e., it can be written as an average over the joint distribution of x, y_1 and y_2.

Now we can evaluate each of the terms of (A.5) using (A.11). Considering the term $\frac{\partial}{\partial w_{2j}} \left\langle \log P(y_1^k, y_2^k|Y_1^{k-1}, Y_2^{k-1}) \right\rangle_{\mathbf{X}^k,\mathbf{Y}_1^k,\mathbf{Y}_2^k}$ first, we get

$$\begin{aligned}&\left\langle \frac{\partial}{\partial w_{2j}} \log P(y_1^k, y_2^k|Y_1^{k-1}, Y_2^{k-1}) \right\rangle_{\mathbf{X}^k,\mathbf{Y}_1^k,\mathbf{Y}_2^k} \\ &\qquad + \left\langle \left[\frac{\partial}{\partial w_{2j}} \log P(Y_2^k|X^k) \right] \log P(y_1^k, y_2^k|Y_1^{k-1}, Y_2^{k-1}) \right\rangle_{\mathbf{X}^k,\mathbf{Y}_1^k,\mathbf{Y}_2^k} .\end{aligned} \tag{A.12}$$

[1] since the outputs y_1 and y_2 are independent given the input x

We find that the first term of (A.12) vanishes because

$$\begin{aligned}
&\left\langle \frac{\partial}{\partial w_{2j}} \log P(y_1^k, y_2^k | Y_1^{k-1}, Y_2^{k-1}) \right\rangle_{\mathbf{X}^k, \mathbf{Y}_1^k, \mathbf{Y}_2^k} = \\
&= \left\langle \left\langle \frac{\partial}{\partial w_{2j}} \log P(y_1^k, y_2^k | Y_1^{k-1}, Y_2^{k-1}) \right\rangle_{\mathbf{y}_1^k, \mathbf{y}_2^k | Y_1^{k-1}, Y_2^{k-1}} \right\rangle_{\mathbf{Y}_1^{k-1}, \mathbf{Y}_2^{k-1}} \\
&= \left\langle \sum_{y_1^k, y_2^k} \left[\frac{\partial}{\partial w_{2j}} \log P(y_1^k, y_2^k | Y_1^{k-1}, Y_2^{k-1}) \right] P(y_1^k, y_2^k | Y_1^{k-1}, Y_2^{k-1}) \right\rangle_{\mathbf{Y}_1^{k-1}, \mathbf{Y}_2^{k-1}} \\
&= \left\langle \frac{\partial}{\partial w_{2j}} \left[\sum_{y_1^k, y_2^k} P(y_1^k, y_2^k | Y_1^{k-1}, Y_2^{k-1}) \right] \right\rangle_{\mathbf{Y}_1^{k-1}, \mathbf{Y}_2^{k-1}} = 0. \qquad \text{(A.13)}
\end{aligned}$$

In the second line of (A.13) we drop the expectation over $\mathbf{X}^k$ since the argument of the expectation operator is independent of the input spike train X^k and use the identity $\langle \cdot \rangle_{\mathbf{Y}_1^k, \mathbf{Y}_2^k} = \left\langle \langle \cdot \rangle_{\mathbf{y}_1^k, \mathbf{y}_2^k | Y_1^{k-1}, Y_2^{k-1}} \right\rangle_{\mathbf{Y}_1^{k-1}, \mathbf{Y}_2^{k-1}}$.

With the same argument it can be shown that

$$\left\langle \frac{\partial}{\partial w_{2j}} \log P(y_i^k | Y_i^{k-1}) \right\rangle_{\mathbf{X}^k, \mathbf{Y}_1^k, \mathbf{Y}_2^k} = \left\langle \frac{\partial}{\partial w_{2j}} \log P(y_i^k | Y_i^{k-1}, X^k) \right\rangle_{\mathbf{X}^k, \mathbf{Y}_1^k, \mathbf{Y}_2^k} = 0 \quad \text{(A.14)}$$

for $i = 1, 2$ (see also (Toyoizumi et al., 2005b)), and $\tilde{P}(y_i^k | Y_i^{k-1})$ is by definition independent of w_{2j}. Hence, the only term that gives a nontrivial contribution in (A.12) is the second one. With an analogous evaluation for the other terms in (A.5) we finally have

$$\begin{aligned}
\frac{\partial}{\partial w_{2j}} \Delta L^k = \Bigg\langle \left[\frac{\partial}{\partial w_{2j}} \log P(Y_2^k | X^k) \right] \Bigg(&\log \frac{P(y_2^k | Y_2^{k-1}, X^k)}{P(y_2^k | Y_2^{k-1})} \\
&- \frac{\gamma_1}{\Delta t} \log \frac{P(y_1^k, y_2^k | Y_1^{k-1}, Y_2^{k-1})}{P(y_1^k | Y_1^{k-1}) P(y_2^k | Y_2^{k-1})} - \gamma_2 \log \frac{P(y_2^k | Y_2^{k-1})}{\tilde{P}(y_2^k | Y_2^{k-1})} \Bigg) \Bigg\rangle_{\mathbf{X}^k, \mathbf{Y}_1^k, \mathbf{Y}_2^k}. \qquad \text{(A.15)}
\end{aligned}$$

From chapter 4 we know that

$$\begin{aligned}
P(y_i^k | Y_i^{k-1}, X^k) &= (\rho_i^k)^{y_i^k} (1 - \rho_i^k)^{(1-y_i^k)}, \\
P(y_i^k | Y_i^{k-1}) &= (\bar{\rho}_i^k)^{y_i^k} (1 - \bar{\rho}_i^k)^{(1-y_i^k)}, \\
\tilde{P}(y_i^k | Y_i^{k-1}) &= (\tilde{\rho}_i^k)^{y_i^k} (1 - \tilde{\rho}_i^k)^{(1-y_i^k)}, \\
P(y_1^k, y_2^k | Y_1^{k-1}, Y_2^{k-1}) &= (\bar{\bar{\rho}}_{12}^k)^{y_1^k y_2^k} (\bar{\bar{\rho}}_1^k - \bar{\bar{\rho}}_{12}^k))^{y_1^k (1-y_2^k)} (\bar{\bar{\rho}}_2^k - \bar{\bar{\rho}}_{12}^k)^{(1-y_1^k) y_2^k} \cdot \\
&\quad \cdot (1 - \bar{\bar{\rho}}_1^k - \bar{\bar{\rho}}_2^k + \bar{\bar{\rho}}_{12}^k)^{(1-y_1^k)(1-y_2^k)},
\end{aligned}$$

where ρ_i^k is the firing probability of neuron i in time step k, $\bar{\rho}_i^k = \left\langle \rho_i^k \right\rangle_{\mathbf{X}^k|Y_i^{k-1}}$ and $\bar{\bar{\rho}}_i^k = \left\langle \rho_i^k \right\rangle_{\mathbf{X}^k|Y_1^{k-1},Y_2^{k-1}}$ are average firing probabilities of neuron i at time k, and $\bar{\bar{\rho}}_{12}^k = \left\langle \rho_1^k \rho_2^k \right\rangle_{\mathbf{X}^k|Y_1^{k-1},Y_2^{k-1}}$ is the average product of firing probabilities of both neurons.

Hence, we can define the factors

$$\begin{aligned} F_2^k &:= \log \frac{P(y_2^k|Y_2^{k-1}, X^k)}{P(y_2^k|Y_2^{k-1})} \\ &= y_2^k \log \frac{\rho_2^k}{\bar{\rho}_2^k} + (1 - y_2^k) \log \frac{1 - \rho_2^k}{1 - \bar{\rho}_2^k}, \end{aligned} \tag{A.16}$$

$$\begin{aligned} G_2^k &:= \log \frac{P(y_2^k|Y_2^{k-1})}{\tilde{P}(y_2^k|Y_2^{k-1})} \\ &= y_2^k \log \frac{\bar{\rho}_2^k}{\tilde{\rho}_2^k} + (1 - y_2^k) \log \frac{1 - \bar{\rho}_2^k}{1 - \tilde{\rho}_2^k}, \end{aligned} \tag{A.17}$$

and

$$\begin{aligned} F_{12}^k &:= \log \frac{P(y_1^k, y_2^k|Y_1^{k-1}, Y_2^{k-1})}{P(y_1^k|Y_1^{k-1}) P(y_2^k|Y_2^{k-1})} \\ &= y_1^k y_2^k \log \frac{\bar{\bar{\rho}}_{12}^k}{\bar{\rho}_1^k \bar{\rho}_2^k} + y_1^k (1 - y_2^k) \log \frac{\bar{\bar{\rho}}_1^k - \bar{\bar{\rho}}_{12}^k}{\bar{\rho}_1^k - \bar{\rho}_1^k \bar{\rho}_2^k} + \\ &\quad + (1 - y_1^k) y_2^k \log \frac{\bar{\bar{\rho}}_2^k - \bar{\bar{\rho}}_{12}^k}{\bar{\rho}_2^k - \bar{\rho}_1^k \bar{\rho}_2^k} + \\ &\quad + (1 - y_1^k)(1 - y_2^k) \log \frac{1 - \bar{\bar{\rho}}_1^k - \bar{\bar{\rho}}_2^k + \bar{\bar{\rho}}_{12}^k}{1 - \bar{\rho}_1^k - \bar{\rho}_1^k + \bar{\rho}_1^k \bar{\rho}_2^k}. \end{aligned} \tag{A.18}$$

Furthermore, we can calculate the derivative in (A.15) which yields the correlation term

$$\begin{aligned} C_{2j}^k &:= \frac{\partial}{\partial w_{2j}} \log P(Y_2^k|X^k) \\ &= \frac{\partial}{\partial w_{2j}} \log \left[\prod_{l=1}^{k} (\rho_2^l)^{y_2^l} (1 - \rho_2^l)^{(1 - y_2^l)} \right] \\ &= \sum_{l=1}^{k} \frac{\partial}{\partial w_{2j}} \left[y_2^l \log(\rho_2^l) + (1 - y_2^l) \log(1 - \rho_2^l) \right] \\ &= \sum_{l=1}^{k} \left[\frac{y_2^l}{\rho_2^l} - \frac{1 - y_2^l}{1 - \rho_2^l} \right] \frac{\partial \rho_2^l}{\partial w_{2j}} \\ &= \sum_{l=1}^{k} \left[\frac{y_2^l}{\rho_2^l} - \frac{1 - y_2^l}{1 - \rho_2^l} \right] \frac{\partial \rho_2^l}{\partial u_2} \sum_{n=1}^{l} \epsilon (t^l - t^n) x_2^n. \end{aligned} \tag{A.19}$$

Summarizing the derived update rule, the weight change at time step k is given by

$$\Delta w_{2j}^k = \alpha \left\langle C_{2j}^k (F_2^k - \frac{\gamma_1}{\Delta t} F_{12}^k - \gamma_2 G_2^k) \right\rangle_{\mathbf{X}^k, \mathbf{Y}_1^k, \mathbf{Y}_2^k}. \tag{A.20}$$

A.2. From Averages to an Online Rule

Now we want to derive an online rule dw_{2j}/dt for $\Delta t \to 0$. First, we note that the terms F_2^k, F_{12}^k and G_2^k depend on postsynaptic variables only and therefore introduce a postsynaptic factor

$$B_{12}^k = \frac{F_2^k - \gamma_2 G_2^k}{\Delta t} - \gamma_1 \frac{F_{12}^k}{(\Delta t)^2}, \tag{A.21}$$

and write

$$\frac{\Delta w_{2j}^k}{\Delta t} = \alpha \left\langle C_{2j}^k B_{12}^k \right\rangle_{\mathbf{X}^k, \mathbf{Y}_1^k, \mathbf{Y}_2^k}. \tag{A.22}$$

Taking the limit $\Delta t \to 0$, according to (Toyoizumi et al., 2005a; Toyoizumi et al., 2005b) the correlation term C_{2j}^k can be transformed into the term $C_{2j}(t)$ given by the differential equation

$$\frac{dC_{2j}(t)}{dt} = -\frac{C_{2j}(t-\delta)}{\tau_C} + \sum_f \epsilon(t - t_j^{(f)}) S_2(t) [\delta(t - \hat{t}_2 - \delta) - g(u_2(t)) R_2(t)]. \tag{A.23}$$

Considering the postsynaptic term B_{12}^k, we make the assumption that the expectations $\langle\rangle_{\mathbf{X}^k|Y_i^{k-1}}$ and $\langle\rangle_{\mathbf{X}^k|Y_1^{k-1},Y_2^{k-1}}$ are approximately equal. This simplifies the expression and we get

$$\begin{aligned} B_{12}^k = {} & \frac{y_2^k}{\Delta t}\left[\log\frac{\rho_2^k}{\bar{\rho}_2^k} - \gamma_2 \log\frac{\bar{\rho}_2^k}{\tilde{\rho}_2^k}\right] + \frac{1-y_2^k}{\Delta t}\left[\log\frac{1-\rho_2^k}{1-\bar{\rho}_2^k} - \gamma_2 \log\frac{1-\bar{\rho}_2^k}{1-\tilde{\rho}_2^k}\right] - \\ & - \gamma_1 \left[\frac{y_1^k y_2^k}{(\Delta t)^2}\log\frac{\bar{\rho}_{12}^k}{\bar{\rho}_1^k \bar{\rho}_2^k} - \frac{(1-y_1^k)y_2^k}{(\Delta t)^2}\log\frac{1-\frac{\bar{\rho}_{12}^k}{\bar{\rho}_2^k}}{1-\bar{\rho}_1^k} - \right. \\ & \left. - \frac{y_1^k(1-y_2^k)}{(\Delta t)^2}\log\frac{1-\frac{\bar{\rho}_{12}^k}{\bar{\rho}_1^k}}{1-\bar{\rho}_2^k} + \frac{(1-y_1^k)(1-y_2^k)}{(\Delta t)^2}\log\frac{1-(\bar{\rho}_1^k + \bar{\rho}_2^k - \bar{\rho}_{12}^k)}{1-(\bar{\rho}_1^k + \bar{\rho}_2^k - \bar{\rho}_1^k \bar{\rho}_2^k)}\right]. \end{aligned} \tag{A.24}$$

We recall the definition

$$\rho_i^k = 1 - \exp[-g(u_i(t^k)) R_i(t^k) \Delta t] \approx g(u_i(t^k)) R_i(t^k) \Delta t \tag{A.25}$$

and define

$$\bar{g}_i(t^k) = \langle g(u_i(t^k)) \rangle_{\mathbf{X}^k|Y_i^{k-1}} \tag{A.26}$$

$$\bar{g}_{12}(t^k) = \langle g(u_1(t^k)) g(u_2(t^k)) \rangle_{\mathbf{X}^k|Y_1^{k-1},Y_2^{k-1}}; \tag{A.27}$$

then,

$$\bar{\rho}_i^k = \bar{g}_i(t^k) R_i(t^k) \Delta t \tag{A.28}$$

$$\bar{\rho}_{12}^k = \bar{g}_{12}(t^k) R_1(t^k) R_2(t^k) (\Delta t)^2. \tag{A.29}$$

Using $\log(1-x) \approx -x$, we get

$$\begin{aligned} B_{12}^k = \frac{y_2^k}{\Delta t}\left[\log\frac{g(u_2(t^k))}{\bar{g}_2(t^k)} - \gamma_2 \log\frac{\bar{g}_2(t^k)}{\tilde{g}}\right] - \\ - (1-y_2^k) R_2(t^k)\left[g(u_2(t^k)) - (1+\gamma_2)\bar{g}_2(t^k) + \gamma_2\tilde{g}\right] - \\ - \gamma_1\Bigg\{\frac{y_1^k y_2^k}{(\Delta t)^2}\log\frac{\bar{g}_{12}(t^k)}{\bar{g}_1(t^k)\bar{g}_2(t^k)} - \\ - (1-y_2^k)\frac{y_1^k}{\Delta t}R_1(t^k)\left[\frac{\bar{g}_{12}(t^k)}{\bar{g}_2(t^k)} - \bar{g}_1(t^k)\right] - \\ - (1-y_1^k)\frac{y_2^k}{\Delta t}R_2(t^k)\left[\frac{\bar{g}_{12}(t^k)}{\bar{g}_1(t^k)} - \bar{g}_2(t^k)\right] + \\ + (1-y_1^k)(1-y_2^k)R_1(t^k)R_2(t^k)\left[\bar{g}_{12}(t^k) - \bar{g}_1(t^k)\bar{g}_2(t^k)\right]\Bigg\}. \end{aligned} \tag{A.30}$$

Taking the limit $\Delta t \to 0$, we replace the term B_{12}^k with the postsynaptic factor

$$B_{12}(t) = B_2^{post}(t) - \gamma_1 B_{12}^{post}(t), \tag{A.31}$$

with terms

$$\begin{aligned} B_2^{post}(t) = \delta(t-\hat{t}_2-\delta)\log\left[\frac{g(u_2(t))}{\bar{g}_2(t)}\left(\frac{\tilde{g}}{\bar{g}_2(t)}\right)^{\gamma_2}\right] - \\ - R_2(t)\left[g(u_2(t)) - (1+\gamma_2)\bar{g}_2(t) + \gamma_2\tilde{g}\right] \end{aligned} \tag{A.32}$$

and

$$\begin{aligned} B_{12}^{post}(t) = \delta(t-\hat{t}_2-\delta)\left\{\delta(t-\hat{t}_1-\delta)\log\frac{\bar{g}_{12}(t)}{\bar{g}_1(t)\bar{g}_2(t)} - R_1(t)\left[\frac{\bar{g}_{12}(t)}{\bar{g}_2(t)} - \bar{g}_1(t)\right]\right\} - \\ - R_2(t)\left\{\delta(t-\hat{t}_1-\delta)\left[\frac{\bar{g}_{12}(t)}{\bar{g}_1(t)} - \bar{g}_2(t)\right] - R_1(t)\left[\bar{g}_{12}(t) - \bar{g}_1(t)\bar{g}_2(t)\right]\right\}. \end{aligned} \tag{A.33}$$

Under the assumption of a small learning rate α the expectations $\langle\cdot\rangle_{\mathbf{X}^k,\mathbf{Y}_1^k,\mathbf{Y}_2^k}$ from (A.20) can be approximated by averaging over a single long trial that allows us to define the online rule:

$$\frac{dw_{2j}(t)}{dt} = \alpha C_{2j}(t) B_{12}(t-\delta). \tag{A.34}$$

B. Notation

w_{ij}	weight value of the synapse connecting presynaptic neuron j to postsynaptic neuron i
u_j	presynaptic acivity (e.g., firing rate) of neuron j
v_i	postsynaptic acivity (e.g., firing rate) of neuron i
α	learning rate ($\alpha > 0$)
$\delta(t)$	Dirac-δ function ($\int_{-\infty}^{+\infty} \delta(t)dt = \int_{-\epsilon}^{+\epsilon} \delta(t)dt = 1$, for any $\epsilon > 0$)
$t_j^{(f)}(t_i^{(f)})$	time of f-th presynaptic (postsynaptic) spike of neuron j (i)
$\mathbf{X}, X$	a random variable ($\mathbf{X}$) and a specific instantiation of this random variable (X)
$P(X)$	the probability that $\mathbf{X}$ takes on value X; or the probability distribution of $\mathbf{X}$
$H(\mathbf{X})$	entropy of the random variable $\mathbf{X}$ (or of the probability distribution $P(X)$)
$H(\mathbf{X}, \mathbf{Y})$	joint entropy of the random variables $\mathbf{X}$ and $\mathbf{Y}$ (i.e., the entropy of the joint distribution $P(X, Y)$)
$H(\mathbf{Y}\|\mathbf{X})$	conditional entropy of $\mathbf{Y}$ given $\mathbf{X}$ (i.e., the average entropy of the distribution $P(Y\|X)$ over all X)
$I(\mathbf{X}, \mathbf{Y})$	mutual information between random variables $\mathbf{X}$ and $\mathbf{Y}$ (or between the distributions $P(X)$ and $P(Y)$)
$I'(\mathbf{X}, \mathbf{Y})$	information rate between $\mathbf{X}$ and $\mathbf{Y}$
$D(P(X)\|\|Q(X))$	Kullback-Leibler divergence between two probability distributions $P(X)$ and $Q(X)$ of the same random variable ($\mathbf{X}$)
$\langle\cdot\rangle_{\mathbf{X}\|Y}$	angular brackets denote an average with respect to the distribution given in the subscript, i.e., in this case $\langle\cdot\rangle_{\mathbf{X}\|Y} = \sum_X \cdot P(X\|Y)$
x_j^k	indicates if there is a input spike at presynaptic neuron j at time step k (i.e., at a time $t_j^{(f)}$ with $t^{k-1} \leq t_j^{(f)} \leq t^k$, with $t^k = k\Delta t$), $x_j^k \in \{0, 1\}$
y_i^k	indicates if there is a output spike at postsynaptic neuron i at time step k, $y_i^k \in \{0, 1\}$
X_j^K	input spike train at synapse j of length $K\Delta t$ (K time bins), $X_j^K = (x_j^1, x_j^2, \ldots, x_j^K)$

X^K	ensemble of input spike trains of length $K\Delta t$ arriving at all synapses $1 \leq j \leq N$, i.e., $X^K = (X_1^K, X_2^K, \ldots, X_N^K)$
Y_i^K	output spike train of neuron i of length $K\Delta t$ (K time bins), $Y_i^K = (y_i^1, y_i^2, \ldots, y_i^K)$
$\mathbf{X}^K, \mathbf{Y}_i^K, \mathbf{y}_i^k$	random variables characterizing the values X^K, Y_i^K, and y_i^k
$\epsilon(t - t_j^{(f)})$	time course of a postsynaptic potential (PSP) caused by a presynaptic spike arrival at synapse j at time $t_j^{(f)}$
$u_i(t)$	membrane potential of neuron i at time t
$R_i(t)$	refractory variable of neuron i at time t, $R_i(t) \in [0; 1]$
$g(u)$	gain function; firing rate in Hz as a smooth increasing function of the membrane potential u
$\tilde{g}$	constant target firing rate
$\bar{g}_i(t)$	average of the output firing rate of neuron i at time t, $\bar{g}_i(t) = \langle g(u_i(t)) \rangle_{\mathbf{X}\|Y_i}$
$\bar{g}_{il}(t)$	average product of firing rates of neurons i and l at time t, $\bar{g}_{il}(t) = \langle g(u_i(t)) g(u_l(t)) \rangle_{\mathbf{X}\|Y_i, Y_l}$
ρ_i^k	firing probability of neuron i at time step k, $\rho_i^k \approx g(u_i(t^k)) R_i(t^k) \Delta t$
$\tilde{\rho}_i^k$	target firing probability of neuron i at time step k, $\tilde{\rho}_i^k = \tilde{g} R_i(t^k) \Delta t$
$\bar{\rho}_i^k$	average firing probability of neuron i at time step k, $\bar{\rho}_i^k = \bar{g}_i(t^k) R_i(t^k) \Delta t$
$\bar{\rho}_{il}^k$	average product of firing probabilities of neurons i and l at time step k, $\bar{\rho}_{il}^k = \bar{g}_{il}(t^k) R_i(t^k) R_l(t^k) (\Delta t)^2$
$C_{ij}(t)$	correlation term sensitive to coincidences between presynaptic input spikes at synapse j and postsynaptic output spikes of neuron i
$B_i^{post}(t)$	postsynaptic term measuring the information transmission between input and output spike trains of neuron i
$B_{il}^{post}(t)$	postsynaptic term reflecting the statistical dependence (mutual information) between the output spike trains of neurons i and l

Bibliography

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

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In Rosenblith, W. A., editor, *Sensory Communication*, pages 217–234. MIT Press.

Barlow, H. B. (1989). Unsupervised learning. *Neural Computation*, 1:295–311.

Bell, A. J. and Parra, L. C. (2005). Maximising information yields spike timing dependent plasticity. In *Advances in Neural Information Processing Systems 17*. MIT Press.

Bell, A. J. and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159.

Bi, G. and Poo, M. (2001). Synaptic modification of correlated activity: Hebb's postulate revisited. *Annu. Rev. Neurosci.*, 24:139–166.

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

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Bohte, S. M. and Mozer, M. C. (2005). Reducing spike train variability: A computational theory of spike-timing dependent plasticity. In *Advances in Neural Information Processing Systems 17*. MIT Press.

Brown, T. H. and Chattarji, S. (1998). Hebbian synaptic plasticity. In Arbib, M. A., editor, *The Handbook of Brain Theory and Neural Networks*, pages 454–459. MIT Press.

Buhl, E. H., Cobb, S. R., Halasy, K., and Somogyi, P. (1995). Properties of unitary IPSPs evoked by anatomically identified basket cells in the rat hippocampus. *Eur. J. Neurosci.*, 7:1989–2004.

Buhl, E. H., Halasy, K., and Somogyi, P. (1994). Diverse sources of hippocampal unitary inhibitory postsynaptic potentials and the number of synaptic release sites. *Nature*, 368:823–828.

Chance, F. S., Abbott, L. F., and Reyes, A. D. (2002). Gain modulation from background synaptic input. *Neuron*, 35:773–782.

Chechik, G. (2003). Spike-timing dependent plasticity and relevant mutual information maximization. *Neural Computation*, 15:1481–1510.

Chechik, G., Horn, D., and Ruppin, E. (2002). Hebbian learning and neuronal regulation. In Arbib, M. A., editor, *The Handbook of Brain Theory and Neural Networks*, pages 511–514. MIT Press.

Cobb, S. R., Buhl, E. H., Halasy, K., Paulsen, O., and Somogyi, P. (1995). Synchronization of neuronal activity in hippocampus by individual GABAergic interneurons. *Nature*, 378:75–78.

Cooper, L. N., Intrator, N., Blais, B. S., and Shouval, H. Z. (2004). *Theory of Cortical Plasticity*. World Scientific Publishing.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley.

Dayan, P. and Abbott, L. F. (2001). *Theoretical Neuroscience*. MIT Press.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification*. Wiley.

Erdi, P. and Somogyvari, Z. (2002). Post-hebbian learning algorithms. In Arbib, M. A., editor, *The Handbook of Brain Theory and Neural Networks*, pages 898–901. MIT Press.

Fregnac, Y. (2002). Hebbian synaptic plasticity. In Arbib, M. A., editor, *The Handbook of Brain Theory and Neural Networks*, pages 515–522. MIT Press.

Froemke, R. C. and Dan, Y. (2002). Spike-timing-dependent synaptic modification induced by natural spike trains. *Nature*, 415:433–438.

Gerstner, W. and Kistler, W. M. (2002). *Spiking Neuron Models*. Cambridge University Press.

Hebb, D. O. (1949). *The Organization of Behavior*. Wiley.

Hyvärinen, A., Hoyer, P. O., Hurri, J., and Gutmann, M. (2005). Statistical models of images and early vision. In *Proceedings of the Int. Symposium on Adaptive Knowledge Representation and Reasoning (AKRR2005)*.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. Wiley.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, 13:411–430.

Intrator, N. and Cooper, L. N. (1998). BCM theory of visual cortical plasticity. In Arbib, M. A., editor, *The Handbook of Brain Theory and Neural Networks*, pages 153–157. MIT Press.

Legenstein, R., Naeger, C., and Maass, W. (2005). What can a neuron learn with spike-timing-dependent plasticity? *Neural Computation*, 17:2337–2382.

Linsker, R. (1989). How to generate ordered maps by maximizing the mutual information between input and output signals. *Neural Computation*, 1:402–411.

Maass, W. and Markram, H. (2002). Synapses as dynamic memory buffers. *Neural Networks*, 15:155–161.

Markram, H., Toledo-Rodriguez, M., Wang, Y., Gupta, A., Silberberg, G., and Wu, C. (2004). Interneurons of the neocortical inhibitory system. *Nat. Rev. Neurosci.*, 5:793–807.

Markram, H., Wang, Y., and Tsodyks, M. (1998). Differential signaling via the same axon of neocortical pyramidal neurons. *Proc. Natl. Acad. Sci. USA*, 95:5323–5328.

Miles, R., Toth, K., Gulyas, A. I., Hajos, N., and Freund, T. F. (1996). Differences between somatic and dendritic inhibition in the hippocampus. *Neuron*, 16:815–823.

Nadal, J.-P. and Parga, N. (1997). Redundancy reduction and independent component analysis: Condition on cumulants and adaptive approaches. *Neural Computation*, 9:1421–1456.

Nelson, S. B., Sjøstrøm, P. J., and Turrigiano, G. G. (2002). Rate and timing in cortical synaptic plasticity. *Phil. Trans. R. Soc. Lond.*, 357:1851–1857.

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

Pfister, J.-P., Toyoizumi, T., Barber, D., and Gerstner, W. (2005). Optimal spike-timing dependent plasticity for precise action potential firing. To appear (arXiv:q-bio.NC/0502037).

Ramón y Cajal, S. (1911). *Histologie du système nerveux de l'homme et des vertèbres*. Paris Maloine.

Rieke, F., Warland, D., de Ruyter van Steveninck, R., and Bialek, W. (1997). *Spikes - Exploring the Neural Code*. MIT Press.

Salinas, E. and Thier, P. (2000). Gain modulation: A major computational principle of the central nervous system. *Neuron*, 27:15–21.

Semyanov, A., Walker, M. C., Kullmann, D. M., and Silver, R. A. (2004). Tonically active GABA-A receptors: modulating gain and maintaining the tone. *Trends in Neurosciences*, 27:262–269.

Shannon, C. E. and Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press.

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

Song, S., Miller, K. D., and Abbott, L. F. (2000). Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *Nature Neurosci.*, 3:919–926.

Squire, L. R. and Kandel, E. R. (1999). *Memory - From Mind To Molecules*. Scientific American Library.

Toledo-Rodriguez, M., Gupta, A., Wang, Y., Wu, C. Z., and Markram, H. (2002). Neocortex: Basic neuron types. In Arbib, M. A., editor, *The Handbook of Brain Theory and Neural Networks*, pages 719–725. MIT Press.

Toyoizumi, T., Pfister, J.-P., Aihara, K., and Gerstner, W. (2005a). Generalized bienenstock-cooper-munro rule for spiking neurons that maximizes information transmission. *Proc. Natl. Acad. Sci. USA*, 102:5239–5244.

Toyoizumi, T., Pfister, J.-P., Aihara, K., and Gerstner, W. (2005b). Generalized bienenstock-cooper-munro rule for spiking neurons that maximizes information transmission - supporting information.

Toyoizumi, T., Pfister, J.-P., Aihara, K., and Gerstner, W. (2005c). Spike-timing dependent plasticity and mutual information maximization for a spiking neuron model. In *Advances in Neural Information Processing Systems 17*. MIT Press.

Triesch, J. (2005). Synergies between intrinsic and synaptic plasticity in individual model neurons. In *Advances in Neural Information Processing Systems 17*. MIT Press.

Turrigiano, G. G. and Nelson, S. B. (2004). Homeostatic plasticity in the developing nervous system. *Nat. Rev. Neurosci.*, 5:97–107.

Wang, Y., Gupta, A., Toledo-Rodriguez, M., Wu, C. Z., and Markram, H. (2002). Anatomical, physiological, molecular and circuit properties of nest basket cells in the developing somatosensory cortex. *Cerebral Cortex*, 12:395–410.

Zhu, Y., Stornetta, R. L., and Zhu, J. J. (2004). Chandelier cells control excessive cortical excitation: Characteristics of whisker-evoked synaptic responses of layer 2/3 nonpyramidal and pyramidal neurons. *J. Neurosci.*, 24:5101–5108.