Correlations strike back (again): the case of associative memory retrieval – supplementary information –

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

Optimal retrieval dynamics for generalized Hebbian learning

Here and later in this supplement, we repeat selected text from the main paper concerning the mathematical argument, and add the details that we lacked the space to discuss.

For retrieval dynamics based on Gibbs sampling a key quantity is the the log-odds ratio, which, for neuron i, is:

$$I_i = \log\left(\frac{\mathrm{P}(x_i = 1|\mathbf{x}_{\neg i}, \mathbf{W}, \tilde{\mathbf{x}})}{\mathrm{P}(x_i = 0|\mathbf{x}_{\neg i}, \mathbf{W}, \tilde{\mathbf{x}})}\right) \tag{1}$$

corresponding to the total current entering the unit. This translates into a probability of firing given by the sigmoid activation function $f(I_i) = \frac{1}{(1+e^{-I_i})}$.

The total current entering a neuron is a sum of two terms: one term from the external input, I_i^{ext}, and the other from recurrent collaterals, I_i^{rec}. The external current has the form: $I_i^{\mathrm{ext}} = c_1 \cdot \tilde{x}_i + c_2$, with constants $c_1 = 2\log\left(\frac{1-r}{r}\right)$ and $c_2 = \log\left(\frac{fr}{(1-f)(1-r)}\right)$.

For computing the recurrent component, let us first consider the case of symmetric additive learning, when reciprocal connections are perfectly correlated. In this case, the probability $\mathrm{P}(\mathbf{W}|\mathbf{x})$ can be written as $\prod_{i,j}\delta(W_{ij}, W_{ji}) \cdot \mathcal{N}\left(\mathbf{W}_{\triangleleft}, \boldsymbol{\mu}_{\triangleleft}, \mathbf{C}_{\triangleleft}\right)$, where the subscript $_{\triangleleft}$ marks the fact that the multivariate normal is defined in the space of the weights from the upper triangular part of matrix $\mathbf{W}$. Writing down the exact expression for the log-odds ratio determining the current in this case yields:

$$I_i^{\mathrm{rec}} = \frac{1}{2(T-1)}\left(\left(\mathbf{W}_{|} - \boldsymbol{\mu}_W^{(0)}\right)^{\mathrm{T}}\mathbf{C}_1^{-1}\left(\mathbf{W}_{|} - \boldsymbol{\mu}_W^{(0)}\right) - \left(\mathbf{W}_{|} - \boldsymbol{\mu}_W^{(1)}\right)^{\mathrm{T}}\mathbf{C}_1^{-1}\left(\mathbf{W}_{|} - \boldsymbol{\mu}_W^{(1)}\right)\right), \tag{2}$$

where $\boldsymbol{\mu}_W^{(0/1)} = \boldsymbol{\Omega}(\mathbf{x}^{(0/1)}) + (T-1)\boldsymbol{\mu}_1$, $\mathbf{x}^{(0/1)}$ is the vector of activities obtained from $\mathbf{x}$ in which the activity of neuron i is set to 0, or 1, respectively, and the subscript $_{|}$ marks the fact that the weights are reorganised as a column vector.

When computing the current to a neuron i all the terms that are not local to i cancel out[1]. Hence, the above expression reduces to:

$$\begin{aligned} I_i^{\text{rec}} &= \sum_{j,k} C^{-1}_{\text{Idx}(i,j),\text{Idx}(i,k)}(W_{ij} - \Omega(1,x_j))(W_{ik} - \Omega(1,x_k)) - \\ & \sum_{j,k} C^{-1}_{\text{Idx}(i,j),\text{Idx}(i,k)}(W_{ij} - \Omega(0,x_j))(W_{i,k} - \Omega(0,x_k)), \end{aligned} \tag{3}$$

where $Idx(i,j)$ represents the index of synapse W_{ij} in the column-version representation $\mathbf{W}_|$. Importantly, the exact optimal dynamics depending only on incoming synapses to the neuron. If the learning rule would not be symmetric, the expression for the current above would include additional terms corresponding to the covariance between other types of pairs – incoming-outgoing, outgoing-outgoing, and reciprocal connections, making the decoder no longer strictly local. It is still possible to construct approximately optimal dynamics that are local by replacing the non-local component, by its expectation, conditioned on the local information:

$$I_i^{\text{total}} = I_i^{\text{local}}\left(\mathbf{W}_{i,\cdot}, \mathbf{x}\right) + \left\langle I_i^{\text{nonlocal}}\left(\mathbf{W}, \mathbf{x}\right) | \mathbf{W}_{i,\cdot}, \mathbf{x} \right\rangle \tag{4}$$

Lastly, what does the remaining local term mean for circuit dynamics? The current in Eq.3 can be separated in one component summing over the terms with $j = k$ and the other for all remaining terms:

$$I_i^{\text{rec},1} = \frac{c_{\text{lin}}}{2} \sum_j \left(2W_{ij}x_j - 2\beta W_{ij} - (1-2\alpha)(1-2\beta)x_j - (1-2\alpha)\beta^2\right) \tag{5}$$

$$I_i^{\text{rec},2} = \frac{v_{\text{nonlin}}(2\alpha - 1)}{2} \left(\sum_{j \neq i} x_j \sum_{k \neq i} W_{ik} - \sum_{j \neq i} x_j \sum_{k \neq i} x_k \right) \tag{6}$$

where we have used the generalised Hebbian learning rule, $\Omega(x_i, x_j) = (x_i - \alpha)(x_j - \beta)$ and factored out $c_{\text{lin}} = C^{-1}_{\text{Idx}(i,j),\text{Idx}(i,j)}$, $c_{\text{nonlin}} = C^{-1}_{\text{Idx}(i,j),\text{Idx}(i,k)}$(same for all terms). In the case of the covariance rule, the second component cancels out, as $c_{\text{nonlin}} = 0$. For any other Hebbian learning rule in the family considered, the total current involves the second term, which is has a quadratic dependence on the total activity $n_b = \sum_j x_j$, translating into nonlinear dynamic inhibition in the neural circuit.

Cascade details

Learning is stochastic and local, with changes in the state of a synapse V_{ij} being determined only by the activation of the pre- and post-synaptic neurons, x_j and x_i. The transition matrices for potentiation, $\mathbf{M}_+$ and depression, $\mathbf{M}_-$ (of size $2n \times 2n$, with n being the cascade depth), are defined using Fusi's cascade model [1], which assumes that the probability of potentiation and depression decays with cascade depth i as a geometric progression, $q_i^{\pm} = \chi^{i-1}$, with $q_n^{\pm} = \frac{\chi^{n-1}}{1-\chi}$ to compensate for boundary effects. The transition between metastates is given by $p_i^{\pm} = \varsigma_{\pm} \frac{\chi^i}{1-\chi}$, with the correction factors $\varsigma_+ = \frac{1-f}{f}$ and $\varsigma_- = \frac{f}{1-f}$ ensuring that different metastates are equally occupied for different pattern sparseness values f [1].

Additionally, we consider three possible mappings from neural activity to potentiation events:

- a post-synaptically gated learning rule, R1: $\mathbf{M}(0,0) = \mathbf{I}$, $\mathbf{M}(0,1) = \mathbf{I}$, $\mathbf{M}(1,0) = \mathbf{M}_-$, $\mathbf{M}(1,1) = \mathbf{M}_+$;
- a pre-synaptically gated learning rule, R2: $\mathbf{M}(0,0) = \mathbf{I}$, $\mathbf{M}(0,1) = \mathbf{M}_-$,$\mathbf{M}(1,0) = \mathbf{I}$, $\mathbf{M}(1,1) = \mathbf{M}_+$
- the XOR-like learning rule of Ben Dayan Rubin et al [2], R3: $\mathbf{M}(0,0) = \mathbf{M}_+$, $\mathbf{M}(0,1) = \mathbf{M}_-$,$\mathbf{M}(1,0) = \mathbf{M}_-$,$\mathbf{M}(1,1) = \mathbf{M}_+$.

[1] This is only true for the class of additive learning rules because the same covariance matrix appears in both of the terms of the above sum; the non-local terms will not cancel out in the palimpsest case.

Estimating the covariance matrix

To estimate the degree of correlations for different synaptic configurations we need to compute the joint distribution of synaptic pairs. As an illustration, we take the case of two incoming synapses to the same neuron; other configurations are very similar. We represent the joint probability of a synaptic states pair as a matrix $\boldsymbol{\rho}$, with $2n$ rows and $2n$ columns, with , $\rho_{ab} = \mathrm{P}(V_{ij} = a, V_{ik} = b)$.

For the first synapse, between neurons i and j, a column v in matrix $\boldsymbol{\rho}$, is proportional to the conditional distribution $\mathrm{P}(V_{ij}|V_{ik} = v)$. Hence, an encoding event for this synapse, means a multiplication by $\mathbf{M}(x_i, x_j)$. For the second synapse the distribution over states is a row vector, so we need an additional transposition, before applying the corresponding transition operator, $\mathbf{M}(x_i, x_k)$. Putting the two together, we obtain the operator for encoding: $(x_{\text{pre1}}, x_{\text{pre2}}, x_{\text{post}})$ as:

$$\boldsymbol{\rho}^{(1)} = \left(\mathbf{M}(x_{\text{post}}, x_{\text{pre2}}) \cdot \left(\mathbf{M}(x_{\text{post}}, x_{\text{pre1}}) \cdot \boldsymbol{\rho}^{(0)}\right)^{\mathrm{T}}\right)^{\mathrm{T}} = \mathbf{M}(x_{\text{post}}, x_{\text{pre1}})\cdot\boldsymbol{\rho}^{(0)}\cdot\mathbf{M}(x_{\text{post}}, x_{\text{pre2}})^{\mathrm{T}}, \tag{7}$$

with $\boldsymbol{\rho}^{(0)}$ the stationary distribution, corresponding to storing an infinite number of triplets from the pattern distribution.[2]

Replacing $\boldsymbol{\pi}^V$ with $\boldsymbol{\rho}$ (which is now a function of the triplet $(x_{\text{pre1}}, x_{\text{pre2}}, x_{\text{post}})$ and the multiplication with $\mathbf{M}$ by the slightly more complicated operator above, we can estimate the evolution of the joint distribution over synaptic states in a manner very similar to the calculations above:

$$\boldsymbol{\rho}^{(t)} = \sum_{x_i} \mathrm{P}_{\text{store}}(x_i) \cdot \hat{\mathbf{M}}(x_i) \cdot \boldsymbol{\rho}^{(t-1)} \cdot \hat{\mathbf{M}}(x_i)^{\mathrm{T}}, \tag{8}$$

where $\hat{\mathbf{M}}(x_i) = \sum_{x_j} \mathrm{P}_{\text{store}}(x_j)\mathbf{M}(x_i, x_j)$ is an analog of $\overline{\mathbf{M}}$, conditioned on the shared neuron's activity. The final joint distribution over states can be mapped in a joint over synaptic weights as $\mathbf{M}_{V\to W} \cdot \boldsymbol{\rho}^{(t)} \cdot \mathbf{M}_{V\to W}^{\mathrm{T}}$, as done above, which see use for constructing covariance matrix $\mathbf{C}$.

For outgoing pairs, we would obtain a very similar expression, just changing the activity component that is shared between the two $\boldsymbol{\rho'}^{(1)} = \mathbf{M}(x_{\text{post1}}, x_{\text{pre}})\cdot\boldsymbol{\rho'}^{(0)}\cdot\mathbf{M}(x_{\text{post2}}, x_{\text{pre}})^{\mathrm{T}}$ (which also implies defining a different average transition matrix $\hat{\mathbf{M}}'(x_i) = \sum_{x_j} \mathrm{P}_{\text{store}}(x_j)\mathbf{M}(x_j, x_i)$), and so on.

TAP procedure for fitting the maximum entropy model

To consider the effect of synaptic correlations, we approximate $\mathrm{P}(\mathbf{W}|\mathbf{x})$ by a maximum entropy distribution with the same marginals and covariance structure, ignoring the higher order moments:

$$\mathrm{P}(\mathbf{W}|\mathbf{x}, t) = \frac{1}{Z(\mathbf{x}, t)} \exp\left(\sum_{ij} k_{ij}(\mathbf{x}, t) \cdot W_{ij} + \frac{1}{2}\sum_{ijkl} J_{ijkl}(\mathbf{x}, t) \cdot W_{ij}W_{kl}\right) \tag{9}$$

We use the TAP mean-field method [3] to compute the model parameters, $\mathbf{k}$ and $\mathbf{J}$, and the partition function, Z, for each possible activity pattern $\mathbf{x}$, given the mean and covariance for the synaptic weights derived in the main text.[3]

Briefly, given the target mean $\mathbf{m}$ and covariance $\mathbf{C}$ in the spin representation, the parameters can be computed by solving the equations[4]:

$$\begin{aligned}
\tanh^{-1}(m_{ij}) &= k_{ij} + \sum_{(i,j)\neq(k,l)} J_{(ij)(kl)} \cdot m_{kl} - \sum_{(i,j)\neq(k,l)} J^2_{(ij)(kl)} \cdot m_{ij}(1 - m^2_{kl}) && (10)\\
(C^{-1})_{(ij)(kl)} &= -J_{(ij)(kl)} - 2m_{ij}m_{kl}J^2_{(ij)(kl)} && (11)
\end{aligned}$$

where we first solve the second equation for J_{ij} (additional continuity constraints determine which of the two solutions to select, see [4]), then solve the first equation for k_i. Lastly, the normalizing

[2] In this case, we estimate the stationary distribution numerically, by repeatedly applying the operator in 8.

[3] The TAP fit uses a spin-based representation for the variables; we subsequently convert the parameters K and J back to a binary representation.

[4] To keep things simple, we keep the dependence of the parameters $\mathbf{x}$ implicit here.

constant Z can be computed as [5]:

$$\log(Z) = \sum_{i,j} \log\left(2\cosh(k_{ij} + L_{ij})\right) - \sum_{i,j} L_{ij} m_{ij} + \frac{1}{2} \sum_{(i,j)\neq(kl)} J_{(ij)(kl)} m_{ij} m_{kl} + \quad (12)$$

$$\frac{1}{4} \sum_{(i,j)\neq(kl)} J^2_{(ij)(kl)} \left(1 - m^2_{ij}\right)\left(1 - m^2_{kl}\right) \quad (13)$$

where $L_{ij} = \frac{1}{2}\left(\sum_{(i,j)\neq(kl)} J_{(ij)(kl)} m_{kl} - m_{ij} \sum_{(i,j)\neq(kl)} J^2_{(ij)(kl)} \left(1 - m^2_{kl}\right)\right)$.

We use the TAP method to find parameters $\mathbf{k}$ and $\mathbf{J}$ and the partition function, Z, for each possible activity pattern $\mathbf{x}$, given the mean and covariance for the synaptic weights matrix. This seems computationally daunting, as the number of distributions needed to be fitted is exponential in the network size. However, due to the all-to-all connectivity assumption, we can use symmetry arguments to reduce this computation from 2^N, corresponding to all possible values of the pattern $\mathbf{x}$, to $N + 1$ TAP runs, corresponding to patterns including 0 to N bits set to 1. Intuitively, because of the all-to-all connectivity, there is no inherent indexing of the neurons; if we have computed $\mathrm{P}(\mathbf{W}|\mathbf{x}_1)$, for a pattern $\mathbf{x}_1 = (1\ 0\ 0)$, then when estimating the evidence for a pattern $\mathbf{x}_2 = (0\ 1\ 0)$, $\mathrm{P}(\mathbf{W}|\mathbf{x}_2)$, we do not need to recompute the max entropy parameters, as they are the same lest for a permutation of neuron indices (the mean and covariance matrices are the same after reordering the indices; this is because we compute the mean and covariance analytically, from the joint $\boldsymbol{\rho}^{(t)}$, so they are the same for all synaptic pairs receiving the same activity pattern). Hence all patterns with the same number of active bits are isomorphic. The TAP procedure is nonetheless very computationally expensive, since the parameter space is of the order N^4 (we could simulate networks of up to 50 neurons).

Simulation parameters

Default parameters: $f = 0.5$, $r = 0.1$. For the additive learning rule, we use $T = 5$; we simulate $N_{\text{run}} = 100$ retrieval episodes; for each we generate a sequence of T random patterns, used for computing W; we pick the first as the target for retrieval and generate a corresponding recall cue; starting from this recall cue we run the Gibbs sampler for 10000 steps, use all the samples for estimating the mean of the posterior (no burn-in), with the final error given by the euclidian distance between the network response and the true pattern. In the palimpsest case, the simulations assume cascade depth $n = 5$, $t = 3$, $N_{\text{run}} = 50$ and the same general procedure, with a sequence of cascade transitions for the encoding (starting from the stationary distribution).

References

[1] Fusi, S., Drew, P.J. & Abbott, L.F. Cascade models of synaptically stored memories. *Neuron* **45**, 599–611 (2005).

[2] Dayan Rubin, B. & Fusi, S. Long memory lifetimes require complex synapses and limited sparseness. *Frontiers in Computational Neuroscience* (2007).

[3] Thouless, D.J., Anderson, P.W. & Palmer, R.G. Solution of 'Solvable model of a spin glass'. *Philosophical Magazine* **35**, 593–601 (1977).

[4] Tanaka, T. Mean-field theory of Boltzmann machine learning. *Phys. Rev. E* **58**, 2302–2310 (1998).

[5] Schaub, M.T. & Schultz, S.R. The Ising decoder: reading out the activity of large neural ensembles. *Journal of computational neuroscience* (2011).

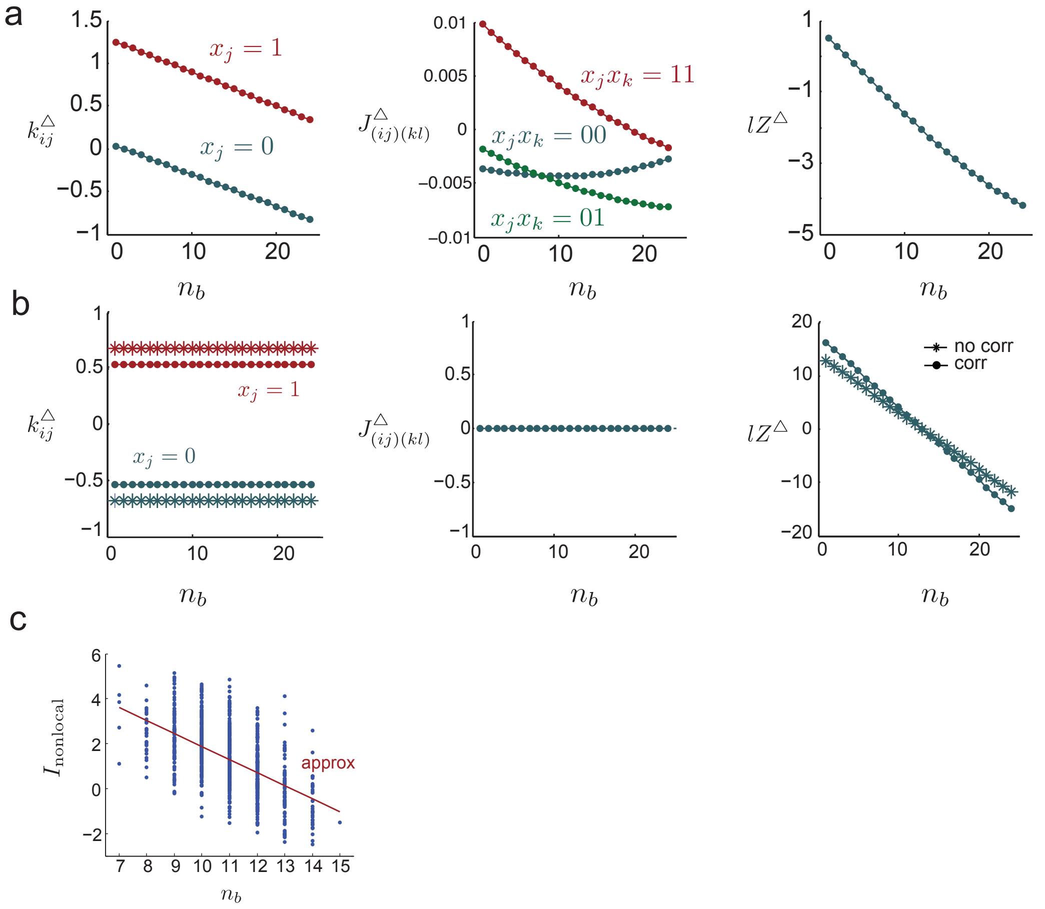

Figure 1: **Implications of synaptic correlations on neural dynamics.** **a**) Parameters for the local component of the neural dynamics for a pre-synaptically gated cascade. **b**) Same for the XOR rule. **c**) Approximation of the non-local term for post-synaptically gated learning. The nonlocal current is replaced by a linear function in n_b, with parameters determined numerically (by linear regression).