Backward reasoning the formation rules

Walter Senn & João Sacramento

Synaptic plasticity during learning is as fundamental as it is hard to study. The underlying synaptic plasticity rule has now been inferred using only the firing rate statistics of visual neurons in monkeys before and after learning.

"There are few people however, who, if you told them a result, would be able to evolve from their own inner consciousness what the steps were which led up to that result," explains Sherlock Holmes to his adjunct Dr. John Watson in *A Study in Scarlet*. "[T]he grand thing," he carries on, "is to be able to reason backwards." Lim *et al.*[1], writing in this issue of *Nature Neuroscience*, reasoned backwards to deduce the synaptic plasticity rules from observed neuronal activities. By inspecting the neuronal responses in monkey visual cortex to novel and familiar stimuli, they inferred the synaptic plasticity rules that explain the response changes from novel to familiar stimuli.

As with Dr. Watson, one vaguely suspects that the recorded activities would be enough to uncover the underlying network and its formation. But only in the course of the authors' arguments do the proofs take shape that a specific synaptic plasticity rule causes the observed data. The analysis confirms what Bienenstock, Cooper and Munro predicted more than 30 years ago, albeit by pure theoretical reasoning at the time[2]. Today, after collecting new data on the neuronal responses in monkey visual cortex to novel and familiar stimuli during a passive and an active viewing task[3], Lim *et al.*[1] are able to draw further conclusions. They carefully disentangled the firing rate statistics of these neurons and inferred the changes in the synaptic strengths w between excitatory neurons of the infero-temporal cortex (ITC) that arise while viewing the same images thousands of times. They concluded that, roughly, the weights change according to

$$\Delta w \approx (r_{\text{post}} - \theta_{\text{post}})r_{\text{pre}} \qquad (1)$$

where r_{post} and r_{pre} are the post- and presynaptic firing rates and θ_{post} represents a dynamic plasticity threshold. Postsynaptic rates less than θ_{post} induce long-term depression; larger rates induce long-term potentiation (LTP). Stability considerations require that this threshold θ_{post} be highly correlated with r_{post}, although the data analysis does not tell whether this covariation of θ_{post} with r_{post} is delayed or not (we return to this point below). Moreover, most of the neurons respond with a rate lower than θ_{post}, so that the neuronal activity across the population is reduced. This may serve as a synaptic explanation for the phenomenon known as repetition suppression[4]. However, a small fraction of neurons responds with a rate larger than θ_{post}, and these neurons enhance their rate, leading to a sharpening of the neuronal representation[5] (**Fig. 1e**). Hence, the new analysis of the *in vivo* data now catches the postulated Bienenstock, Cooper and Munro rule *in flagrante*.

But how is it possible to infer the synaptic learning rule from only the activities of a few dozen putatively excitatory and inhibitory neurons recorded in a viewing task? To understand this, a mathematical excursion is in order. We consider the population of ITC neurons that respond to the visual stimuli, ordered from left to right by their average firing rates r (**Fig. 1b**). Let $p_{\text{nov}}(r)$ be the fraction of neurons responding with rate r to novel stimuli and $p_{\text{fam}}(r)$ be the corresponding fraction responding to familiar stimuli (**Fig. 1a**). With repeated stimulus presentations, the neurons change their firing rates by virtue of the synaptic plasticity: neurons reducing their rate move to the left, whereas neurons increasing their rate move to the right (**Fig. 1b**).

At some equilibrium rate r_o, the neurons neither move left nor right. The number of neurons that have lower rates than r_o (that is, are left of this point) must therefore be the same before and after the repeated stimulus presentation. In other words, the equilibrium point is at the crossing of the cumulative distribution functions for the novel and familiar stimuli, $P_{\text{nov}}(r)$ and $P_{\text{fam}}(r)$, and this defines the plasticity threshold θ_{post} (r_o, with the addition of a shift; see below) arising in the rule in equation (1) (**Fig. 1c**). Next, to calculate the rate changes induced by the stimulus presentations, one needs to look for the rate r_{fam} toward which a neuron that initially fires with r_{nov} is pushed; formally, this means transforming P_{nov} backwards with the inverse of P_{fam} to get the desired rate $r_{\text{fam}} = P_{\text{fam}}^{-1}\left(P_{\text{nov}}(r_{\text{nov}})\right)$. This rate change calculation assumes that the ordering of firing rates is preserved: that is, if a neuron responds on average to novel stimuli with a lower rate than another one, it will still do so for familiar stimuli. This order preservation allows us to assign a change $r_{\text{nov}} \to r_{\text{fam}}$ to an individual neuron that shifts its firing rate (**Fig. 1b**).

Finally, the rate changes need to be explained by a change in the excitatory-to-excitatory synaptic strengths; these synapses are believed to be the most plastic ones. For this, Lim *et al.*[1] back-engineered the current h that leads to a given firing rate r by assuming that many independent synaptic inputs produce a (truncated) Gaussian current distribution. The median of that current distribution is then aligned with the median of the firing rate distribution, and the current-to-rate transfer function $r = \phi(h)$ is read off. The change in the postsynaptic current then becomes $\Delta h(r) = \phi^{-1}\left(P_{\text{fam}}^{-1}\left(P_{\text{nov}}(r)\right)\right) - \phi^{-1}(r)$. By shifting this function $\Delta h(r)$ downward to compensate for the recurrent feedback and squeezing it by a factor that compensates for the presynaptic rate, one obtains the desired modification of the synaptic weight $\Delta w(r)$, expressed as a function of $r = r_{\text{post}}$ (**Fig. 1d**).

To test their findings, Lim *et al.*[1] constructed a recurrently connected network of rate-based excitatory and inhibitory neurons using the extracted postsynaptic dependency for the excitatory-to-excitatory weight changes. Because the overall firing rates decrease from novel to familiar stimuli, the plasticity rule extracted so far, however, would continuously decrease the average rate in the network. To compensate for this, the authors assume that the total excitatory synaptic strength onto a postsynaptic neuron remains constant during learning[6]. This invariance can be achieved by subtracting the average presynaptic firing rate $\overline{r_{\text{pre}}}$ in the presynaptic factor of the weight change. The final learning rule the authors infer then has the form

$$\Delta w = \sigma\left(r_{\text{post}} - \theta_{\text{post}}\right)\left(r_{\text{pre}} - \overline{r_{\text{pre}}}\right) \qquad (2)$$

where the function $\sigma(r_{\text{post}} - \theta_{\text{post}})$ acting on the deviation of r_{post} from θ_{post} represents the

Walter Senn and João Sacramento are in the Department of Physiology, University of Bern, Bern, Switzerland, and Walter Senn is also at the Center for Learning, Cognition and Memory, University of Bern, Bern, Switzerland.
e-mail: senn@pyl.unibe.ch or sacramento@pyl.unibe.ch

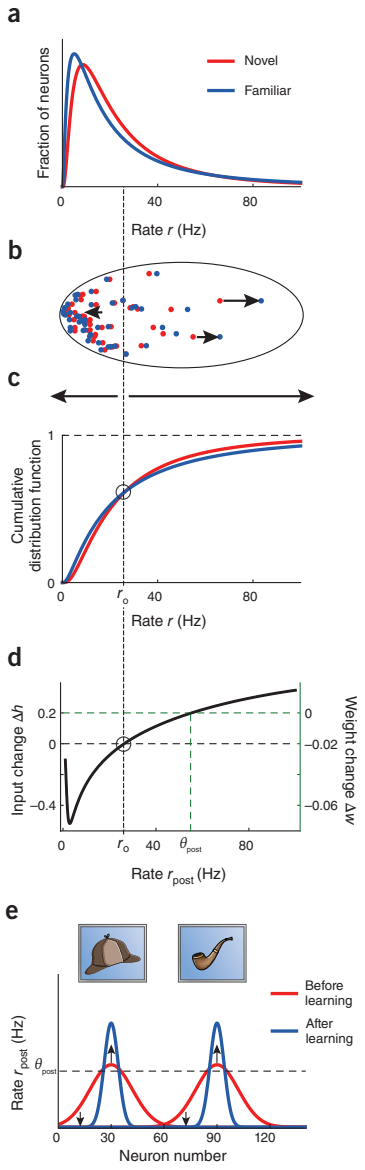

Figure 1 Inferring the plasticity rule from firing rate distributions. (**a**) Distribution of rate responses before (red) and after (blue) repeated stimulus presentations, shown here as log-normal to approximate the ITC recordings[1]. (**b**) Population of neurons (dots), aligned by their average response rates before (red) and after (blue) learning. (**c**) Neurons left of the intersection point (θ_{post}) of the two cumulative distribution functions decrease their rates; neurons to the right increase them (arrows). (**d**) Intuitively, the rate change induced by learning is the difference between the two cumulative distribution functions (red minus blue curve from **c**). The rate changes are converted into a change of input currents (Δh, left axis) via a rate-to-current transfer function. The weight change Δw for fixed presynaptic rate is equal to Δh downshifted and squeezed (right axis; equation (2)). (**e**) In a network (here of 130 neurons), the plasticity rule (equations (1) or (2)) sharpens the selectivity for recurring visual stimuli (top) by increasing firing rates that are above the threshold and decreasing those that are below.

postsynaptic factor shown in **Figure 1d**. This type of plasticity applied to the excitatory-to-excitatory synapses alone was able to reproduce the observed changes in the firing rates, both the differential change in the excitatory neurons and the overall firing rate reduction in the inhibitory neurons. The fact that the excitatory neurons with the largest firing rates enhance their rates even further while most other excitatory neurons decrease their rate (**Fig. 1e**) rules out, at least in the example chosen by Lim *et al.*[1], a dominant role of homeostatic inhibitory-to-excitatory synaptic plasticity[7].

What Lim *et al.*[1] offer is, in Holmes's words, a method to "observe and to draw inferences from our observations" that can now be applied to existing and future data. Lim *et al.*[1] have already applied it to recordings from a passive and active object viewing task, with the active task requiring attention to report subtle differences in luminance. Consistent with the common knowledge that active engagement improves learning[8], plasticity in ITC appears to be enhanced and shifted toward LTP for the active viewing task, although the details of how engagement modulates plasticity remain to be analyzed. As an intermediate result, the method also infers the current-to-rate transfer functions of the ITC neurons during the tasks. For the passive viewing task, the transfer functions look qualitatively similar to the ones obtained from *in vivo* recordings in cat visual cortex[9] or from *in vitro* recordings in rat barrel cortex[10]. It would be interesting to evaluate whether the ITC transfer functions increase their gain during active engagement[11].

So far the method has been applied to either novel or familiar stimuli without tracking intermediate familiarity levels. Such intermediate levels could give hints to how the plasticity rule changes in time and, in particular, how the plasticity threshold θ_{post} varies during learning. In the simulations, the threshold was fixed for a given pair of neurons, and presynaptic normalization (equation (2)) was introduced instead. But theoretical reasoning, both for rate-based[2] and spike-based[12] plasticity, predicts that it should slowly adapt to yield stimulus selectivity and stability. Yet the threshold may vary even on the fast, neuronal time scale, on the order of 10 ms, as it arises when interpreting θ_{post} as the dendritic prediction of somatic activity[13]. This would cast the Bienenstock, Cooper and Munro rule as an error-correcting rule, and the characteristic covariance reported by Lim *et al.*[1] of θ_{post} with the mean and even the s.d. of r_{post} would still be satisfied.

Now that we know how to access to the plasticity rule from activity distributions, inferring the connectivity pattern could soon be in our reach. With Dr. Watson, however, we may further need to take to heart the master's advice that "[i]n solving a problem of this sort, the grand thing is to be able to reason backwards"[14].

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Lim, S. *et al. Nat. Neurosci.* **18**, 1804–1810 (2015).
2. Bienenstock, E., Cooper, N. & Munro, W. *J. Neurosci.* **2**, 32–48 (1982).
3. Woloszyn, L. & Sheinberg, D.L. *Neuron* **74**, 193–205 (2012).
4. Summerfield, C., Trittschuh, E.H., Monti, J.M., Mesulam, M.M. & Egner, T. *Nat. Neurosci.* **11**, 1004–1006 (2008).
5. Freedman, D.J., Riesenhuber, M., Poggio, T. & Miller, E.K. *Cereb. Cortex* **16**, 1631–1644 (2006).
6. Bourne, J.N. & Harris, K.M. *Hippocampus* **21**, 354–373 (2011).
7. Vogels, T.P., Sprekeler, H., Zenke, F., Clopath, C. & Gerstner, W. *Science* **334**, 1569–1573 (2011).
8. Freeman, S. *et al. Proc. Natl. Acad. Sci. USA* **111**, 8410–8415 (2014).
9. Anderson, J.S., Lampl, I., Gillespie, D.C. & Ferster, D. *Science* **290**, 1968–1972 (2000).
10. Rauch, A., La Camera, G., Lüscher, H.-R., Senn, W. & Fusi, S. *J. Neurophysiol.* **90**, 1598–1612 (2003).
11. Maunsell, J.H. & Treue, S. *Trends Neurosci.* **29**, 317–322 (2006).
12. Gjorgjieva, J., Clopath, C., Audet, J. & Pfister, J.-P. *Proc. Natl. Acad. Sci. USA* **108**, 19383–19388 (2011).
13. Urbanczik, R. & Senn, W. *Neuron* **81**, 521–528 (2014).
14. Young, G. *Reasoning Backwards: How Sherlock Holmes Can Make You a Better Problem Solver* (Young Associates, 2008).