SI: Learning of chunking sequences in cognition and behavior

Jordi Fonollosa¹,², Emre Neftçi¹,³,*, Mikhail Rabinovich¹

1 Biocircuits Institute, University of California San Diego, La Jolla, USA
2 Institute for Bioengineering of Catalonia, Barcelona, Spain
3 Department of Cognitive Sciences, University of California Irvine, USA

These authors contributed equally to this work.

* nemre@ucsd.edu

1 Details to the learning rule Eq. (1)

The update terms are governed by an asymmetric temporal window $K_A(t)$:

$$A_{x_j}(t) = A^+ \int_{-\infty}^{t} K_A(t-s)x_j(s)ds$$
$$= A^+ \int_{0}^{\infty} K_A(\Delta)x_j(t-\Delta)d\Delta,$$

$$A_{x_i}(t) = A^- \int_{-\infty}^{t} K_A(s-t)x_i(s)ds$$
$$= A^- \int_{0}^{\infty} K_A(-\Delta)x_j(t-\Delta)d\Delta,$$

where $A^+ > 0$, $A^- > 0$ define the magnitude of the weight update.

For simplicity, we choose an exponential temporal window $K_A(\Delta) = \exp(-|\Delta/\tau_{STDP}|)$ with decay rate $\tau_{STDP} \ll T$. This rule is consistent with the requirement that $V_{ij}$ depotentiates when a transition from $x_i$ to $x_j$ occurs. As long as potentiation and depression are matched, this does not depend critically on this window, as we demonstrate below.

The condition that potentiation and depression are matched can be written:

$$\int_{0}^{\infty} K_A(\Delta)d\Delta = -\int_{0}^{\infty} K_A(-\Delta)d\Delta$$

(2)

We assume that the sign of $K_A(\Delta)$ is fixed at each side of the $\Delta = 0$ axis:

$$K_A(\Delta < 0) \leq 0, \ K_A(\Delta > 0) \geq 0$$

(3)

The state typically transitions sharply, such that $x_i(t)$ and $x_j(t)$ are monotonic around the transition times. For a transition, this can be written:

$$\frac{d}{dt} x_i(t) \leq 0, \ \frac{d}{dt} x_j(t) \geq 0, \forall t.$$ 

(4)

Under the above assumptions, we show that during a transition from $x_i$ to $x_j$, $V_{ij}$ depotentiates and $V_{ji}$ potentiates. The weight change is:

$$\tau_V \frac{d}{dt} V_{ij} = x_j(t) \int_{0}^{\infty} K_A(s)x_i(t-s)ds$$
$$+ x_j(t) \int_{-\infty}^{0} K_A(s)x_j(t+s)ds.$$
Fig. SI 1. A asymmetric learning window (left) causes the weight to change when a transition between two units take place (right).

With a change in sign in the second integral, the above equality can be written:

$$\tau_V \frac{d}{dt} V_{ij} = \int_{0}^{\infty} x_i(t-\Delta)x_j(t)K_A(\Delta) + x_i(t)x_j(t-\Delta)K_A(-\Delta)d\Delta.$$ 

Adding two terms that sum to zero under the integral:

$$\tau_V \frac{d}{dt} V_{ij} = \int_{0}^{\infty} x_i(t-\Delta)x_j(t)K_A(\Delta) - x_i(t)x_j(t)(K_A(\Delta) + K_A(-\Delta)) + x_i(t)x_j(t-\Delta)K_A(-\Delta)d\Delta.$$ 

The matching of the potentiation and depression in Eq. (2) guarantees that the middle terms vanishes.

The terms under the integral can be regrouped as follows:

$$\tau_V \frac{d}{dt} V_{ij} = \int_{0}^{\infty} x_j(t)(x_i(t-\Delta) - x_i(t))K_A(\Delta) + x_i(t)(x_j(t-\Delta) - x_j(t))K_A(-\Delta)d\Delta.$$ 

It is clear that, under the assumptions above (Eq. (2), (Eq. (3)) and (Eq. (4))), the integrand is positive or zero, leading to $\frac{d}{dt} V_{ij} \geq 0$. Similarly, $\frac{d}{dt} V_{ji} \leq 0$. Fig. 1 illustrates how the asymmetric learning windows causes the weight to change when a transition between two units takes place.

2 Network Dynamics Influence Chunking Rate

The chunking rate is defined as the number of transitions in the chunking layer while a pattern of the sequence is presented in the learning phase. This rate can be modulated, for example by biasing the chunking layer or its auxiliary variables $z_k$. To illustrate this, we added a global, step-wise varying input to the auxiliary variables $z_k$, and proceeded with the learning protocol similarly to the experiments in the main text (100 epochs). Results show that a larger number of chunks transition around the steps, and that the input magnitude drastically alters the chunking rate.

3 Learning with Noisy Stimuli

Noisy patterns $S'_k$ were obtained by adding noise to each pattern of the sequence:

$$S'_k[i] = S_k[i] + \max(0, \eta_k[i]), \quad i \in \mathbb{N}, k \in 1, \ldots, M$$

where $\eta_k[i] \sim N(0, \sigma_S)$, and $S_k[i]$ are the original patterns consisting of horizontal bars. The noise term changes from one presentation of the sequence to the other, but it remains constant during the presentation. In the main text, we report the amplitude of the noise as the ratio:

$$\text{Noise amplitude} = \frac{\sum_k \langle \eta_k[i] \rangle_S}{\sum_k S_k}$$

where $\langle \cdot \rangle_i$ is the expectation over realizations of $\max(0, \eta_k[i])$. Fig. 1 shows examples of the stimuli with noise amplitudes matching those used in the main text.
Fig. SI 2. **Chunking rate is modulated by a time-varying bias in the chunking layer.** (Top) We added a global, step-wise varying input $b_z$ to the auxiliary variables $z_k$. (Middle) Chunking rate computed as the number of transitions in the chunking layer during the presentation of each sequence element, averaged over 60 different runs of the training, and averaged over epochs 50 to 100. The average chunking rate was 0.071 from time 35 to 60, and 1.24 from time 95 to 120. Very few transitions occurred during the phase where $b_z$ was strongly positive, compared to chunking rate .314, when $b_z$ was zero. For strongly negative $b_z$, chunking is nearly absent, as the chunking layer transitions almost once every presentation of a sequence element (the chunking rate is close to 1). Furthermore, the chunking rate is high at the points where $b_z$ changes, which illustrates how the chunking has a tendency to synchronize with changes in $b_z$. (Bottom) Illustration of the activity in the chunking layer at trial 50 for all 60 runs. The boundaries of the chunks are clearly located at the time points where $b_z$ changed.

Fig. SI 3. **Examples of noisy stimuli,** drawn for noise parameters $\sigma_S = 0, .1, .3, .5$, with noise amplitudes estimated at 0%, 38%, 115% and 191%, respectively.

4 **Parameters of the learning model**

In Tab. 1, we detail all the parameters and values of the learning model so that the dynamics can be reproduced.
Fig. SI 4. An Example of Weight Evolution during Learning, for the run shown in Fig. 6 top right.
Table 1. Parameters of the hierarchical network

| Parameter                        | Description                                      | Figure(s)                  | Value(s)                  |
|----------------------------------|--------------------------------------------------|----------------------------|---------------------------|
| $N_e$                            | Number of Elementary Mode (EM)                    | Fig. 3, 4, 5               | 24                        |
| $N_y$                            | Number of Chunking Mode (CM)                      | All other figures          | 3                         |
| $\tau_x$                         | Time constant EM                                 | All figures                | 0.05 au                   |
| $\tau_y$                         | Time constant CM                                 | All figures                | 6.5                       |
| $\tau_z$                         | Synaptic time constant                           | All figures except Fig. 7 | 4.5                       |
| $b_x$                            | Growth term EM                                   | All figures, training      | 0                         |
| $b_y$                            | Growth term CM                                   | All figures, training      | 1                         |
| $b_t(t)$                         | Bias term synaptic states                        | All figures, recall        | 0.2                       |
| $C$                              | Total input magnitude of PMs                     | Training, all figures      | 15                        |
| $\tau_P$                         | Learning time constant $P$                       | All figures                | 10                        |
| $\tau_V$                         | Learning time constant $V$                       | All figures                | 28.6                      |
| $\tau_W$                         | Learning time constant $W$                       | All figures                | 125                       |
| $\tau_R$                         | Learning time constant $R$                       | All figures                | 333                       |
| $\tau_Q$                         | Learning time constant $Q$                       | All figures                | 125                       |
| $\epsilon_H$                     | Heterosynaptic competition                       | Fig. 2, 3, 4, 5            | 0                         |
| $m_H$                            | Total efferent weight from each EM               | Fig. 6, 7                  | 3                         |
| $\alpha_V$                       | Scaling of bistable term $V$                     | All figures                | 0.02                      |
| $\alpha_W$                       | Scaling of bistable term $W$                     | All figures                | 0.02                      |
| $\theta_V^d, \theta_W^d$         | Depotentiation threshold $V, W$                  | Fig. 3, 4, 5               | 0.6                       |
| $\theta_V^p, \theta_W^p$         | Potentiation threshold $V, W$                    | Fig. 3, 4, 5               | 0.6                       |
| $V(t=0), W(t=0)$                 | Initial cond. $V, W$                            | All figures                | off-diagonal 2.1, otherwise 1 |
| $V^+, W^+$                       | Positive boundary                                | Fig. 3, 4, 5               | 2.55                      |
| $V^-, W^-$                       | Negative boundary                                | Fig. 3, 4, 5               | 0.6                       |
| $V^*, W^*$                       | Boundary of the basins                           | Fig. 3, 4, 5               | 1.77                      |
| $\gamma_Q^d$                     | Depotentiation factor $Q$                        | All figures                | .75                       |
| $\gamma_Q^d$                     | Potentiation factor $Q$                          | All figures except Fig. 7  | .35                       |
| $\theta_Q^d, \theta_Q^p$         | Depotentiation threshold $Q$                     | All figures                | 0.2                       |
| $\theta_Q^d, \theta_Q^p$         | Potentiation threshold $Q$                       | All figures                | 0.42                      |
| $Q^+$                            | Positive boundary                                | All figures                | 1                         |
| $Q^-$                            | Negative boundary                                | All figures                | 0                         |
| $Q^*$                            | Boundary of the basins                           | All figures                | 0.4                       |
| $\alpha_Q$                       | Scaling of bistable term $Q$                     | All figures                | 2                         |
| $\alpha_R$                       | Scaling of bistable term $R$                     | All figures                | 0.02                      |
| $\theta_R^d, \theta_R^p$         | Depotentiation factor $R$                        | All figures                | 1.2                       |
| $\gamma_R^d, \gamma_R^p$         | Potentiation factor $R$                          | All figures                | 0.5                       |
| $\theta_R^d, \theta_R^p$         | Depotentiation threshold $R$                     | All figures                | 0.2                       |
| $\theta_R^d, \theta_R^p$         | Potentiation threshold $R$                       | All figures                | 0.25                      |
| $R^+$                            | Positive boundary                                | All figures                | 0.4                       |
| $R^-$                            | Negative boundary                                | All figures                | 0                         |
| $R^*$                            | Boundary of the basins                           | All figures                | 0.1                       |
| $\sigma_X$                       | Noise amplitude EM                               | All figures, training      | .02                       |
| $\sigma_V$                       | Noise amplitude CM                               | All figures, recall        | $10^{-6}$                 |
| $\sigma_Y$                       | Noise amplitude CM                               | All figures, recall        | $10^{-6}$                 |