Two is better than one: distinct roles for familiarity and recollection when retrieving palimpsest memories – supplementary text –

Cristina Savin[1]
cs664@cam.ac.uk

Peter Dayan[2]
dayan@gatsby.ucl.ac.uk

Máté Lengyel[1]
m.lengyel@eng.cam.ac.uk

[1]Computational & Biological Learning Lab, Dept. of Engineering, University of Cambridge, UK
[2]Gatsby Computational Neuroscience Unit, University College London, UK

1 Recall dynamics for single module

The Gibbs sampler dynamics assume that at each step the activity of one unit i is updated by sampling the probability $\mathrm{P}(x_i|x_{\setminus i}, \tilde{\mathbf{x}}, \mathbf{W})$. Since the neurons have binary activations, this is equivalent to sampling $\sigma(I)$, where $\sigma(x) = \frac{1}{1+e^{-x}}$, with the total current given as the log-odds ratio:

$$I_i \quad = \quad \log \frac{\mathrm{P}(x_i = 1|\mathbf{x}_{\setminus i}, \mathbf{W}, \tilde{x}_i)}{\mathrm{P}(x_i = 0|\mathbf{x}_{\setminus i}, \mathbf{W}, \tilde{x}_i)} = I_i^{\text{rec,in}} + I_i^{\text{rec,out}} + a\tilde{x}_i + b, \tag{1}$$

with parameters $a = 2\log\left(\frac{1-r}{r}\right)$, and $b = \log\left(\frac{fr}{(1-f)(1-r)}\right)$.

The contribution of the recurrent weights has the form:

$$I_i^{\text{rec}} = \sum_j \left(c_1 \cdot W_{ij}\, x_j + c_2 \cdot W_{ij} + c_3 \cdot x_j + c_4\right), \tag{2}$$

where the constants c_i can be computed as $c_1 = s_{11} + s_{00} - s_{01} - s_{10},\ c_2 = s_{10} - s_{00},\ c_3 = s_{01} - s_{00},\ c_4 = s_{00}$, with $s_{wx} = \log\left(\frac{\mathrm{P}(W_{ij}=w|x_i=1,x_j=x)}{\mathrm{P}(W_{ij}=w|x_i=0,x_j=x)}\right)$ for incoming weights and $s_{wx} = \log\left(\frac{\mathrm{P}(W_{ji}=w|x_i=1,x_j=x)}{\mathrm{P}(W_{ji}=w|x_i=0,x_j=x)}\right)$ for outgoing weights, respectively. These values are uniquely determined by the parameters defining the learning rule, n, the pattern distribution, f, and the average pattern age, $\bar{t}$.

2 Tempered transitions

The tempered transitions sampling procedure uses annealing, and systematically increases and decreases the temperature to ensure a better exploration of the parameter space [1]. To sample from distribution $\mathrm{P}\left(\mathbf{x}|\mathbf{W}, \tilde{\mathbf{x}}\right)$, one needs to define the intermediate probability distributions for a set of S (inverse) temperatures β_s. Here, we assume linear variations in inverse temperature, between $\beta_S = 1$, for the distribution of interest, and $\beta_0 = 0$; $\beta_s = s \cdot ds$, with $ds = \frac{1}{S}$ ($ds = 0.1$ for the results presented in the main text).

We assume the temperature modulates only the contribution of the recurrent weights to the posterior as:

$$\mathrm{P}_s\left(\mathbf{x}|\mathbf{W}, \tilde{\mathbf{x}}\right) \propto \mathrm{P}_{\text{store}}(\mathbf{x}) \cdot \mathrm{P}_{\text{noise}}(\tilde{\mathbf{x}}|\mathbf{x}) \cdot \left(\mathrm{P}(\mathbf{W}|\mathbf{x})\right)^{\beta_s}. \tag{3}$$

The transition operator for each step is a single component Gibbs sampler, $T_s(\mathbf{x}' \leftarrow \mathbf{x})$, with the reverse transition operator $\tilde{T}_s(\mathbf{x} \leftarrow \mathbf{x}')$ being the same single component Gibbs sampler (since the operator is reversible); the index of the component is selected at random for each temperature level during a cycle.

Given a current state $\mathbf{x}$ of the Markov chain, sampling proceeds by applying the sequence of transition operators $T_{S-1}...T_0\tilde{T}_0...\tilde{T}_{S-1}$, with the final state $\mathbf{x}'$ being accepted with a probability given by the ratios of probabilities of intermediate states [1].

3 The dual memory system

If we ignore the contribution of information that cannot be accessed across modules, $\mathbf{W}$, the posterior over pattern ages can be computed as:

$$\mathrm{P}\left(t|\mathbf{x}, \mathbf{W}^{\mathrm{fam}}\right) = \frac{1}{Z}\mathrm{P}\left(\mathbf{W}^{\mathrm{fam}}|\mathbf{x}, t\right), \tag{4}$$

where Z is the unknown partition function.

Using again the assumption that the weight distribution factorizes and taking the logarithm of the above expression we obtain the total input to a neuron: $I_i^{\mathrm{fam}} = \log \mathrm{P}(t = i|\mathbf{x}, \mathbf{W}^{\mathrm{fam}})$ which translates into a simple linear activation function:

$$\begin{aligned} I_i^{\mathrm{fam}} &= \sum_j \log \mathrm{P}(W_{ij}^{\mathrm{fam}}|x_i = 1, x_j, t) + \log \mathrm{P}(t) - \log(Z) & (5) \\ &= \sum_j \left[c_{1,i}^{\mathrm{fam}} W_{ij}^{\mathrm{fam}} x_j + c_{2,i}^{\mathrm{fam}} W_{ij}^{\mathrm{fam}} + c_{3,i}^{\mathrm{fam}} x_j + c_{4,i}^{\mathrm{fam}}\right] + \log \mathrm{P}(t) - \log(Z) & (6) \end{aligned}$$

with constants c_i^{fam} computed as before $c_{1,i}^{\mathrm{fam}} = s_{11}^i + s_{00}^i - s_{01}^i - s_{10}^i$, $c_{2,i}^{\mathrm{fam}} = s_{10} - s_{00}$, $c_{3,i}^{\mathrm{fam}} = s_{01}^i - s_{00}^i$, $c_{4,i}^{\mathrm{fam}} = s_{00}^i$, with $s_{wy}^i = \log\left(\mathrm{P}\left(W = w|x_i = 1, x_j, t = i\right)\right)$.

4 Default simulation parameters

We use balanced patterns, $f = 0.5$, the mean pattern age $\bar{t} = 25$, recall cue noise $r = 0.1$ (for recognition, $r = 0.001$, to be closer to previous models, e.g. [2], which assume the true $\mathbf{x}$ is used for familiarity detection). This translates into an average mutual information between a synaptic weight and the stored pattern of 0.0034 bits (this can be used as a palimpsest memory equivalent to the memory load traditionally used in Hopfield-like autoassociative memory networks). Performance is estimated in all cases by storing and recalling 250 patterns, with t distributed according to the prior. The error estimates are based on the posterior mean (minimizing mean-squared error) computed from 40000 samples (without burn-in, at least 10000 samples are needed for a reliable mean estimate). For the GP classifier, we use the GPML toolbox, with the logistic regression likelihood function; each class has 125 data points.

5 Beyond sampling-based representations

To derive the mean-field equivalent dynamics, we assume the N neurons in the network now have analog activations μ_i, which define a probability distribution over patterns $\mathbf{x}$ as a set of independent Bernoulli random variables. Neural dynamics during recall optimize the parameters $\boldsymbol{\mu}$ to bring this distribution as close as possible to the true posterior, $\mathrm{P}(\mathbf{x}|\mathbf{W}, \tilde{\mathbf{x}})$. The similarity between the two is measured using the Kullback-Leibler divergence and the recall dynamics are constructed as to minimize this cost function. Note that the distribution represented by the neural population preserves the marginals, but ignores all higher order structure of the original posterior.

If using coordinate descent [3] and restricting inputs to incoming synapses yields simple neural-like network dynamics:

$$\mu_i(t+1) = \sigma\left(I_i^{\mathrm{rec,in}} + I_i^{\mathrm{rec,out}} + a\tilde{x}_i + b\right), \tag{7}$$

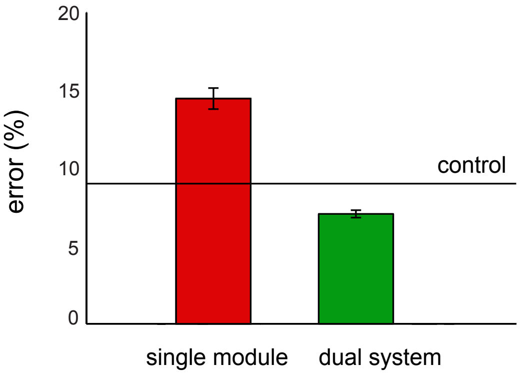

Figure 1: Comparison of the recall performance of single vs. dual memory system using a mean-field representation.

with $\sigma(x) = \frac{1}{1+e^{-x}}$ and $a_1 < 0$ and the recurrent current contributions:

$$I_i^{\text{rec,in}} = \sum_j \left(c_1^{\text{in}} \cdot W_{ij}\, \mu_j + c_2^{\text{in}} \cdot W_{ij} + c_3^{\text{in}} \cdot \mu_j + c_4^{\text{in}} \right) \tag{8}$$

$$I_i^{\text{rec,out}} = \sum_j \left(c_1^{\text{out}} \cdot W_{ji}\, \mu_j + c_2^{\text{out}} \cdot W_{ji} + c_3^{\text{out}} \cdot \mu_j + c_4^{\text{out}} \right). \tag{9}$$

Note that the total input current has a very similar form to that obtained for the sampling based dynamics (Eq. 7 in main text), with the same scaling constants a, b and $c_{1-4}^{\text{in/out}}$ as before.

A dual memory system implementation in this case will have a similar architecture as in the sampling case, with the activity in the familiarity module given by:

$$\begin{aligned} I_i^{\text{fam}} &= \sum_j \log \left(\mathrm{P}(W_{ij}^{\text{fam}} | x_i = 1, x_j = 1, t) \cdot \mu_j + \mathrm{P}(W_{ij}^{\text{fam}} | x_i = 1, x_j = 0, t) \cdot (1 - \mu_j) \right) \\ &+ \log(\mathrm{P}(t)) - \log(Z). \end{aligned} \tag{10}$$

Again, softmax competition ensures that the corresponding distribution of the pattern ages in properly normalized. In this case, the signal transmitted to the recognition module can be the MAP estimate of this posterior or the average t, translating into a corresponding set of values for the parameters $c_{1-4}^{\text{in/out}}$, as before.

The Gibbs sampling and the mean-field implementation behave similarly in the monolithic memory system, as both are poor at representing correlated probability distributions (Fig.1). Moreover, both benefit from the explicit estimation of pattern age, suggesting that the effects presented are not specific to the selected representation.

References

[1] Neal, R.M. Sampling from multimodal distributions using tempered transitions. *Statistics and Computing* **6**, 353–366 (1996).

[2] Bogacz, R. Comparison of computational models of familiarity discrimination in the perirhinal cortex. *Hippocampus* (2003).

[3] Sommer, F.T. & Dayan, P. Bayesian retrieval in associative memories with storage errors. *IEEE transactions on neural networks* **9**, 705–713 (1998).