Attractor dynamics in networks with learning rules inferred from *in vivo* data - Mathematical Note

Ulises Pereira and Nicolas Brunel

May 22, 2018

Contents

1 The Model

We consider a network of N neurons with firing rates represented by a vector of analog variables $\vec{r}$. Standard normal patterns of current $\{\vec{\xi}^k\}_{k=1}^p$ with $\xi_i^k \overset{iid}{\sim} \mathcal{N}(0,1)$ are imprinted in the connectivity matrix as the corresponding firing rates elicited by these current patterns, neglecting contributions of the recurrent connections. Hence, the firing rate patterns corresponding to these current patterns are given by $\phi(\xi_i^k)$, where ϕ is the static transfer function of single neurons. In other words, the stored firing rate patterns are standard normal patterns of current $\{\vec{\xi}^k\}_{i=1}^p$ passed through the static transfer function ϕ. Note that in the limit where h_0 is large, these firing rate patterns become distributed according to a log-normal distribution, since the transfer function is exponential in that limit. The rate dependent learning rule is given by two firing rate dependent functions: 1) g which characterizes the dependence on the firing rate of the pre synaptic neuron; 2)f which characterizes the dependence on the firing rate of the post synaptic neuron. With this learning rule, assuming a linear summation of terms corresponding to the different patterns, as in the Hopfield model (Hopfield, 1982) and many of its generalizations, the connectivity matrix is given by

$$J_{ij} = \frac{Ac_{ij}}{cN} \sum_{k=1}^{p} f[\phi(\xi_i^k)] g[\phi(\xi_i^k)], \tag{1}$$

where c_{ij} is a sparse directed Erdős-Rényi structural connectivity with each synapse present with probability c, and the pair of functions f and g define together the learning rule. This is a generalization of classical Hebbian learning rules such as the covariance (Sejnowski, 1977) and BCM (Bienenstock et al., 1982) since the synaptic strength of the connections between pre and post synaptic neurons is proportional to the product of two functions of their activities. This feature allows a nonlinear dependence of the synaptic strength with the pre and post synaptic activity, but maintains the separability of the learning rule. The operation of f and g under a vector $\vec{r}$, i.e. $f(\vec{r})$ or $g(\vec{r})$, is element-wise. We assume that

$$\int_{-\infty}^{\infty} \mathcal{D}z g(\phi(z)) = 0 \tag{2}$$

which ensures that the average change in connection strength due to learning of a single pattern is zero. This could be enforced by a homeostatic mechanism that controls the mean changes in the incoming inputs due to learning (Toyoizumi et al., 2014; Vogels et al., 2011). In our model we assume that both functions f and g are bounded above and below by q_f/q_g and $q_f - 1/q_g - 1$, respectively, where $0 < q_f < 1$, $0 < q_g < 1$. The constant A in Eq. (1) controls the strength of the changes in the connectivity due to the learning rule.

The firing rate $r_i(t)$ of each neuron evolve according to standard rate equations (Hopfield, 1984), i.e.

$$\tau \dot{r}_i = -r_i + \phi\left(I_i + \sum_{i \neq j}^{N} J_{ij} r_j\right). \tag{3}$$

Thus, the steady or attractor state for the dynamics is given by

$$r_i = \phi\left(\sum_{i \neq j}^{N} J_{ij} r_j\right) \quad i = 1, \ldots, N. \tag{4}$$

2 Mean Field Analysis

2.1 Order parameters - delay period

Throughout this paper, we will perform a mean field analysis of the steady states of the network in the limits N, cN and p going to infinity, $1 \ll Nc \ll N$ and $p = \alpha/cN$ where α remains of order 1. We consider exclusively steady states that are correlated with a single pattern $\vec{\xi}^1$ but uncorrelated with all other patterns $\vec{\xi}^\mu$ for $\mu > 1$. States with a non-zero correlation with one of the patterns are termed 'retrieval states', while the state with no correlation with any of the patterns is termed 'background state'. The steady state $\vec{r}$ given by Eq. (4) depends on the pattern being

retrieved $\vec{\xi}^1$ (the 'signal') but also on two sources of frozen noise: 1) the disorder due to the random patterns stored in the connectivity; 2) the disorder given by the structural connectivity C (where C is a binary matrix with entries $c_{ij} \in \{0, 1\}$). The goal of the mean-field analysis is to compute whether and how the network state $\vec{r}$ is correlated with $\vec{\xi}^1$, together with other quantities of interest such as the distribution of firing rates.

The first step in the mean field analysis consists in computing the statistics of the synaptic inputs,

$$h_i = I_i + \sum_{i \neq j}^{N} J_{ij} r_j, \tag{5}$$

where the connectivity matrix J_{ij} is given by Eq. (1). We first start by the situation in which there are no external inputs, $I_i = 0$. In a delay match to sample experiment, this describes the intervals before presentation of the stimulus, and after this presentation (delay period)

To compute the statistics of synaptic inputs, it is useful to separate the contribution due to the first pattern ξ_i^1 that the network is trying to retrieve, with the contributions of all other patterns, which will act as noise on the retrieval of the first pattern,

$$h_i = A f(\xi_i^1) \frac{1}{cN} \sum_j c_{ij} g(\phi(\xi_j^1)) r_j + Y_i \tag{6}$$

where Y_i describes the 'noise' term,

$$Y_i = \frac{A}{cN} \sum_{\mu > 1} \sum_j c_{ij} f(\xi_i^\mu) g(\phi(\xi_j^\mu)) r_j$$

In the large cN limit, due to the law of large numbers, the first term in Eq. (6) converges in probability to $A f(\xi_i^1) q$, where q is given by

$$q = \frac{1}{N} \sum_i g(\phi(\xi_i^1)) r_i. \tag{7}$$

q is our first order parameter. It describes how correlated the network state is with a non-linear transformation of the stored pattern ξ_i^1, $g(\phi(\xi_i^1))$. This is a natural generalization of the overlap defined in classical models (Amit et al., 1985) for networks with generalized Hebbian learning rules.

It is instructive to consider first the case in which ξ^1 is the only stored pattern in the connectivity matrix. In this case, the synaptic input to neuron i is uniquely determined by the learning rate A, the post-synaptic function f taken at the firing rate induced by the pattern $\phi(\xi_i^1)$, and q. To compute q, we can use Eq. (7), replace r_i by $\phi(h_i)$ where $h_i = A f(\xi_i^1) q$, and replace $1/N \sum_i$ by an integral over the distribution of ξ_i,

$$q = \int \mathcal{D}\xi g(\phi(\xi)) \phi(A f(\phi(\xi)) q), \tag{8}$$

where $\mathcal{D}\xi$ denotes the Gaussian measure $d\xi e^{-\xi^2/2}/\sqrt{2\pi}$. Eq. (8) can be solved to obtain the possible values of q given f, g and A. Note that $q = 0$ (corresponding to the background state) is always a solution to this equation, due to Eq. (2).

In the case in which many patterns are stored in the connectivity matrix, we need to compute the statistics of the noise term Y_i. In the large p, N limits, this term becomes distributed according to a Gaussian distribution with zero mean (since the average of $g(\phi(\xi))$ over the distribution of ξs is zero) and a variance given by

$$\mathrm{Var}(Y_i) = \alpha \gamma M$$

where

$$\gamma \equiv A^2 \int_{-\infty}^{\infty} \mathcal{D}\xi f^2(\phi(\xi))) \int_{-\infty}^{\infty} \mathcal{D}\xi g^2(\phi(\xi))), \tag{9}$$

and M is our second order parameter, which is equal to the average squared firing rate over the network,

$$M = \frac{1}{N} \sum_i r_i^2. \tag{10}$$

The final step is to compute the order parameters self-consistently. For this, we use the fact that Y_i is a Gaussian random variable with zero mean and variance $\alpha\gamma M$, replace r_i by $\phi(qAf(\phi(\xi_i^1)) + Y_i$ in Eqs. (7,10) and replace the sums over i by a double integral over the distributions of ξ_i and Y_i, leading to

$$q = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{D}z\mathcal{D}y g(\phi(z))\phi(qAf(\phi(z)) + \sqrt{\alpha\gamma M}y) \tag{11}$$

$$M = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{D}z\mathcal{D}y \phi^2(qAf(\phi(z)) + \sqrt{\alpha\gamma M}y). \tag{12}$$

2.2 Distributions of firing rates - delay period

To compute the distribution of firing rates, we use the fact that the distribution of synaptic inputs conditioned on the pattern being retrieved is Gaussian,

$$p\left(h|\xi^1 = z\right) = \mathcal{N}\left(Af(\phi(z))q, \alpha\gamma M\right), \tag{13}$$

where the order parameters q and M are determined by the self-consistent equations (11) and (12).

Using the fact that the transfer function is non-decreasing, we obtain the distribution of steady state firing rates conditional to the pattern $\vec{\xi}^1$ presented during the delay period

$$p_r(r|\xi^1 = z) = \frac{1}{\sqrt{2\pi\alpha\gamma M}} \exp\left(-\frac{\left(\phi^{-1}(r) - Af(z)q\right)^2}{2\alpha\gamma M}\right) \frac{d\phi^{-1}(r)}{dr}. \tag{14}$$

From this conditional probability distribution, we obtain the marginal distribution of firing rates at the steady state, r,

$$p_r(r) = \int_{-\infty}^{\infty} \mathcal{D}z \frac{1}{\sqrt{2\pi\alpha\gamma M}} \exp\left(-\frac{\left(\phi^{-1}(r) - Af(z)q\right)^2}{2\alpha\gamma M}\right) \frac{d\phi^{-1}(r)}{dr}. \tag{15}$$

2.3 Order parameters and distributions of firing rates - presentation period

A similar analysis can be done in the situation when an external stimulus is presented to the network. We consider here two scenarios, one in which the presented stimulus is one of the stored patterns, $I_i = \xi_i^1$ (a 'familiar' stimulus), and the other in which the stimulus is uncorrelated with the stored patterns (a 'novel' stimulus).

In the 'novel' case, the synaptic inputs are

$$h_i = I_i + Y_i \tag{16}$$

where the external stimulus $\{I_i\}$ is independently sampled from a normal distribution with mean zero and variance I_0 (i.e. $I_i \overset{iid}{\sim} \mathcal{N}(0, I_0^2)$), where I_0 is the amplitude of the stimulation. For consistency reasons we use $I_0 = 1$ in all the results shown in this paper, but show here calculations for arbitrary I_0. The stimulus $\vec{I}$ is independent of all the previous patterns learned $\{\vec{\xi}^k\}_{k=1}^p$. Therefore, the synaptic inputs are the sum of two uncorrelated Gaussian random variables, one with variance I_0^2, the other with variance $\alpha\gamma M$. Hence, they are distributed according to a Gaussian of variance $\sqrt{I_0^2 + \alpha\gamma M}$.

Since the stimulus is uncorrelated with all stored patterns, the overlap q is equal to zero, while the other order parameter M is given by

$$M = \int_{-\infty}^{\infty} \mathcal{D}z \phi^2(\sqrt{I_0^2 + \alpha\gamma M}z). \tag{17}$$

The distribution of firing rates during the presentation period for a novel stimulus is a distribution of a Gaussian of mean zero and variance $\sqrt{I_0^2 + \alpha\gamma M}$ passed through the non-linear function ϕ and is therefore given by

$$p_{\text{pres}}^{\text{nov}}(r) = \frac{1}{\sqrt{2\pi(I_0^2 + \alpha\gamma M)}} \frac{d\phi^{-1}(r)}{dr} \exp\left(-\frac{(\phi^{-1}(r))^2}{2(I_0^2 + \alpha\gamma M)}\right). \tag{18}$$

In the 'familiar' case, the synaptic inputs during presentation of the pattern become

$$h_i = I_0\xi_i^1 + qAf(\phi(\xi_i^1)) + Y_i \tag{19}$$

where the first term in the r.h.s. of Eq. (19) is due to the external input, and the two other terms are identical to the situation analyzed in the previous section. Again, we use in all results shown in this paper $I_0 = 1$ but show the calculations for arbitrary I_0.

The distribution of the synaptic inputs, conditioned on the pattern ξ_i^1, has now a mean $I_0\xi_i^1 + qAf(\phi(\xi_i^1))$, and a variance $\alpha\gamma M$. This leads to the following equations for the order parameters q and M,

$$q = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{D}z\mathcal{D}y g(\phi(z))\phi(I_0 z + Af(\phi(z))q + \sqrt{\alpha\gamma M}y) \tag{20}$$

$$M = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{D}z\mathcal{D}y \phi^2(I_0 z + Af(\phi(z))q + \sqrt{\alpha\gamma M}y), \tag{21}$$

while the distribution of firing rates is

$$p_{\rm pres}^{\rm fam}(r) = \frac{1}{\sqrt{2\pi\alpha\gamma M}}\frac{d\phi^{-1}(r)}{dr}\int_{-\infty}^{\infty}\mathcal{D}z \exp\left(-\frac{(\phi^{-1}(r) - I_0 z - Af(\phi(z))q)^2}{2\alpha\gamma M}\right). \tag{22}$$

3 MFT when f and g are described by step functions

3.1 Equations for arbitrary A, q_f and q_g

Here we take f and g to be step functions (i.e. $\beta_f, \beta_g \to \infty$) with the same threshold, i.e.:

$$f(\eta) = \begin{cases} q_f & \eta \geq x_f \\ -(1-q_f) & \eta < x_f \end{cases} \tag{23}$$

and

$$g(\eta) = \begin{cases} q_g & \eta \geq x_f \\ -(1-q_g) & \eta < x_f. \end{cases} \tag{24}$$

The condition $\int_{-\infty}^{\infty}\mathcal{D}\xi g(\phi(\xi))) = 0$ implies that

$$q_g = \int_{-\infty}^{xf} dr \frac{\frac{d\phi^{-1}(r)}{dr}}{\sqrt{2\pi}} e^{-\frac{(\phi^{-1}(r))^2}{2}}.$$

The mean field equations simplify to

$$q = q_g(1-q_g)\left\{\int_{-\infty}^{\infty}\mathcal{D}y\phi\left(A\sqrt{\tilde{\gamma}}\left[\left(\frac{q_f}{\sqrt{\tilde{\gamma}}}\right)q + \sqrt{\alpha M}y\right]\right) - \int_{-\infty}^{\infty}\mathcal{D}y\phi\left(A\sqrt{\tilde{\gamma}}\left[-\left(\frac{1-q_f}{\sqrt{\tilde{\gamma}}}\right)q + \sqrt{\alpha M}y\right]\right)\right\} \tag{25}$$

$$M = (1-q_g)\int_{-\infty}^{\infty}\mathcal{D}y\phi^2\left(A\sqrt{\tilde{\gamma}}\left[q\left(\frac{q_f}{\sqrt{\tilde{\gamma}}}\right) + \sqrt{\alpha M}y\right]\right) + q_g\int_{-\infty}^{\infty}\mathcal{D}y\phi^2\left(A\sqrt{\tilde{\gamma}}\left[-\left(\frac{1-q_f}{\sqrt{\tilde{\gamma}}}\right)q + \sqrt{\alpha M}y\right]\right) \tag{26}$$

where

$$\tilde{\gamma} = \int_{-\infty}^{\infty}\mathcal{D}\xi\left(g(\phi(\xi))^2\right)\int_{-\infty}^{\infty}\mathcal{D}\xi\left(f(\phi(\xi))^2\right) = q_g(1-q_g)\left[q_f^2(1-q_g) + (1-q_f)^2 q_g\right].$$

Defining

$$
\begin{aligned}
m_0 &\equiv \frac{q}{r_m q_g (1-q_g)} && (27)\\
M_0 &\equiv \frac{M}{r_m^2} && (28)\\
\bar{A} &\equiv A\sqrt{\tilde{\gamma}} && (29)\\
\psi(x) &\equiv \frac{\phi(x)}{r_m} && (30)\\
p &\equiv 1-q_g && (31)\\
\eta &\equiv \sqrt{\frac{q_g(1-q_g)}{q_f^2(1-q_g)+(1-q_f)^2 q_g}}, && (32)
\end{aligned}
$$

we obtain

$$
\begin{aligned}
m_0 &= \int_{-\infty}^{\infty} \mathcal{D}y\psi\left(\bar{A}\left[q_f\eta m_0 + \sqrt{\alpha M_0}y\right]\right) - \int_{-\infty}^{\infty} \mathcal{D}y\psi\left(\bar{A}\left[-(1-q_f)\eta m_0 + \sqrt{\alpha M_0}y\right]\right) && (33)\\
M_0 &= p\int_{-\infty}^{\infty} \mathcal{D}y\psi^2\left(\bar{A}\left[q_f\eta m_0 + \sqrt{\alpha M_0}y\right]\right) + (1-p)\int_{-\infty}^{\infty} \mathcal{D}y\psi^2\left(\bar{A}\left[-(1-q_f)\eta m_0 + \sqrt{\alpha M_0}y\right]\right). && (34)
\end{aligned}
$$

3.2 Equations for $q_g = q_f$

When $q_f = q_g$, the mean field equations read

$$
\begin{aligned}
m_0 &= \int_{-\infty}^{\infty} \mathcal{D}y\psi\left(\bar{A}\left[(1-p)m_0 + \sqrt{\alpha M_0}y\right]\right) - \int_{-\infty}^{\infty} \mathcal{D}y\psi\left(\bar{A}\left[-pm_0 + \sqrt{\alpha M_0}y\right]\right) && (35)\\
M_0 &= p\int_{-\infty}^{\infty} \mathcal{D}y\psi^2\left(\bar{A}\left[(1-p)m_0 + \sqrt{\alpha M_0}y\right]\right) + (1-p)\int_{-\infty}^{\infty} \mathcal{D}y\psi^2\left(\bar{A}\left[-pm_0 + \sqrt{\alpha M_0}y\right]\right). && (36)
\end{aligned}
$$

Solutions to this equation are numerically explored in Fig. S6C,D of Data S1.

3.3 Recovering Tsodyks equations in the large A limit

In the limit $\bar{A} \to \infty$, the function $\psi(\bar{A}x)$ become a step (Heaviside) function, $\psi(\bar{A}x) \to 1$ if $x > 0$, 0 otherwise. Consequently, the mean field equations become

$$
\begin{aligned}
m_0 &= \Phi\left(\frac{-(1-p)m_0}{\sqrt{\alpha M_0}}\right) - \Phi\left(\frac{pm_0}{\sqrt{\alpha M_0}}\right) && (37)\\
M_0 &= p\Phi\left(\frac{-(1-p)m_0}{\sqrt{\alpha M_0}}\right) + (1-p)\Phi\left(\frac{pm_0}{\sqrt{\alpha M_0}}\right), && (38)
\end{aligned}
$$

where $\Phi(x) = \int_x^\infty dx e^{-x^2/2}/\sqrt{2\pi}$. These equations are identical to equations (20) and (21) derived by Tsodyks (1988) in a sparsely connected network of binary 0,1 neurons (with a threshold θ_0) storing binary random patterns with coding level p, with $\theta_0 = 0$. Note that the full equations derived by Tsodyks can be recovered when the threshold of the transfer function scales as $h_0 = \bar{A}\theta_0$.

Using these equations, Tsodyks found that the capacity diverges in the sparse coding limit as $\alpha_c \approx \frac{\theta_0^2}{2p\log(1/p)}$ (Tsodyks, 1988). In our network, the capacity cannot diverge in the $p \to 0$ limit due to the fact that $\theta_0 = 0$, since h_0 is a fixed parameter and therefore does not scale with $\bar{A}$. However, optimizing the threshold of the transfer function together with the parameters of the learning rule would allow one to reach the same scaling as the one obtained by Tsodyks (1988). This would require setting $h_0 = \bar{A}\theta_0$.

To obtain the capacity of our network, i.e. the largest value of α for which we can find a solution of Eqs. (37,38) with $m_0 > 0$, we analyze the Jacobian of the right side of equations (37) and (38) in the limit $m_0 \to 0$ (i.e. when the overlap approaches to zero) which gives

$$\mathbb{J} = \begin{pmatrix} \frac{1}{\sqrt{\pi\alpha}} & 0 \\ 0 & 0 \end{pmatrix}.$$

For equations (37) and (38) to have a stable solution in the limit $m_0 \to 0$, the eigenvalues of the Jacobian have to be less than one. This leads to the maximal capacity

$$\alpha_c = \frac{1}{\pi} \approx 0.318, \tag{39}$$

for all p.

Since the trace of the Jacobian is zero at the critical point, then the phase transition is of the second order (see S7 A and B). The parameter p has no effect on the capacity for this limit and the capacity is much lower than what has been found for the best-fit median parameters. For $q_f \neq q_g$, it is straightforward to show that the capacity is

$$\alpha_c = \frac{\eta^2}{\pi} \tag{40}$$

for all p.

This is always lower or equal than what is found in Eq. (39) since $\max_{q_f \in [0,1]}(\eta) = 1$ with $\text{argmax}_{q_f \in [0,1]}(\eta) = q_g = 1 - p$.