Model-Based Reinforcement Learning by Pyramidal Neurons: Robustness of the Learning Rule

Michael Eisele & Terry Sejnowski
Howard Hughes Medical Institute
Computational Neurobiology Laboratory
The Salk Institute
San Diego, CA 92186-5800

Abstract

Reinforcement learning of control is studied in computational models of pyramidal neurons. We have previously demonstrated that a neuron can not only learn direct control, in which the neuron's activity is directly associated with later rewards, but also indirect control, in which the association is based on a model of the controlled dynamics. Here we will summarize these learning principles: the central task of the learning rule consists in detecting small transient changes in the postsynaptic firing rate. Then the robustness of the learning rule is demonstrated by applying it to a set of cyclically activated synapses, which can detect cyclic changes in the postsynaptic firing rate. The learned response is insensitive to inhibition, dendritic nonlinearities, dendritic time constants, overlaps of distributed input patterns, or input noise. Predictions are made for the learning principles employed by pyramidal neurons.

1 Introduction

It has been demonstrated experimentally [1, 2] that single pyramidal neurons may be capable of reinforcement learning, if some external reinforcement signal is provided. Several algorithms for reinforcement learning by single neurons are known (see review in [3]). In the terminology of control theory [4], all these algorithms are classified as "direct" reinforcement learning, because they directly associate a neuron's output with the reinforcement signals. If the correlation between the neuron's output and the reinforcement signal is small, then these algorithms can learn only very slowly.

A potentially more efficient type of algorithms is known as "indirect" or "model based" reinforcement learning. Agents which employ these algorithms need feedback from the dynamics that they attempt to control. They use this feedback to learn a model of how the dynamics is affected by their output. Such a model can be learned even in the absence of reinforcements. And given enough time, the agents can detect even small effects. In a second learning phase, the agents can use this model to respond correctly to reinforcement signals: they detect the states of the dynamics whose occurrence is associated with reinforcement signals and act on the dynamics such that the probability of these states is increased. The known algorithms for model-based reinforcement learning are, however, so complex, that only a whole network, but not an individual neuron could act as controlling agent.

We have recently introduced a new form of model-based reinforcement learning, in which only a very reduced model of the dynamics is learned by the agent. We demonstrated that the new learning principles can be implemented within a single neuron. We also argued that the anatomy and physiology of neocortical neurons is especially suited for this task[5]. Our numerical simulations were, however, based on a very simplified neuron model: synaptic transmission was not stochastic, excitatory postsynaptic potentials lasted for only one time step, and signal transmission by dendrites was linear. Here we will demonstrate that these assumptions are not essential.

Section 2 contains a short summary of our learning principles for model-based reinforcement. The focus is on the first part of the learning process, the learning of the model in the absence of reinforcement, because the second part, the reinforcement-induced learning, may be transmitted by biochemical pathways which are quite insensitive to all the disturbances mentioned above. The first part of the learning process turns out to be equivalent to the detection of small transient changes in the postsynaptic firing rate. In section 3 the robustness of this learning process is demonstrated in a numerical example. A small set of cyclically stimulated synapses detects transient changes in the postsynaptic firing rate. The learned responses are shown to be insensitive to inhibition, dendritic nonlinearities, the duration of excitatory postsynaptic potentials, overlaps between input patterns, or synaptic transmission noise,

2 Learning Principles

2.1 Direct and Model-Based Reinforcement

This section summarizes our version of model-based reinforcement learning and demonstrates how the learning of the model reduces to the detection of small transient changes in the neuron's firing rate. As mentioned above, our algorithm learns only a reduced model of the controlled dynamics. How far can one reduce the model, without losing the advantage over direct reinforcement learning? Consider, as an illustrative example, a population of neurons whose output induces some motor response, like the movement of a finger. Assume that a subpopulation of "extension" neurons extends the finger and another subpopulation of "flexion" neurons flexes the finger. Assume that in response to some stimuli more "extension" neurons than "flexion" neurons fire and that this causes the finger to be extended. Assume, furthermore, that this response is followed by a reinforcement signal. The best way to reinforce this response would be to strengthen synapses on all the extension neurons and to weaken those on flexion neurons. Direct reinforcement learning does not achieve this goal: it strengthens all the recently active neurons, including active flexion neurons, and weakens all the the passive neurons, including the passive extension neurons. As more extension neurons have been active than flexion neurons, the response of the network is shifted into the right direction, but many such learning trials will be necessary before the two classes of neurons are clearly separated.

In model-based reinforcement, on the other hand, learning neurons use their previous experience to determine whether they are extending or flexing. They then use this model of their own effect to assign the credit so that only extension neurons are reinforced. They can make the distinction between extension and flexion without having a complete model of the network and motor dynamics; all they have to do is observe whether extension of the finger is more likely to occur after they were active or after they were passive. If the effect of a single neuron on the finger's movement is small, than it may take a long time to detect this effect. But this learning step can take place in an unsupervised way, even in the absence of reinforcements. And using this experience, neurons can respond to reinforcements much faster than in the case of direct reinforcement learning.

2.2 Positive and Negative Effects

The same learning principle can be applied in a more general setting. All the neuron needs are feedback projections from those brain regions that its output is acting on. Then it can observe

which of the synaptic input patterns in the feedback are more likely to occur after it fires. It can take this as evidence that its firing has a "positive" effect on whatever response the brain has just generated, like the effect that extension neurons exert on extending the finger. Likewise, if the probability of an input pattern is reduced by a neuron's firing, this can be taken as evidence for a "negative" effect, like the effect of a flexion neuron on extending the finger. Once a neuron knows which of its input patterns it effects positively or negatively, it can easily respond to reinforcements, by simply increasing synaptic strengths whenever positive effects are associated with reinforcements and decreasing synaptic strengths whenever negative effects are associated with reinforcements.

But how can a neuron store its knowledge about positive and negative effects? In our version of model-based reinforcement learning, this information is stored in distal apical synapses, such that neurons with different effects show different dendritic excitation patterns: During extension of the finger, extension neurons would show a large excitation in the apical tuft and flexion neurons a small one (while during flexion of the finger the roles would be reversed). There is a plausible biochemical substrate which could associate such an excitation pattern with a reinforcement signal, for example, a dopaminergic signal: Adenyl cyclases may be activated both by dopamine and local membrane depolarization[6], and via the protein kinase A pathway they may subsequently strengthen synaptic weights. Further support for this idea comes from the observation that both dopaminergic afferents in primates[7] and feedback projections from higher levels in the cortical hierarchy[8] target preferably layer 1, which is filled with apical tufts of pyramidal neurons.

2.3 Learning Rule

How can a neuron learn to respond with large apical excitation whenever it has a positive effect? As in other rules of the Hebbian type, the only signals that can be associated at a synapse are the local excitation a and the backpropagating action potential y, which show up as the slow and fast component of the postsynaptic membrane depolarization, the presynaptic input x, and local chemical messengers that are too slow to contribute much to this learning task. The postsynaptic firing y can of course depend on the excitation a, but this dependency will not be important in the following (this might explain why apical tufts may be far away from the soma). What's important, is the dependency of the input x on the firing y: If some synapse signals a positive effect of the neuron, then this synapses will be preferably active shortly after the neuron's firing. It can detect the correlation, if it maintains some signal trace $\langle y \rangle$ of recent postsynaptic action potentials and associates it with the present input x.

This association is done by the first term in the learning rule:

$$\frac{dw}{dt} = \eta \cdot w \cdot (\alpha \cdot x \cdot \langle y \rangle - \alpha \cdot \langle x \rangle \cdot y + \langle x \rangle \cdot (1 - a^{\beta})) \qquad (1)$$

$$\text{with the trace } \langle y \rangle \text{ defined as: } \frac{d\langle y \rangle}{dt} = (y - \langle y \rangle) / \tau \qquad (2)$$

This learning-rule is a simplification of the rule that was used previously[5] to demonstrate reinforcement learning. In the numerical simulations the parameters were chosen as $\tau = 100$ msec, $\alpha = 250$ msec, $\eta = 1/(50$ msec), and $\beta = 1$.

The learning rule contains two anti-Hebbian terms. The term that is proportional to $-\langle x \rangle y$ associates the postsynaptic firing y with a signal trace $\langle x \rangle$ of the input (which is defined in analogy to eq. (2)). If x and y are uncorrelated, then $\langle y \rangle x - \langle x \rangle y = 0$ on average and the first two terms in the learning rule cancel. The other term $\langle x \rangle (1 - a^{\beta})$ limits the growth of synaptic weights by weakening synapses which are active while the excitation a exceeds the mean excitation 1. In real synapses, the local excitation may affect the learning process, for example, through the voltage-dependent magnesium block of NMDA channels.

The use of the time scale $\tau = 100$ msec for all the signal traces in eq. (1) is justified by the fact that both the opening of NMDA channels and the calcium transients in spines[9] take place on a time

scale of 100 msec. Furthermore, it has been measured directly that excitatory synapses are able to associate pre- and postsynaptic action potentials that occur within about 100 msec, to detect their temporal order, and to change their weight accordingly [10, 11, 12]. These experiments do indicate that the parameter α is negative on proximal synapses; see ref. [5] for a further discussion of the biological plausibility of the learning rule. Of course, animals can learn to produce responses that bridge delays of many seconds during which no external feedback occurs, but this does not necessarily disagree with our choice of the small association time $\tau = 100$ msec: As such a behavioural response is due to a whole series of neural activations, each neuron in the series may receive internal feedback from the next stage of the series in less than 100 msec.

2.4 Simplified Learning Task

The learning rule (1) is sensitive to the temporal derivative of the neuron's firing. If the firing rate prior to the synaptic input $x > 0$ is larger than the subsequent firing rate, then $\langle y \rangle x$ will be larger than $\langle x \rangle y$ and the synapse will grow stronger. In contrast to temporal difference learning, however, the main task of the learning rule (1) is not to level out such transient changes of the firing rate, but rather to mark them by large deviations of the local excitation a from its mean value 1. It does not matter whether a positive term $\langle y \rangle x - \langle x \rangle y > 0$ is due to the neuron having a positive effect on the input x or whether it is due to a transient decrease in the neuron's firing rate: In both cases, the same a-deviation will be learned. We can use this to test the robustness of the learning rule in an especially simple setting, in which the neuron exerts no effects and all the learned a-deviations are due to transient changes in the firing rate.

3 Robustness of the learning rule

3.1 Standard Learning Task

The learning rule (1) is tested here for a simple, representative synaptic input. 20 synapses are simulated. They experience the same postsynaptic excitation a and action potential y, but different inputs x_i. The postsynaptic action potential y is set equal to a real number, the firing probability $P_y \approx 10$ Hz. The same presynaptic inputs are paired again and again with the same pre-set firing probabilities P_y. They change every 50 msec and are presented in a cyclic order. For half of the cycle P_y is fixed at 11 Hz, for the other half at 9 Hz. The learning tasks consists in producing a large excitation a whenever P_y decreases and a small a whenever $P(y)$ increases. Starting from random synaptic weights, a learning time of 10^4 sec was sufficient to assure convergence to a fixed, cyclic response.

In the simplest case, only one afferent was taken to be active during each 50 msec time interval. The corresponding input x_i was defined to be a square pulse of length 50 msec, which describes the average input received during random presynaptic spiking in this time interval. The excitation is defined as the weighted sum $a = \sum_i w_i f[x_i]$. Here $f[x_i]$ is the convolution of the input x_i with an α-function, which has a time constant of 10 msec and corresponds to an excitatory postsynaptic potential.

The result of the learning process is shown in fig. 1. The synapse has succeeded to mark the transition between different postsynaptic firing rates by strong deviations of a from its mean value 1. Around the time $t \approx 250$ msec, for example, where the firing rate decreases, the learned excitation a is large (except for the neglegibly small opposite response at time $t \approx 200$ msec, which is due to the Anti-Hebbian term $-\langle x \rangle a^\beta$ and the large a at times $t > 200$ msec.)

3.2 Inhibition

How does the learning process perform under more realistic conditions? The membrane depolarization of a real dendrite will depend not only on the local excitatory input, but also on more distant input and on inhibition. Subtracting a constant inhibition from the excitatory input $\sum_i w_i f[x_i]$

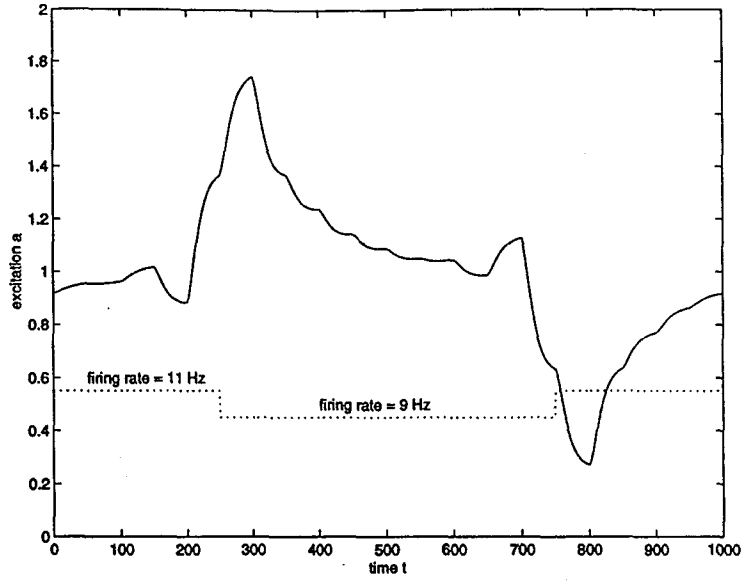

Figure 1: Standard result of the learning process. The excitation (solid line, arbitrary units) produced by 20 synapses, which are stimulated cyclically for 50 msec each, is plotted against time (in milliseconds, time = 0 denotes the end of the learning process). The learning process detects changes in the pre-set postsynaptic firing rate (dotted line, changes at 250 msec and 750 msec) and marks them by large deviations of the excitation from its mean value. These learned deviations start even slightly before the postsynaptic firing rate changes, because synapses learn to predict the rate changes, which occur always at the same points in the cycle. See text for details.

did, however, not change the result of the learning process: the synaptic weights w just grew stronger and compensated the inhibition exactly. An analogous compensation was also learned for time dependent excitatory or inhibitory input, unless this input changed so abruptly that no possible choice of local excitatory weights could compensate for it.

3.3 Nonlinearities

A real dendrite will also respond to its input in a nonlinear way, and even if it would respond linearly, than the dependence of the learning process on the local membrane depolarization will probably be nonlinear. This was modeled by introducing a nonlinearity $\beta \neq 1$ into the learning rule (1). As fig. 2 shows, the learned a–deviations still occur at the right times. Their amplitude, however, could change. These amplitudes could be kept roughly constant, if the parameter α was multiplied by the same factor as β. Thus the nonlinear learning rule is as powerful as the original linear rule.

3.4 Time Constants

The learning process is also robust against changes in the speed of excitatory postsynaptic potentials. The dashed-dotted line in fig. 3 shows an example where the α-function that describes the excitatory postsynaptic potential was given a time constant of 100 msec instead of 10 msec. To assure convergence, the learning process was simulated for a longer time of $5 \cdot 10^4$ sec with a larger learning rate $\eta = 1/(10 \text{ msec})$. The learned a–deviations are similar to the original ones.

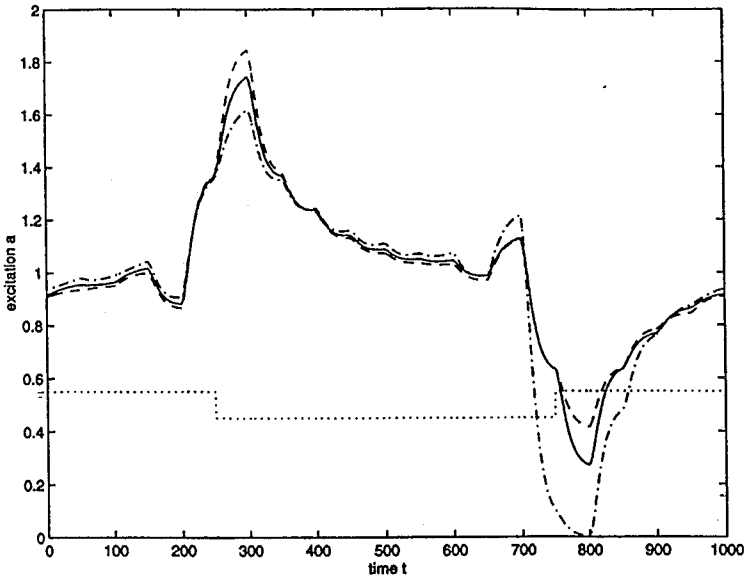

Figure 2: Robustness against dendritic nonlinearities. The solid line shows the standard response from fig. 1 ($\beta = 1, \alpha = 250$ msec). Similar responses were learned, if the learning rule depended nonlinearly on the synaptic excitation: Dashed line: $\beta = 0.5, \alpha = 125$ msec. Dash-dotted line: $\beta = 2, \alpha = 500$ msec.

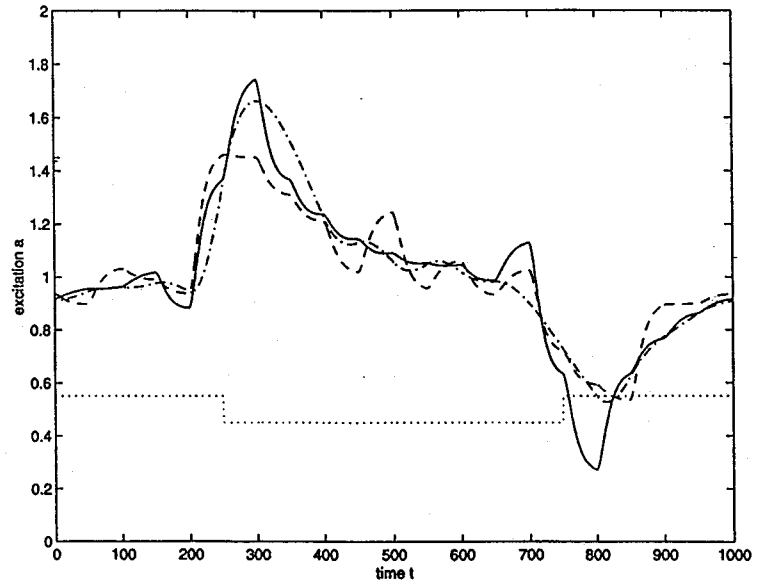

Figure 3: Robustness against slower postsynaptic potentials and distributed synaptic input. The solid line shows the standard response from fig. 1. It was produced by synaptic input patterns that were sparse and excitatory postsynaptic potentials that had a time scale of 10 msec. Similar responses were learned, if this time scale was increased to 100 msec (dashed-dotted line) or if the synaptic input patterns were distributed and overlapping (dotted line).

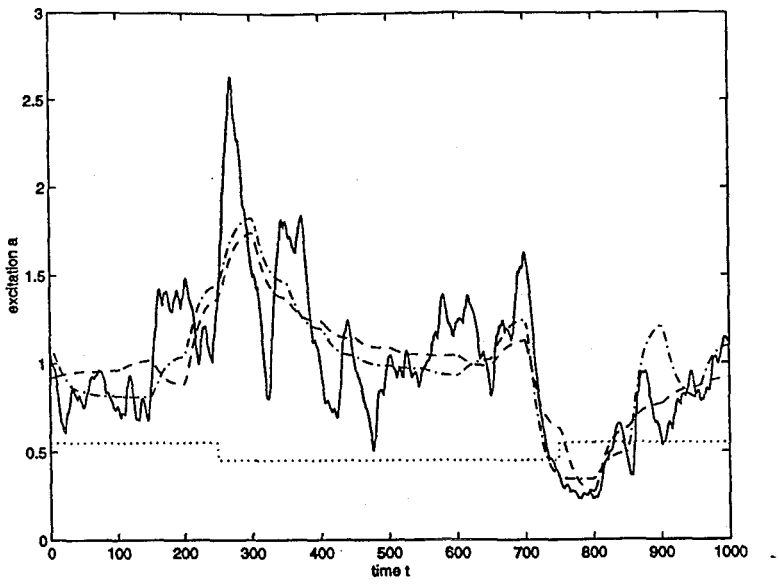

Figure 4: Robustness against input and output noise. The dashed line shows the standard response from fig. 1. It is produced by inserting the noiseless pre- and postsynaptic firing rates into the learning rule. A similar response was learned, when random trains of pre- and postsynaptic action potentials were inserted into the learning rule. The dashed-dotted line shows the average response, the solid line the noisy response during one particular cycle.

3.5 Distributed Input

A more serious challenge to the learning rule arises from distributed synaptic input patterns. As the correlation between the firing of neocortical neurons depends on the global state of the network[13], it is likely that the effect of a neuron's firing on any one of its feedback x_i will depend on the global activity pattern, as signaled by all the other feedbacks x_j. To mark effects, the a-deviations should depend not so much on any particular input x_i, but rather on the whole synaptic input pattern. Thus the learning rule was tested with overlapping input patterns. 100 synapses were simulated, each of which had randomly chosen but fixed afferent firing rates in each of the 50 msec time intervals. The learning process in 87 synapses led to weights closer and closer to zero, which eventually could be neglected. The 13 synapses that survived produced the response of fig. 3 (dashed line), which is similar to the response learned in the case of non overlapping input patterns.

3.6 Noisy Input and Output

Finally, the learning rule was tested in the presence of input and output noise. The postsynaptic action potential y was modeled by delta functions that occurred at random times with a rate of 9 or 11 Hz. The spatial input patterns were again chosen to be sparse and non-overlapping. The input x_i was chosen to be the sum of temporally overlapping α-functions. Their starting times were chosen at random, so that the local excitation a differed from cycle to cycle, even if the weights remained constant. The α-functions occurred at a rate of 500Hz. As this rate is too high to be realistic, each synapse should be regarded as representing a whole group of synapses that are active independently during the same 50 msec time interval. Even with this large rate of excitatory postsynaptic potentials, the stochastic fluctuations in the excitation a (solid line in fig. 4) are still quite large, but not too large to hide the learned a–deviations. And by averaging over several cycles, one gets an average response (dash–dotted line) that is close to the response learned in the absence of noise (dashed line).

4 Conclusion

These numerical examples demonstrate the robustness of learning process against inhibition, dendritic nonlinearities, changes in dendritic time scale, overlap between input patterns, and, to a lesser degree, input noise. This robustness is not entirely surprising, as the learning rule is related to another rule, which can be derived from an objective function[5]. The objective function allows one to express the learned excitation in terms of the postsynaptic temporal firing patterns, independent of the spatial form and the noisiness of the input patterns. However, this derivation works only for the case of linear independent input patterns, whose excitatory postsynaptic potentials last only one time step. Here we demonstrated that the learning rule maintains its robustness under more realistic conditions. This supports the hypothesis that pyramidal neurons are capable of model-based reinforcement learning.

References

[1] L. Stein and J. D. Belluzzi. Cellular investigations of behavioral reinforcement. *Neurosci. Biobeh. Rev.*, 13:69–80, 1989.

[2] N. A. Otmakhova and J. E. Lisman. D1/D5 dopamine receptor activation increases the magnitude of early long-term potentiation at CA1 hippocampal synapses. *J. Neurosci*, 16:7478–86, 1996.

[3] A. G. Barto, S. J. Bradtke, and S. P. Singh. Learning to act using real-time dynamic programming. *Art. Intell.*, 72:81–138, 1995.

[4] G. C. Goodwin and K. S. Sin. *Adaptive Filtering, Prediction, and Control.* Prentice-Hall, Englewood Cliffs, N.J., 1984.

[5] M. Eisele. Self determination of reinforcement learning in single neurons. submitted to *Network.*

[6] R. Reddy et al. Voltage–sensitive adenylyl cyclase activity in cultured neurons. A calcium-independent phenomenon. *J. Biol. Chem.*, 270:14340–14346, 1995.

[7] B. Berger, P. Gaspar, and C. Verney. Dopaminergic innervation of the cerebral cortex: Unexpected differences between rodents and primates. *Trends Neurosci.*, 14:21–27, 1991.

[8] D. J. Felleman and D. C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex*, 1:1–47, 1991.

[9] R. Yuste and W. Denk. Dendritic spines as basic functional units of neuronal integration. *Nature*, 375:682–684, 1995.

[10] B. Gustafsson et al. Long-term potentiation in the hippocampus using depolarizing current pulses as the conditioning stimulus to single volley synaptic potentials. *J. Neurosci.*, 7:774–780, 1987.

[11] D. Debanne, B. H. Gahwiler, and S. M. Thompson. Asynchronous pre– and postsynaptic activity induces associative long–term depression in area CA1 of the rat hippocampus in vitro. *Proc. Natl. Acad. Sci. USA*, 91:1148–1152, 1994.

[12] H. Markram et al. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, 275:213–215, 1997.

[13] E. Vaadia et al. Dynamics of neuronal interactions in monkey cortex in relation to behavioural events. *Nature*, 373:515–518, 1995.