Proc Natl Acad Sci USA 102(14):5239-44 (2005).
Generalized Bienenstock-Cooper-Munro rule for spiking neurons that maximizes information transmission.
Taro Toyoizumi, Jean-Pascal Pfister, Kazuyuki Aihara, and Wulfram Gerstner

Supporting Text

Evolution of the Average Synaptic Update Rule

In this appendix we evaluate the derivative of Eq. **9** in the main text, i.e., we need to calculate

$$\frac{\partial}{\partial w_j}\left\langle \log\frac{P(y^k|Y^{k-1},X^k)}{P(y^k|Y^{k-1})} - \gamma\log\frac{P(y^k|Y^{k-1})}{\tilde{P}(y^k|Y^{k-1})}\right\rangle_{\boldsymbol{Y}^k,\boldsymbol{X}^k}. \tag{1}$$

Before we start let us recall some notation. The average of an arbitrary function f_w with arguments x and y is by definition

$$\langle f_w(x,y)\rangle_{\mathbf{x},\mathbf{y}} = \sum_x\sum_y p_w(x,y)\, f_w(x,y) \tag{2}$$

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

By definition, we have $p_w(x,y) = p_w(y|x)p(x)$ where $p(x)$ is a given input distribution and $p_w(y|x)$ the (parameter-dependent) conditional probability of generating an output y given x. Hence Eq. **2** can be transformed into

$$\langle f_w(x,y)\rangle_{\mathbf{x},\mathbf{y}} = \sum_x p(x)\sum_y p_w(y|x)\, f_w(x,y) = \left\langle \sum_y p_w(y|x)\, f_w(x,y)\right\rangle_{\mathbf{x}} \tag{3}$$

If we now take the derivative with respect to the parameter w, the product rule yields two terms

$$\begin{aligned}\frac{\partial}{\partial w}\langle f_w(x,y)\rangle_{\mathbf{x},\mathbf{y}} &= \left\langle \sum_y p_w(y|x)\,\frac{\partial}{\partial w} f_w(x,y)\right\rangle_{\mathbf{x}} \\ &\quad + \left\langle \sum_y p_w(y|x)\,\left[\frac{\partial}{\partial w}\log p_w(y|x)\right]\, f_w(x,y)\right\rangle_{\mathbf{x}}\end{aligned} \tag{4}$$

The first term contains the derivative of the function f_w whereas the second term contains the derivative of the conditional probability p_w. We note that Eq. **4** can also be written in the form

$$\frac{\partial}{\partial w}\langle f_w(x,y)\rangle_{\mathbf{x},\mathbf{y}} = \left\langle \frac{\partial}{\partial w} f_w(x,y)\right\rangle_{\mathbf{x},\mathbf{y}} + \left\langle \left[\frac{\partial}{\partial w}\log p_w(y|x)\right]\, f_w(x,y)\right\rangle_{\mathbf{x},\mathbf{y}}, \tag{5}$$

i.e., as an average over the joint distribution of x and y. This formulation will be useful for the problem at hand.

The gradient in Eq. **1** contains several terms and for the moment we pick only one of these. The others will then be treated analogously. Let us focus on the term $\left\langle \log P(y^k|Y^{k-1}, X^k) \right\rangle_{\boldsymbol{Y}^k, \boldsymbol{X}^k}$ and apply steps completely analogous to those leading from Eqs. **2**-**5**.

$$\begin{aligned} &\frac{\partial}{\partial w_j} \left\langle \log P(y^k|Y^{k-1}, X^k) \right\rangle_{\boldsymbol{Y}^k, \boldsymbol{X}^k} \\ &= \left\langle \frac{\partial}{\partial w_j} \log P(y^k|Y^{k-1}, X^k) \right\rangle_{\boldsymbol{Y}^k, \boldsymbol{X}^k} \\ &\quad + \left\langle \left[\frac{\partial}{\partial w_j} \log P(Y^k|X^k) \right] \log P(y^k|Y^{k-1}, X^k) \right\rangle_{\boldsymbol{Y}^k, \boldsymbol{X}^k} \end{aligned} \tag{6}$$

We now evaluate the averages using the identity $\langle \cdot \rangle_{\boldsymbol{Y}^k, \boldsymbol{X}^k} = \langle \langle \cdot \rangle_{\mathbf{y}^k|Y^{k-1}, X^k} \rangle_{\boldsymbol{Y}^{k-1}, \boldsymbol{X}^k}$. We find that the first term on the right-hand side of Eq. **6** vanishes, since

$$\begin{aligned} &\left\langle \frac{\partial}{\partial w_j} \log P(y^k|Y^{k-1}, X^k) \right\rangle_{\mathbf{y}^k|Y^{k-1}, X^k} \\ &= \sum_{y^k \in \{0,1\}} \frac{\partial}{\partial w_j} \left[\log P(y^k|Y^{k-1}, X^k) \right] P(y^k|Y^{k-1}, X^k) \\ &= \frac{\partial}{\partial w_j} \left[\sum_{y^k \in \{0,1\}} P(y^k|Y^{k-1}, X^k) \right] = 0 \end{aligned} \tag{7}$$

because of the normalization of probabilities. The same argument can be repeated to show that $0 = \left\langle \frac{\partial}{\partial w_j} \log P(y^k|Y^{k-1}) \right\rangle_{\mathbf{y}^k|Y^{k-1}, X^k}$. The reference distribution $\tilde{P}(y^k|Y^{k-1})$ is by definition independent of w_j.

Hence the only term that gives a non-trivial contribution on the right-hand side of Eq. **6** is the second term. With an analogous argument for the other factors in Eq. **1** we have

$$\begin{aligned} &\frac{\partial}{\partial w_j} \left\langle \log \frac{P(y^k|Y^{k-1}, X^k)}{P(y^k|Y^{k-1})} - \gamma \log \frac{P(y^k|Y^{k-1})}{\tilde{P}(y^k|Y^{k-1})} \right\rangle_{\boldsymbol{Y}^k, \boldsymbol{X}^k} \\ &= \left\langle \left[\frac{\partial \log P(Y^k|X^k)}{\partial w_j} \right] \left(\log \frac{P(y^k|Y^{k-1}, X^k)}{P(y^k|Y^{k-1})} - \gamma \log \frac{P(y^k|Y^{k-1})}{\tilde{P}(y^k|Y^{k-1})} \right) \right\rangle_{\boldsymbol{Y}^k, \boldsymbol{X}^k} \end{aligned} \tag{8}$$

An identification of the factors C, F, and G in the main text is straightforward. From Eq. **4** in the main text we have

$$\log P(y^k|Y^{k-1}, X^k) = y^k \log(\rho^k) + (1 - y^k) \log(1 - \rho^k) \tag{9}$$

Hence we can evaluate the factors

$$\begin{aligned} F^k &= \log \frac{P(y^k|Y^{k-1}, X^k)}{P(y^k|Y^{k-1})} = y^k \log \frac{\rho^k}{\bar{\rho}^k} + (1-y^k) \log \frac{1-\rho^k}{1-\bar{\rho}^k} \\ G^k &= \log \frac{P(y^k|Y^{k-1})}{\tilde{P}(y^k|Y^{k-1})} = y^k \log \frac{\bar{\rho}^k}{\tilde{\rho}} + (1-y^k) \log \frac{1-\bar{\rho}^k}{1-\tilde{\rho}} \end{aligned}$$

Furthermore we can calculate the derivative needed in Eq. **8** using the chain rule from Eq. **6** of the main text, i.e.,

$$P(Y^k|X^k) = \prod_{l=1}^{k} P(y^l|Y^{l-1}, X^l) \tag{10}$$

which yields

$$\begin{aligned} \frac{\partial \log P(Y^k|X^k)}{\partial w_j} &= \frac{\partial}{\partial w_j} \sum_{l=1}^{k} \log P(y^l|Y^{l-1}, X^l) &(11)\\ &= \sum_{l=1}^{k} \left[\frac{y^l}{\rho^l} - \frac{1-y^l}{1-\rho^l} \right] \rho'^l \sum_n \epsilon(t^l - t^n)\, x_j^n &(12) \end{aligned}$$

We note that in Eq. **8** the factor $\frac{\partial}{\partial w_j} \log P(Y^k|X^k)$ has to be multiplied with F^k or with G^k before taking the average. Multiplication generates terms of the form $\langle y^l\, y^k \rangle_{\boldsymbol{Y}^k, \boldsymbol{X}^k} = \langle \langle y^l\, y^k \rangle_{\boldsymbol{Y}^k|X^k} \rangle_{\boldsymbol{X}^k}$ For any given input X^k, the autocorrelation $\langle y^l\, y^k \rangle_{\boldsymbol{Y}^k|X^k}$ with $l < k$ of the postsynaptic neuron will have a trivial value

$$\langle y^l\, y^k \rangle_{\boldsymbol{Y}^k|X^k} = \langle y^l \rangle_{\boldsymbol{Y}^k|X^k} \langle y^k \rangle_{\boldsymbol{Y}^k|X^k} \quad \text{for} \quad k - l > k_a \tag{13}$$

where $k_a\, \Delta t$ is the width of the autocorrelation. As a consequence

$$\left\langle \left[\frac{y^l}{\rho^l} - \frac{1-y^l}{1-\rho^l} \right] \left(F^k - \gamma G^k \right) \right\rangle_{\boldsymbol{Y}^k, \boldsymbol{X}^k} = 0 \quad \text{for} \quad k - l > k_a \tag{14}$$

Hence, for $k > k_a$, we can truncate the sum over l in Eq. **8**, i.e., $\sum_{l=1}^{k} \to \sum_{l=k-k_a}^{k}$ which yields exactly the coincidence measure C_j introduced in the main text; cf. Eq. **11** in the main text, and which we repeat here for convenience

$$C_j^k = \sum_{l=k-k_a}^{k} \left[\frac{y^l}{\rho^l} - \frac{1-y^l}{1-\rho^l}\right] \rho'^l \sum_n \epsilon(t^l - t^n)\, x_j^n \tag{15}$$

From Averages to an Online Rule

The coincidence measure C_j^k counts coincidences in a rectangular time window. If we replace the rectangular time window by an exponential one with time constant τ_C and go to continuous time, the summation $\sum_{l=k-k_a}^{k} \dots$ in Eq. **15** turns into an integral $\int_{-\infty}^{t} dt' \exp[-(t-t')/\tau_C] \dots$ which can be transformed into a differential equation

$$\frac{dC_j(t)}{dt} = -\frac{C_j(t-\delta)}{\tau_C} + \sum_f \epsilon(t - t_j^{(f)}) S(t) \big[\delta(t - \hat{t} - \delta) - g(u(t)) R(t)\big] \;; \tag{16}$$

cf. Eq. **15** in the main text. Based on the considerations in the previous paragraph, the time constant τ_C should best be chosen in the range $k_a \Delta t \leq \tau_C \leq 10\, k_a \Delta t$.

Similarly, the average firing rate $\bar{\rho}(t) = \bar{g}(t)\, R(t)$ can be estimated using a running average

$$\tau_{\bar{g}} \frac{d\bar{g}(t)}{dt} = -\bar{g}(t) + g(u(t)) \tag{17}$$

with time constant $\tau_{\bar{g}}$.

In Fig. 6, we compare the performance of three different update schemes in numerical simulations. In particular, we show that (i) the exact value of the truncation of the sum in Eq. **15** is not relevant, as long as $k_a \Delta t$ is larger than the width of the autocorrelation; and (ii) that the online rule is a good approximation to the exact solution.

To do so we take the scenario from Fig. 3 of the main text. For each segment of 1 s, we simulate one hundred pairs of input and output spike trains. We evaluate numerically Eq. **8** by averaging over the 100 samples. After each segment of 1 second (=1,000 time steps) we update the weights

using a rule without truncation in the sum of Eq. **15**. We call this the full batch update; compare. Fig. 6 (*Top*).

Second, we use the definition of C_j^k with the truncated sum and repeat the above steps; Fig. 6 (*Middle*). The truncation is set to $k_a \, \Delta t = 200$ms which is well above the expected width of the autocorrelation function of the postsynaptic neuron. We call this the truncated batch rule.

Third, we use the online rule discussed in the main body of the paper with $\tau_C = 1$s; Fig. 6 (*Bottom*).

Comparison of top and center graphs of Fig. 6 shows that there is no difference in the evolution of mean synaptic efficacies, i.e., the truncation of the sum is allowed, as expected from the theoretical arguments. A further comparison with Fig. 6 *Bottom* shows that updates based on the online rule add some fluctuations to the results, but its trend captures nicely the evolution of the batch rules.

Supplement to the Pattern Detection Paradigm

In Fig. 3 we presented a pattern detection paradigm where patterns defined by input rates were chosen randomly and applied for one second. After learning, the spike count over one second is sensitive to the index of the pattern. Fig. 7*A* shows the histogram of spike counts for each pattern. Optimal classification is achieved by choosing for each spike count the pattern which is most likely. With this criterion 81 percent of the patterns will be classified correctly.

The update of synaptic efficacies depends on the choice of the parameter γ in the learning rule. According the the optimality criterion in Eq. **8** of the main text, a high level of γ implies a strong homeostatic control of the firing rate of the postsynaptic neuron whereas a low level of γ induces only a weak homeostatic control. In order to study the role of γ, we repeated the numerical experiments for the above pattern detection paradigm with a value of $\gamma = 100$ instead of our standard value of $\gamma = 1$. Fig. 7*B* shows that the output firing rate is still modulated by the pattern index, the modulation at $\gamma = 100$ is, however, weaker than that at $\gamma = 1$. As a results, pattern detection is less reliably with 45 percent correct classification only. We note that this is still significantly higher than the chance level of 25 percent.

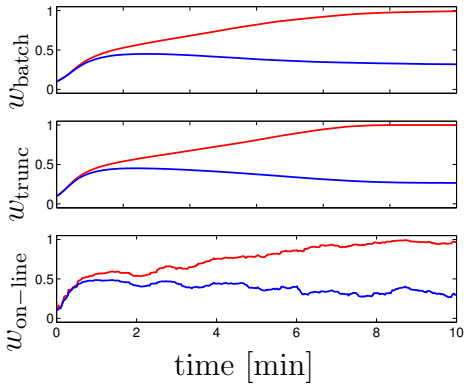

Figure 6: Evolution of the synaptic efficacies for the pattern detection paradigm of Fig. 3 during the first 10 minutes of simulated time. Red: mean synaptic efficacy of the 25 synapses that received pattern-dependent input rates. Blue: mean synaptic efficacy of the remaining 75 synapses. The batch update rule (top), the truncated batch rule (middle) and the online rule (bottom) yield comparable results.

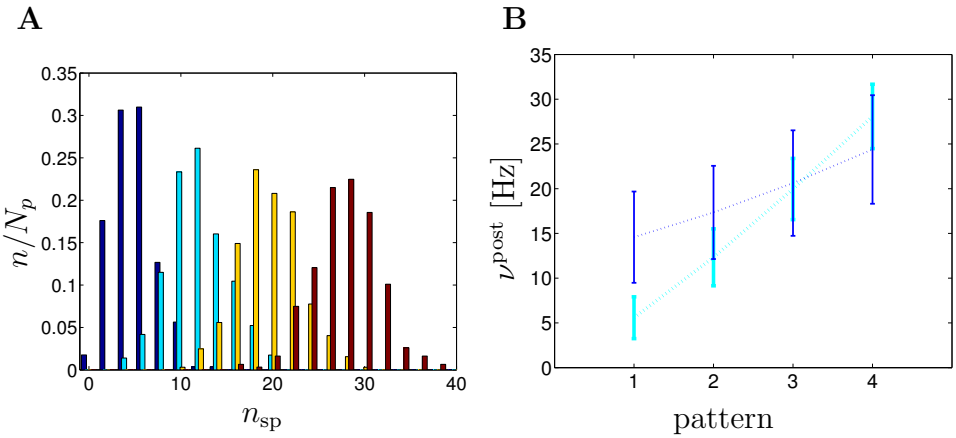

Figure 7: Pattern detection. **A** Histograms of spike counts n_{sp} over one second (horizontal axis, bin size 2) during presentation of pattern 1 (dark blue), pattern 2 (light blue), pattern 3 (yellow), and pattern 4 (red). Vertical scale: number of trials n with a given spike count divided by total number N_p of trials for that pattern. **B** Spike count during one second (mean and variance) for each of the four patterns with a parameter value $\gamma = 1$ (light blue) and $\gamma = 100$ (dark blue). The values for $\gamma = 1$ are redrawn from Fig. 3.