Maximising Information yields Spike Timing Dependent Plasticity

Anthony J. Bell,
Redwood Neuroscience Institute
1010 El Camino Real, Suite 380
Menlo Park, CA 94025
tbell@rni.org

Lucas C. Parra
Biomedical Engineering Department
City College of New York
New York, NY 10033
parra@ccny.cuny.edu

Abstract

Experiments show a synaptic weight potentiating if its presynaptic spike just preceded its postsynaptic one, and depressing if it came just after, with a sharp transition at synchrony. To understand why, we would like to derive this rule from first principles. To do this, we first calculate the dependency of the postsynaptic spike timing on the presynaptic spike timing in a linear spiking model called the Spike Response Model. We then use this to calculate the gradient of the information transfer in a spiking network. This produces a non-local learning rule for the weights which has the correct signs of potentiation and depression, but without the sharp transition. Since the rule is analogous to ICA, we follow Amari in transforming the gradient by an approximate Hessian to get a Natural Gradient rule. This yields an almost-local learning algorithm in which a sharp transition between potentiation and depression now appears. The main mismatch between our rule and the experiment is an offset of the sharp transition from synchrony. We believe this is due to a mismatch between the Spike Response Model and the real neuron. We propose that information maximisation occurs across time through a causal network of spike timings.

1 Introduction

Progress in machine learning has not produced agreement on what, if anything, real neural systems optimise. Meanwhile, artificial neural learning rules are usually derived for 'real-valued neurons' (i.e. rate models). However, real neurons use spikes, and *in vitro* experiments strongly suggest that Hebbian plasticity at excitatory cortical and hippocampal neurons depends critically on the relative timing of pre- and post-synaptic spikes [6,7,11]. This 'spike-timing dependent plasticity' (STDP) can be loosely interpreted as strengthening synapses to the degree that the input spike helped the output spike to occur, and weakening them to the degree that this contribution is small (see Figure 1a).

In this paper, we show that the qualitative form of these STDP curves can be explained by the principle of maximisation of information transfer in a deterministic system. We use the 'Natural Gradient Infomax' procedure [1,4] to derive a learning rule for a linear spiking network called the Spike Response Model [8]. This model is similar to the Integrate-

and-Fire model except it models, with linear kernels, the dynamics of EPSPs (excitatory post-synaptic potentials) and post-spike repolarisation.

The infomax principle has previously been used in firing-rate models to explain the learning of many properties of early visual receptive fields [5,9]. Applied to spiking syatems, the effect of spike timings on each other is maximised, rather than the effect of firing rates on each other. The ability of this basic theory to reproduce key aspects of experimental data provides further support to the notion that the brain optimises information transfer. In addition, the learning rule opens up the possibility of doing machine learning with spikes.

2 Information maximization

We start with a brief recapitulation of the basic argument of [4]. If a system has inputs $\mathbf{x}$, the same number of outputs $\mathbf{y}$, and parameters $\mathbf{W}$ that deterministically and invertibly relate $\mathbf{y}$ to $\mathbf{x}$, then the dependence of the mutual information between $\mathbf{y}$ and $\mathbf{x}$ on the parameters $\mathbf{W}$, can be written as follows:

$$\nabla_{\mathbf{W}} I(\mathbf{y}, \mathbf{x}) = \nabla_{\mathbf{W}} H(\mathbf{y}) = \langle \nabla_{\mathbf{W}} \log |\mathbf{J}| \rangle \ , \tag{1}$$

where I means mutual information, H means entropy, $\langle . \rangle$ means expected value over the distribution of $\mathbf{x}$, the Jacobian matrix of the transformation is $\mathbf{J} = [\partial y_i / \partial x_j]_{ij}$, and $|.|$ means absolute value of the determinant of this matrix. We now apply this argument to spike timings.

2.1 For two spikes.

Consider a spike occurring in neuron j at time t_l that has an effect on the timing of another spike occurring in neuron i at time t_k. The neurons are connected by a weight w_{ij}. Often, we will write $i(k)$ or $j(l)$ to denote the neuron index of the corresponding spike. Similarly, we will use the shorthand $w_{kl} \equiv w_{i(k)j(l)} = w_{ij}$ to denote the constant part of the weight between two spikes. The variable part, which models the time-course of the EPSP is written as $R_{kl}(t_k - t_l) \equiv R_{i(k)j(l)}(t_k - t_l)$. In general, this R_{kl} models both synaptic and dendritic linear responses to an input spike, and thus models synapse type and location. An example is shown in Figure 2.

In this simplest version of the Spike Response Model [8], a neuron adds up its spiking inputs until its 'potential' $u_i(t)$ reaches threshold at time t_k. This threshold we will often write $u_k \equiv u_{i(k)}(t_k, \{t_l\})$, and it is given by a sum over spikes l:

$$u_k = \sum_l w_{kl} R_{kl}(t_k - t_l) \, . \tag{2}$$

To maximise information transfer, we need to determine the effect of a small change in the input firing time t_l on the output firing time t_k. (A related problem is tackled in [3].) When t_l is changed by a small amount dt_l the membrane potential will change as a result. This change in the membrane potential leads to a change in the time of threshold crossing dt_k. The contribution to the membrane potential, du, due to dt_l is $(\partial u_k / \partial t_l) dt_l$, and the change in du corresponding to a change dt_k is $(\partial u_k / \partial t_k) dt_k$. We can relate these two effects by noting that the total change of the membrane potential du has to vanish because u_k is defined as the potential at threshold. ie:

$$du = \frac{\partial u_k}{\partial t_k} dt_k + \frac{\partial u_k}{\partial t_l} dt_l = 0 \, . \tag{3}$$

This is the *total differential* of the function $u_k = u(t_k, \{t_l\})$, and is a special case of the implicit function theorem. Rearranging this:

$$\frac{dt_k}{dt_l} \quad = \quad -\frac{\partial u_k}{\partial t_l} \Big/ \frac{\partial u_k}{\partial t_k} \tag{4}$$

$$= \quad -w_{kl}\dot{R}_{kl}/\dot{u}_k \,. \tag{5}$$

Now, utilising the time derivative of (2), the weight dependence of the spike-to-spike information transfer is easily calculated.

$$\frac{\partial}{\partial w_{kl}} I(t_k, t_l) \quad = \quad \frac{\partial}{\partial w_{kl}} \log \left| \frac{dt_k}{dt_l} \right| \tag{6}$$

$$= \quad \frac{1}{w_{kl}} - \frac{\dot{R}_{kl}}{\dot{u}_k} \,. \tag{7}$$

This has the same form as the one-input/one-output infomax derivation in [4], except that now the input variable and output variable are the values at threshold-crossing of, respectively, the rate of change of the synaptic response and the inverse of the rate of change of the potential. A learning rule utilising this gradient (for random simulated input) produces a typical scatterplot of weight changes shown in Figure 1b. A weight potentiates or depresses if, in doing so, it increases the influence of t_l on t_k, just as in infomax.

Two important differences with the typical experimental result (which is reproduced in Figure 1a) are that (1) the transition between potentiation and depression is graded rather than discontinuous, and (2) this transition occurs at a time offset by the rise time of the EPSP (ie: the time to reach a zero-crossing of $\dot{R}$). The non-zero offset we will come to in the discussion. But first, in order to understand the discontinuous STDP transition, it will be necessary to develop the argument for mappings from N spikes to N spikes, as in [4], and to apply the Natural Gradient arguments of Amari [1] to achieve an almost-local and much more physiological learning rule than that given by straightforward gradient ascent in mutual information. We need to do this at any rate because, as the astute reader will have noticed, when there are only two spikes, the gradient in (7) vanishes.

2.2 For $N \to N$ spike mappings

To maximise the information transfer in a 'square' mapping between spikes, we must, according to (1) maximise the log determinant of a Jacobian matrix, $\mathbf{T}$, the entries of which are the timing dependencies $\mathbf{T}_{kl} \equiv \partial t_k / \partial t_l$. The calculation of this gradient for the full Spike Response Model is in the Appendix. It yields a learning rule in which every interaction between a presynaptic spike at t_l and a postsynaptic spike at t_k causes a weight change:

$$\Delta w_{kl} \propto \frac{\partial \log |\mathbf{T}|}{\partial w_{kl}} = \frac{\mathbf{T}_{kl}}{w_{kl}} \left[\mathbf{T}^{-1}_{lk} - 1 \right] \,, \tag{8}$$

where $\mathbf{T}^{-1}_{lk}$ is the lk-th index of the matrix inverse of $\mathbf{T}$. Equation (8) is the N-dimensional equivalent of Equation (7). Like the original information gradient in Infomax/ICA [4], this learning rule is non-local, and requires a matrix inverse at each step. In addition, although the signs of the weight changes for different relative spike timings (Figure 1b) do match the general pattern of the weight changes observed in experiment (Figure 1a), the shape of the curves do not match.

Inspired by Amari's Natural Gradient version of the algorithm, we now seek to multiply the gradient by a positive definite matrix which does not change the location or sign of the solutions to the learning, but which should make the learning rule simpler and much more local, as well as greatly speeding its convergence. The correct matrix to use is the inverse of the Fisher Information matrix, which is the Riemannian metric tensor of the manifold of weight-matrix-parameterised probability models [1]. In practice, in ICA, we actually multiply the gradient by a simpler data-independent matrix which approximates this effect. To find this matrix in our new situation, we derive, much in the manner of [12], an approximation of the inverse of the Hessian. This yields a 'covariant' gradient update rule.

In the appendix we show that an approximate inverse of the Hessian of $\log|\mathbf{T}|$ is given by:

$$\mathbf{M}_{(kl)(ab)} = -\frac{w_{kl}}{\mathbf{T}_{kl}}\mathbf{T}_{al}\mathbf{T}_{kb}\frac{w_{ab}}{\mathbf{T}_{ab}}\,. \tag{9}$$

An approximate Newton step is then [12]:

$$\Delta w_{kl} \propto -\sum_{ab}\mathbf{M}_{(kl)(ab)}\frac{\partial\log|\mathbf{T}|}{\partial w_{ab}} = w_{kl}\left[1-\frac{\sum_a \mathbf{T}_{al}}{\mathbf{T}_{kl}}\right]\,. \tag{10}$$

In the simple Spike Response Model, this gives a compellingly simple learning rule:

$$\Delta w_{kl} \propto w_{kl} - \frac{\dot{u}_k}{\dot{R}_{kl}}\sum_a \mathbf{T}_{al}\,, \tag{11}$$

where the remaining non-local term is a sum of the sensitivities, $\mathbf{T}_{al} = -w_{al}\dot{R}_{al}/\dot{u}_a$, to spike t_l, of the other post-synaptic neurons when they spike, at times t_a. Many non-local effects perhaps similar to this have been observed when STDP is induced at a single synapse [6,10].

The resulting learning rule is shown for different delays $\Delta t = t_k - t_l$ in Figure 1c. It resembles the biological learning rule with one important difference. The sharp transition (singularity) of this rule is at the maximum of the EPSP, i.e. some $\Delta t > 0$. In the physiological rule the singularity is at $\Delta t = 0$.

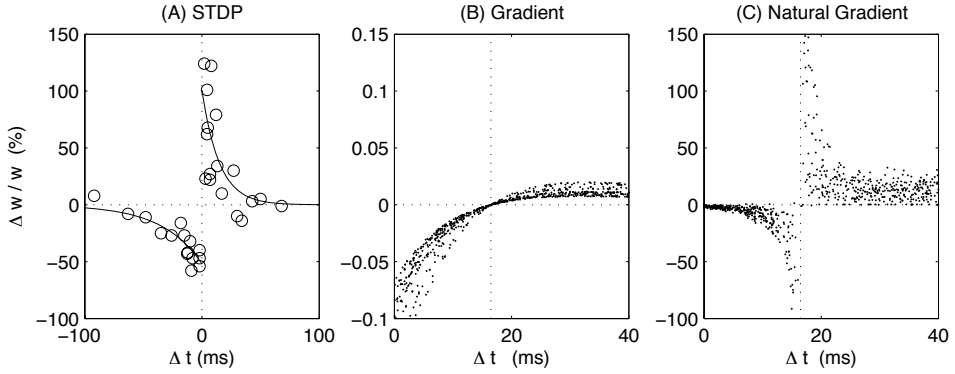

Figure 1: Spike time dependent plasticity (STDP) for an excitatory synapse. (A) Change in synaptic strength observed in the visual cortical slice in response to pre- and postsynaptic spike pairs separated by Δt (adapted with permission from [7]). (B) Maximum information gradient according to (8). (C) Approximate Newton update according to (10). The gradients are shown in an arbitrary scale. For graphs (B) and (C) we used random spikes with $\Delta t = t_k - t_l$. We ignored the effect of previous spikes t_{k-1} and dynamic threshold, i.e. we used T_{kl} according to equation (5) and the spike response $R_{kl} = R(t_k - t_l,)$ as shown in Figure 2. We generated random t_k and t_l with $t_l < t_k$ and set $w_{kl} = 1$ for $t_l < t_k < t_l + 30$ms, and $w_{kl} = 0$ otherwise. We chose random positive $\dot{u}(t_k)$ with $0.5 \le \dot{u}(t_k) < 1$ to simulate the fact that threshold crossing happens preferentially at increased positive slopes. (Appropriate network simulations are in progress).

3 Discussion

In summary, we have written the rule for maximizing information transfer in a deterministic model network of spiking neurons. To get the qualitative similarities to experimental

observations, the rule must follow the natural gradient, which implies that parameter optimization in the brain may implicitly take place in a metric space that is optimal for speed and computational simplicity of learning.

The main dissimilarity with experiments, the temporal offset of the sharp transition from depression to potentiation, we believe to be a consequence of the difference between the spiking model and the real neuron, rather than a flaw in the objective function. The spike response model has a 'causal' direction from synapse to soma. The input spikes propagate to the cell body where the learning rule evaluates the coincidence involving the spike response and the somatic membrane potential. The capacitive membrane filtering of the spike along this path, we believe, is the source that offset. In contrast, in real neurons, the information flow is bidirectional (through linear diffusion, back-propagating action potentials, and other active membrane properties). In addition, the particular coincidence detection machinery known to be involved in LTP (the NMDA receptor complex) actually resides at the synapse. These considerations suggest that a more realistic biophysical model may optimize the information flow without producing such an offset. We are investigating this.

A second concern is that the learning rule has been derived assuming that the spiking network is a 'square' network with N input spikes and N output spikes. At fist glance this seems quite irrelevant to the brain, which has multiple cortical areas with irregular patterns of spiking and connectivity. However, an idea we are exploring is that the information flow is actually maximised across time, which permits us to consider a square Jacobian. To see this, imagine running the network for a very long time to produce a long vector of spike times, $\mathbf{t}$. Now create the vectors $\overleftarrow{\mathbf{t}}$ and $\overrightarrow{\mathbf{t}}$ by removing enough (say n) spikes from the end and beginning of $\mathbf{t}$, such that the Jacobian $\mathbf{T}^{fut}_{past} = \partial\overrightarrow{\mathbf{t}}/\partial\overleftarrow{\mathbf{t}}$ is square, non-singular, and has a log-determinant that depends on the weights. Basically, here, we are removing from consideration the sets of spikes which are not caused or do not cause, and a few more if these sets are not equal-sized. (The problem with just using the whole Jacobian, $\partial\mathbf{t}/\partial\mathbf{t}$, is that its determinant, being 1, does not depend on the weights.) We might be able to optimise 'predictive information' flow in this causal $\mathbf{T}^{fut}_{past}$-network, somewhat along the lines proposed by Crutchfield and Bialek, amongst others. This optimisation involves polysynaptic $\partial t_k/\partial t_l$ terms with multi-path dependencies, because a spike can be caused by another spike through several different synaptic chains. Maximising information flow across time in this slightly more complex matrix is an objective that is consistent with the memory and prediction abilities of natural nervous systems, and we believe it can be done using an appropriately augmented version of the learning rule we propose.

4 Appendix: Gradient and Natural Gradient of $\log|\mathbf{T}|$ for the full Spike Response Model.

As a model for a spiking neuron we will consider the *full* Spike Response Model introduced by Gerstner [8]. This is a fairly general model for which the integrate-and-fire model is a special case. In this model the effect of a presynaptic spike at time t_l on the membrane potential at time t is described by a post synaptic potential or spike response, which may also depend on the time that has passed since the last output spike t_{k-1}, hence the spike response is written as $R(t-t_{k-1}, t-t_l)$. This response is weighted by the synaptic strength w_l. Excitatory or inhibitory synapses are determined by the sign of w_l. Refractoriness is incorporated by adding a hyper-polarizing contribution (spike-afterpotential) to the membrane potential in response to the last preceding spike $\eta(t-t_{k-1})$. The membrane potential as a function of time is therefore given by

$$u(t) = \eta(t-t_{k-1}) + \sum_l w_l R(t-t_{k-1}, t-t_l)\,. \tag{12}$$

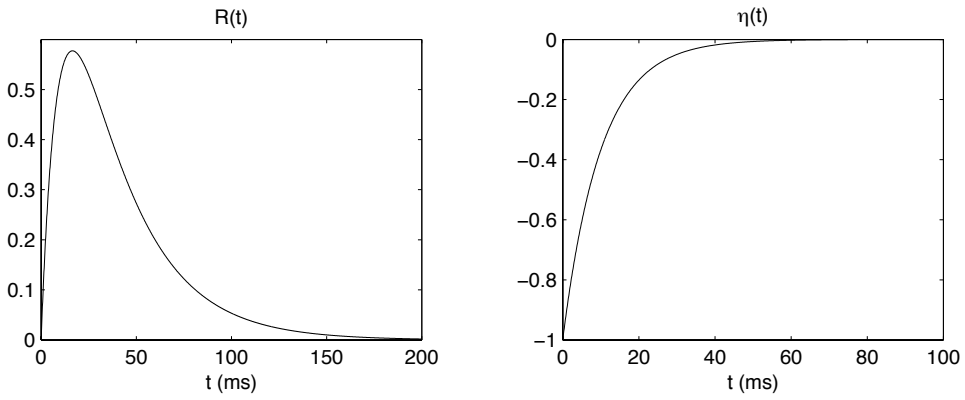

Figure 2: Spike response and refractory hyper-polarization. The spike response shown to the left is given by, $R(t) = (1 - \tau_s/\tau_m)^{-1}(e^{-t/\tau_m} - e^{-t/\tau_s})\Theta(t)$, where $\Theta(t)$ is the unit step function. This spike response corresponds to a EPSP of a capacitively integrating neuron with time constant τ_m and an α-function for the post-synaptic current with time constant τ_s. We used $\tau_m = 30ms$ and $\tau_s = 10ms$. The spike-afterpotential shown on the right, $\eta(t) = -e^{t/\tau_m}\Theta(t)$, mimics the effect of a reset in an leaky integrate-and-fire model.

We have ignored here potential contributions from external currents which can easily be included without modifying the following derivations. The output firing times t_k are defined as the times for which $u(t)$ reaches firing threshold from below. We consider a dynamic threshold, $\vartheta(t - t_{k-1})$, which may depend on the time since that last spike t_{k-1}, together then output spike times are defined implicitly by:

$$t = t_k : u(t) = \vartheta(t - t_{k-1}) \text{ and } \frac{du(t)}{dt} > 0\,. \tag{13}$$

For this more general model $\mathbf{T}_{kl}$ is given by

$$\mathbf{T}_{kl} = \frac{dt_k}{dt_l} = -\left(\frac{\partial u}{\partial t_k} - \frac{\partial \vartheta}{\partial t_k}\right)^{-1}\frac{\partial u}{\partial t_l} = \frac{w_{kl}\dot{R}(t_k - t_{k-1}, t_k - t_l,)}{\dot{u}(t_k) - \dot{\vartheta}(t_k - t_{k-1})}\,, \tag{14}$$

where $\dot{R}(s,t), \dot{u}(t)$, and $\dot{\vartheta}(t)$ are derivatives with respect to t. The dependence of $\mathbf{T}_{kl}$ on t_{k-1} should be implicitly assumed. It has been omitted to simplify the notation.

Now we compute the derivative of $\log|\mathbf{T}|$ with respect to w_{kl}. For any matrix $\mathbf{T}$ we have $\partial \log|\mathbf{T}|/\partial \mathbf{T}_{ab} = \mathbf{T}^{-1}_{ba}$. Therefore:

$$\frac{\partial \log|\mathbf{T}|}{\partial w_{kl}} = \sum_{ab}\frac{\partial \log|\mathbf{T}|}{\partial \mathbf{T}_{ab}}\frac{\partial \mathbf{T}_{ab}}{\partial w_{kl}}\sum_{ab}\mathbf{T}^{-1}_{ba}\frac{\partial \mathbf{T}_{ab}}{\partial w_{kl}}\,. \tag{15}$$

Utilising the Kronecker delta $\delta_{ab} = (1 \text{ if } a = b, \text{ else } 0)$, the derivative of (14) with respect to w_{kl} gives:

$$\begin{aligned}
\frac{\partial \mathbf{T}_{ab}}{\partial w_{kl}} &= \frac{\partial}{\partial w_{kl}}\left[\frac{w_{ab}\dot{R}(t_a - t_{a-1}, t_a - t_b)}{\eta(t_a - t_{a-1}) + \sum_c w_{ac}\dot{R}(t_a - t_{a-1}, t_a - t_c) - \dot{\vartheta}(t_a - t_{a-1})}\right] \\
&= \delta_{ak}\delta_{bl}\frac{\dot{R}(t_a - t_{a-1}, t_a - t_b)}{\dot{u}(t_a) - \dot{\vartheta}(t_a - t_{a-1})} \\
&\quad - \frac{w_{ab}\dot{R}(t_a - t_{a-1}, t_a - t_b)\delta_{ak}\dot{R}(t_a - t_{a-1}, t_a - t_l)}{\left(\dot{u}(t_a) - \dot{\vartheta}(t_a - t_{a-1})\right)^2}
\end{aligned}$$

$$= \quad \delta_{ak}\mathbf{T}_{ab}\left[\frac{\delta_{bl}}{w_{ab}} - \frac{\mathbf{T}_{al}}{w_{al}}\right] . \tag{16}$$

Therefore:

$$\frac{\partial \log|\mathbf{T}|}{\partial w_{kl}} = \sum_{ab}\mathbf{T}^{-1}_{ba}\delta_{ak}\mathbf{T}_{ab}\left[\frac{\delta_{bl}}{w_{ab}} - \frac{\mathbf{T}_{al}}{w_{al}}\right] \tag{17}$$

$$= \frac{\mathbf{T}_{kl}}{w_{kl}}\left[\mathbf{T}^{-1}_{lk} - \sum_{b}\mathbf{T}^{-1}_{bk}\mathbf{T}_{kl}\right] = \frac{\mathbf{T}_{kl}}{w_{kl}}\left[\mathbf{T}^{-1}_{lk} - 1\right] . \tag{18}$$

The Hessian of $\log|\mathbf{T}|$ can be computed by taking derivatives of (18):

$$\mathbf{H}_{(ab)(kl)} = \frac{\partial \log|\mathbf{T}|}{\partial w_{ab}\partial w_{kl}} = \frac{\partial}{\partial w_{ab}}\left(\frac{\mathbf{T}_{kl}}{w_{kl}}\left[\mathbf{T}^{-1}_{lk} - 1\right]\right) . \tag{19}$$

We use now:

$$\frac{\partial}{\partial w_{ab}}\frac{\mathbf{T}_{kl}}{w_{kl}} = -\delta_{ak}\frac{\mathbf{T}_{kl}}{w_{kl}}\frac{\mathbf{T}_{ab}}{w_{ab}} . \tag{20}$$

which can be derived similarly to the second term in (16) and obtain:

$$\mathbf{H}_{(ab)(kl)} = -\delta_{ak}\frac{\mathbf{T}_{kl}}{w_{kl}}\frac{\mathbf{T}_{ab}}{w_{ab}}\left[\mathbf{T}^{-1}_{lk} - 1\right] + \frac{\mathbf{T}_{kl}}{w_{kl}}\sum_{nm}\frac{\partial \mathbf{T}^{-1}_{lk}}{\partial \mathbf{T}_{nm}}\frac{\partial \mathbf{T}_{nm}}{\partial w_{ab}} . \tag{21}$$

Insert now (16) and use for the inverse:

$$\frac{\partial \mathbf{T}^{-1}_{lk}}{\partial \mathbf{T}_{nm}} = -\mathbf{T}^{-1}_{ln}\mathbf{T}^{-1}_{mk} , \tag{22}$$

which is true for any invertible matrix $\mathbf{T}_{kl}$. This gives:

$$\mathbf{H}_{(ab)(kl)} = -\delta_{ak}\frac{\mathbf{T}_{kl}}{w_{kl}}\frac{\mathbf{T}_{ab}}{w_{ab}}\left[\mathbf{T}^{-1}_{lk} - 1\right] - \frac{\mathbf{T}_{kl}}{w_{kl}}\sum_{nm}\mathbf{T}^{-1}_{ln}\mathbf{T}^{-1}_{mk}\delta_{na}\mathbf{T}_{nm}\left[\frac{\delta_{mb}}{w_{nm}} - \frac{\mathbf{T}_{nb}}{w_{nb}}\right] . \tag{23}$$

The first term of the first parenthesis cancels with the second term in the second parenthesis after executing the sums, and we obtain finally:

$$\mathbf{H}_{(ab)(kl)} = \frac{\mathbf{T}_{kl}}{w_{kl}}\frac{\mathbf{T}_{ab}}{w_{ab}}\left[\delta_{ak} - \mathbf{T}^{-1}_{la}\mathbf{T}^{-1}_{bk}\right] . \tag{24}$$

We show now that $\mathbf{M}$ defined in (9) is an approximate inverse of $\mathbf{H}$. The product $\mathbf{MH}$ gives:

$$(\mathbf{MH})_{(xy)(kl)} = \sum_{ab}\mathbf{M}_{(xy)(ab)}\mathbf{H}_{(ab)(kl)} = \delta_{xk}\delta_{yl} - \mathbf{T}_{ky}\sum_{b}\mathbf{T}_{xb} . \tag{25}$$

The first term is the required identity. We will now argue that in the average and for a large number of input spikes this is:

$$\left\langle(\mathbf{MH})_{(xy)(kl)}\right\rangle \approx \delta_{xk}\delta_{xy} - 1/N . \tag{26}$$

where N is the number of input spikes, and $\langle . \rangle$ is the expectation. For large N the second term vanishes and $\langle\mathbf{MH}\rangle$ becomes the identity. We argue this as follows: Assume for simplicity $w_{ky} = const_k$, then $\mathbf{T}_{ky} = \dot{R}(t_k - t_{k+1}, t_k - t_y)/(\dot{\theta}(t_k - t_{k-1}) + \sum_l \dot{R}(t_k - t_{k-1}, t_k - t_l))$, with $\theta(s) = \eta(s) - \vartheta(s)$. For simplicity we also assume that the last output spike t_{k-1} occurred quite some time ago, i.e. $s \to \infty$. For most reasonable choices of η and ϑ wE have $\dot{\theta}(\infty) = 0$. In this case, $\sum_b \mathbf{T}_{xb} = 1$. Assume further that we have N input spikes with spikes times t_l arriving independently and randomly. Then $\dot{R}(\infty, t_k - t_y)$ are independent random numbers with a possible mean r_k. For large N, the denominator, $\sum_l \dot{R}(\infty, t_k - t_l)$, is approximately Nr_k (law of large numbers). In the average over all input spikes $\left\langle \dot{R}(\infty, t_k - t_y)\right\rangle = r_k$. We have then $\langle \mathbf{T}_{ky}\sum_b \mathbf{T}_{xb}\rangle = r_k/(Nr_k) = 1/N$.

Acknowledgments

We are grateful for inspirational discussions with Nihat Ay (on temporal infomax), with Michael Eisele (on STDP), with Hong Hui Yu (on Ansatz (3) which explained our formula (14)), and with Arunava Banerjee, who helped stimulate the work. AJB thanks RNI colleagues for many such discussions.

References

[1] Amari S-I. 1997. Natural gradient works efficiently in learning, *Neural Computation*, 10, 251-276

[2] Ay N. & Wennekers T. 2003. Dynamical properties of strongly interacting Markov chains, *Neural networks*, 16: 1483-97

[3] Banerjee A. 2001. On the Phase-Space Dynamics of Systems of Spiking Neurons. *Neural Computation*, 13, 161-225

[4] Bell A.J. & Sejnowski T.J. 1995. An information maximization approach to blind separation and blind deconvolution, *Neural Computation*, 7, 1129-1159

[5] Bell A.J. & Sejnowski T.J. 1997. The 'Independent Components' of natural scenes are edge filters, *Vision Research*, 37: 3327-3338

[6] Bi G.Q. & Poo M.M. 1998. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength and post-synaptic cell type, *J. Neurosci.*, 18, 24, 10464-10472, see also the review: *Annu. Rev. Neurosci.*, 24: 139-166

[7] Froemke R.C. & Dan Y. 2002. Spike-timing-dependent synaptic modification induced by natural spike trains. *Nature*, 28, 416: 433-8

[8] Gerstner W. & Kistner W.M. 2002. *Spiking neuron models*, Camb. Univ. Press

[9] Hyvärinen A., Hurri J., and Väyrynen J. 2003. Bubbles: a unifying framework for low-level statistical properties of natural image sequences. *J. Optical Soc. Am. A*, 20,7:1237-1252

[10] Li C, Lu J, Wu C, Duan S, Poo M. 2004: Bidirectional modification of presynaptic neuronal excitability accompanying spike timing-dependent synaptic plasticity. *Neuron*, 41(2):257-68.

[11] Markram H., Lubke J., Frotscher M. & Sakmann B. 1997. Regulation of synaptic efficacy by coincidences of post-synaptic action potentials and EPSPs, *Science*, 275: 213

[12] MacKay D.M. 2003. *Information Theory, Inference, and Learning Algorithms*, Camb. Univ. Press