[11pt]article [english]babel [utf8]inputenc [T1]fontenc float lmodern,amsmath,amssymb,amstext,amsfonts,ma
subcaption [width=14cm]geometry [colorlinks,pdfpagelabels,pdfstartview = FitV,bookmarksnumbered
= true, bookmarksopenlevel=section, linkcolor = black,hypertexnames = false,citecolor
= black,pdfpagelabels=false]hyperref tablefootnote textcmds, enumitem sidecap
[labelfont=bf,sf,font=small,labelsep=space]caption chngcntr

1 Methods

In this chapter, we will give an overview of the recurrent network (RNN) models for exploration of the feedforward recurrent alignment hypothesis that is evolved in this work. The firstly introduced symmetric network model builds the basis for modifications and extensions in other further models. The modified models will be introduced subsequently. Finally, we consider the role of learning could play in the feedforward recurrent alignment hypothesis.

1.1 Symmetric Recurrent Network Model

Due to the well-understood mathematical characteristics of symmetric RNNs, they are often applied in models for neuroscience for a better understanding of certain dynamics. Therefore, we first consider the basic case of having a symmetric recurrent network, which has a symmetric interaction matrix. For symmetric RNNs, if there is a connection between two neurons n_i and n_j, the strength of the directed connection from the neuron n_i to the neuron n_j equals the directed connection from n_j to n_i.

1.1.1 Symmetric Recurrent Interaction

In the model, we consider a full rank real symmetric recurrent interaction matrix J with Gaussian distributed entries with mean 0 and variance 1,

$$J_{ij} \sim \mathcal{N}(0, 1). \tag{1.1}$$

Besides, J has full rank equals the number of neurons n involved in the RNN,

$$rank(J) = n\,, \tag{1.2}$$

The eigenvalues of J are limited by parameter $R < 1$ through normalization with the maximal original eigenvalue λ_{max} form J,

$$\tilde{\lambda_i} = \frac{R\lambda_i}{\lambda_{max}} \quad \forall i\,, \tag{1.3}$$

where $\{\lambda_i\}_{i=1,\dots,n}$ are the original eigenvalues of J and $\{\tilde{\lambda_i}\}_{i=1,\dots,n}$ the re-scaled eigenvalues. As a result, the maximal eigenvalues after the re-scaling would take value $R < 1$.

1.1.2 Response Steady State

Existence of Steady State When considering the relationship between firing rate and synaptic current as linear, the dynamic system of the RNN illustrated in Figure 1.1 could be described by the ordinary differential equation [**?**]:

$$\tau_r \frac{\mathrm{d}r}{\mathrm{d}t} = -r + J \cdot r + h \, \tau_r = 1 \Rightarrow \frac{\mathrm{d}r}{\mathrm{d}t} = -r + J \cdot r + h \,, \tag{1.4}$$

with the vector $r \in R^{n \times 1}$ describing firing rate of neurons in the recurrent layer, the vector $h \in R^{n \times 1}$ as feedforward inputs, and τ_r the time constant controlling the speed of dynamic. The steady state of the dynamic system eq.(1.4) can be received by setting the ordinary differential equation to zero. For simplicity, the time constant is set to 1. We then have the formulation for steady-state response r^*:

$$\frac{\mathrm{d}r}{\mathrm{d}t} = -r + J \cdot r + h = 0 \Rightarrow r = (I_n - J)^{-1} \cdot h =: r^* \,, \tag{1.5}$$

where $I_n \in R^{n \times n}$ the identity matrix. Since J is full rank, the matrix $(I_n - J)$ is invertible. Therefore, the steady state exists.

Stability of Steady State The dynamic eq.(1.4) could also be written in an elementary expression:

$$f_i(r_1, ..., r_n) := \frac{\mathrm{d}r_i}{\mathrm{d}t} = -r_i + \sum_{j=1}^{n} J_{ij} r_j + h_i for i = 1, ..., n \,. \tag{1.6}$$

The partial differentiation of f_i to r_j is

$$\frac{\partial f_i}{\partial r_j} = \{ -1 + J_{ij} if \ i = j J_{ij} if \ i \neq j \ . \tag{1.7}$$

The Jacobian matrix A of the dynamic system (eq.1.4) is then

$$A := (\ ) \partial f_1 \partial x_1 \cdots \partial f_1 \partial x_n \vdots \ddots \vdots \partial f_n \partial x_1 \cdots \partial f_n \partial x_n = -I_n + J \,. \tag{1.8}$$

Therefore, the Jacobian matrix A is a linear transformation of the symmetric recurrent interaction matrix J, which is independent of the steady-state response. So, A has the same set of eigenvectors [1] as J. With $E := \{e_i\}_{i=1,...,n}$ the matrix containing eigenvectors of J column-wise, it follows the reformulation

$$(-I_n + J)E = -I_n \cdot E + J \cdot E = -I_n \cdot E + \Lambda \cdot E = (-I_n + \Lambda)E \,, \tag{1.9}$$

Λ the diagonal matrix with eigenvalues $\{\lambda_i\}_{i=1,...,n}$ of J on its diagonal. This means, $\{-1 + \lambda_i\}_{i=1,...,n}$ are eigenvalues for the Jacobian matrix A.

[1] For a symmetric matrix, the set of left eigenvectors equal the set of right eigenvectors

The eigenvalues of the Jacobian matrix A determine the stability of steady states. Here, since the matrix A is symmetric, all its eigenvalues $-1+\lambda_i, i = 1, ..., n$ are from R. Because the eigenvalues λ_i of matrix J is limited by the parameter $R < 1$, defined in eq.(1.3), we have

$$-1 + \lambda_i(1.3) < -1 + 1 = 0\,. \tag{1.10}$$

That is, all eigenvalues of the Jacobian matrix A are negative. This indicates that the steady state r^* is stable. Under the assumption that the system reaches its steady state quickly enough, we could apply the steady state r^* for further analysis.

1.1.3 Feedforward Recurrent Alignment for Symmetric Interactions

Generally, the feedforward inputs can be considered as a firing rate distribution with a certain mean value and disturbed with some noise captured by variance. The mean firing rate is essential for characterizing the strength of inputs. We therefore consider mainly the mean firing rate of inputs. For the rest of the work, if mentioning feedforward inputs without further definition, we refer to the mean firing rate of inputs.

The alignment of a feedforward input $h \in R^{n\times 1}$ with the recurrent network J is defined as [**?**]

$$\nu := \frac{h^T J h}{\parallel h \parallel^2}\,, \tag{1.11}$$

where we consider the Euclidean norm without loss of generality. T denotes the vector transpose. If the inputs are aligned to the eigenvector e_i of the recurrent interaction J, i.e. the feedforward input is proportional to the aligned eigenvector,

$$h \propto e_i\,, \tag{1.12}$$

the feedforward recurrent alignment ν is proportional to the eigenvalues λ_i, because inserting the proportionality eq.(1.12) in eq.(1.11) leads to

$$\nu = \frac{h^T J h}{\|h\|^2} \propto \frac{{e_i}^T J e_i}{\|e_i\|^2} = \frac{\lambda_i {e_i}^T e_i}{\|e_i\|^2} = \lambda_i\,. \tag{1.13}$$

It is therefore observed that the maximal alignment was attained when the input is aligned to the eigenvector e_{max} with maximal eigenvalue λ_{max} [**?**].

1.1.4 Response Properties for Evaluation

Trial-to-trial correlation Given the feedforward inputs that are from the same distribution for multiple trials, the correlation between responses from different trials indicates the reliability of the responses. A large correlation implies high reliability of the response generated by the RNN.

Modeling the inputs $h \in R^{n\times 1}$ as multivariate normal distributions with mean vector $\mu \in R^{n\times 1}$ and covariance matrix $\Sigma \in R^{n\times n}$

$$h \sim \mathcal{N}(\mu, \Sigma)\, with \Sigma := \sigma_{trial} I_n. \tag{1.14}$$

Then, the steady state response $r^* = (I_n - J)^{-1} \cdot h$ from eq.(1.5) has the linearly transformed normal distribution

$$r^* \sim \mathcal{N}\left((I_n - J)^{-1}\mu, (I_n - J)^{-1}\Sigma(I_n - J)^{-T}\right) , \tag{1.15}$$

where the mean vector and covariance matrix are linearly transformed. The property could be proved analogously as in [**?**] with the moment-generating function M_h of the multivariate normal distributed input h as following

$$M_h(t) = E\left[exp(t^T h)\right] = exp\left[t^T\mu + \frac{1}{2}t^T\Sigma t\right] . \tag{1.16}$$

The moment-generating function of the vector $r^* = (I_n - J)^{-1} \cdot h$ becomes

$$M_{r^*}(t) = M_h\left((I_n - J)^{-T}t\right) = exp\left[t^T\left((I_n - J)^{-1}\mu\right) + \frac{1}{2}t^T(I_n - J)^{-1}\Sigma(I_n - J)^{-T}t\right] , \tag{1.17}$$

which indicates the linearly transformed distribution of r^* as in eq.(1.15).

As defined in [**?**], the trial to trial correlation β_s for one stimulus s is calculated by taking the mean of correlations between N response trials that evoked by this stimulus. That is

$$\beta_s = \frac{2}{N(N-1)} \sum_{i=1,j=i+1}^{N} corr(r_i^s, r_j^s) , \tag{1.18}$$

where r_i^s is the i-th response trial that evoked by stimulus s and $corr(r_i^s, r_j^s)$ the Pearson correlation between two response patterns.

Intra-trial stability It was observed that when presenting ongoing visual grating stimuli, the responses in the visually naive cortex have a stronger variation than they are after the visual experience. To reflect the variation of responses during the stimulation period, the quantity of "intra-trial stability" was defined [**?**].

To model the time-dependent input $h(t) \in R^{n\times 1}$ with mean vector μ and its evoked steady-state responses $r(t) \in R^{n\times 1}$, the following stochastic differential equations are formulated $\mathrm{d\ h} = \mu \mathrm{dt} + \sigma_{time}\mathrm{d}W$
$\mathrm{d}r = (-r + J \cdot \mu)\mathrm{d}t + \sigma_{time}\mathrm{d}W$, with W the Wiener process, which is a continuous-time stochastic process with independent Gaussian increments.

To approximate the evoked response $r(t)$, eq.(1.1.4b) is solved numerically with Euler-Maruyama scheme

$$r_{t+1} = r_t + (-r_t + J \cdot \mu)\Delta t + \sigma_{time}\sqrt{\Delta t}\Delta\tilde{W}_t , \tag{1.19}$$

with r_t the response at time point t, Δt the step width for iteration, and $\Delta\tilde{W}_t \in R^{n\times 1}$ the Gaussian increment at time point t defined by the multivariate normal distribution with zero vector 0_v as mean vector and identity matrix I_n as covariance matrix,

$$\Delta\tilde{W}_t \sim \mathcal{N}(0_v, I_n) . \tag{1.20}$$

For an another step width $\Delta\tilde{t}$, the intra-trial stability $c(t, \Delta\tilde{t})$ was defined by the correlation between z-scored responses $\bar{r}$ at time t and its delay at time $t + \Delta\tilde{t}$

$$c(t, \Delta\tilde{t}) := \bar{r}(t)^T \bar{r}(t + \Delta\tilde{t}) \,, \tag{1.21}$$

where the z-scored response is defined as the variation from mean value normalized by variance,

$$\bar{r}(t) := \frac{r - \langle r \rangle}{\sigma_r} \,, \tag{1.22}$$

with mean value of r denoted by $\langle r \rangle$ and standard deviation by σ_r.

The final intra-trial stability for a time period T is the time-averaged value over all time points $0 \leq t \leq T - \Delta\tilde{t}$

$$\bar{c}(\Delta\tilde{t}) = \frac{1}{T - \Delta\tilde{t}} \int_0^{T-\Delta\tilde{t}} c(t, \Delta\tilde{t}) \mathrm{d}t = \frac{1}{T - \Delta\tilde{t}} \int_0^{T-\Delta\tilde{t}} \bar{r}(t)^T \bar{r}(t + \Delta t) \mathrm{d}t \,. \tag{1.23}$$

Dimensionality Reduction in the diversity of modular patterns following the onset of experience was observed in the experiments with the primary visual cortex of ferrets. Through the projection of activity patterns into the space spanned by their leading principal components, it could be quantified that the early modular responses are more diverse, residing in a higher-dimensional linear manifold than those found in the experienced cortex following the onset of visual experience. This suggested that these initial modular patterns reflect a more flexible repertoire of network responses to visual input, fundamentally different from the experience cortex. To capture this trend, the linear "dimensionality" was defined and computed [**?**].

Given the multivariate normal distributed inputs $h \in R^{n \times 1}$

$$h \sim \mathcal{N}(0_v, \Sigma^{Dim}) \,, \tag{1.24}$$

the linear transformed responses (analogously as (1.15)) are

$$r \sim \mathcal{N}(0_v, (I_n - J)^{-1} \Sigma^{Dim} (I_n - J)^{-T}) \tag{1.25}$$

with

$$\Sigma^{Dim} := \sum_{i=L}^{L+M_{dim}} exp\left(\frac{-2(i-L)}{\beta_{dim}}\right) e_i e_i^T \,, \tag{1.26}$$

in which the parameter $M_{dim} := \kappa\beta_{dim}$ and β_{dim} reflects the dimensionality [**?**] and κ for determining the number of directions e_i that contribute to the dimensionality. Since the eigenvectors of J build a set of basis for R^n, they could be chosen as basis vectors e_i for covariance matrix Σ^{Dim} in eq.(1.26). Hereby, the eigenvectors are ordered according to their eigenvalues in descending order,

$$e_{max}, ..., e_i, \mathbf{e}_j, ..., e_{min} \; suchthat \; \lambda_{max} > ... > \lambda_i > \lambda_j > ... > \lambda_{min} \,. \tag{1.27}$$

The exponential factor in eq.(1.26) simulates the exponential decay of variance ratio observed in data [**?**].

The linear effective dimensionality for quantifying the change of pattern diversity during visual maturation is defined based on participation ratio

$$d_{eff} := \frac{\left(\sum_i \lambda_i\right)^2}{\left(\sum_i \lambda_i^2\right)}\,, \tag{1.28}$$

where λ_i the eigenvalues of a certain covariance matrix Σ from an activity pattern distribution. Since as defined in eq.(1.26), eigenvectors e_i of Σ^{Dim} are also eigenvectors for J. The eigenvectors of a covariance matrix are also known as principal components. Therefore, the eigenvalues λ_i^{Dim} for the covariance matrix Σ^{Dim}, also known as variance ratio, are re-scaled eigenvalues λ_i of J expressed as

$$\lambda_i^{Dim} = exp\left(\frac{-2(i-L)}{\beta_{dim}}\right)\lambda_i\,. \tag{1.29}$$

The covariance of the responses shares the same eigenvectors as Σ^{Dim} based on its distribution eq.(1.25) and therefore also the same as J. The eigenvalues λ_i^{Act} for the responses can be obtained through re-scaling eq.(1.29)

$$\lambda_i^{Act} = exp\left(\frac{-2(i-L)}{\beta_{dim}}\right)\frac{1}{(1-\lambda_i)^2}\,, \tag{1.30}$$

for $i = L, ..., L+M$.

Insert the eigenvalues of responses eq.(1.30) in the formula for effective dimensionality to get the final formulation of dimensionality for responses with eigenvalues of J,

$$d_{eff,ana} = \frac{\left(\sum_{i=L}^{L+M} exp\left(-2\frac{i-L}{\beta_{dim}}\right)(1-\lambda_i)^{-2}\right)^2}{\sum_{i=L}^{L+M} exp\left(-4\frac{i-L}{\beta_{dim}}\right)(1-\lambda_i)^{-4}}\,. \tag{1.31}$$

The vector of explained variance ratios in the principal component analysis (PCA) is the normalized vector containing eigenvalues of the covariance matrix re-scaled by the largest eigenvalue in descending order, which then explains how much variance the corresponding principal component contributes. Therefore, another way to access the dimensionality is to empirically determine the explained ratio of generated data samples through PCA and insert the variance ratio into the definition of effective dimensionality, i.e.,

$$d_{eff,emp} = \frac{\left(\sum_{i=L}^{L+M} var_i\right)^2}{\sum_{i=L}^{L+M} var_i^2} \tag{1.32}$$

with var_i the i-th variance ratio.

Alignment with spontaneous activity The alignment of activity patterns to spontaneous activity reflects principally the size of overlaps between activity

patterns and spontaneous activity pattern. Assuming having the evoked response pattern as V and the spontaneous activity pattern as S, the projection of V to S could be quantified as the covariance matrix of V explained by the principal components of S, which results a variance ratio vector with elements v_i calculated as

$$v_i = \frac{p_{i,S}^T \cdot \Sigma_V \cdot p_{i,S}}{Tr(\Sigma_V)}\,, \tag{1.33}$$

for $i = 1, ..., n$. $p_{i,S}$ are the principal component of spontaneous activity and Σ_V the covariance matrix of evoked activity. Consider the projection of activity pattern V in all directions of spontaneous pattern S together to reflect the overall overlaps between two patterns, we calculate the alignment between a response trial $r_{i,V}$ form V to the spontaneous pattern S

$$\gamma_i = \frac{r_{i,V}^T \cdot \Sigma_S \cdot r_{i,V}}{\|r_{i,V}\|^2 Tr(\Sigma_S)}\,, \tag{1.34}$$

where Σ_S is the covariance of pattern S. The final alignment, denoted as γ, between S and V is then the average value of alignment between spontaneous activity S and all trials of evoked activity.

To model the inputs and responses, we assume that spontaneous activity was evoked by inputs from broad sources. Besides, since the spontaneous activity already exists almost a week before eye-opening for ferrets, we assume that they already fit into the activity space generated by recurrent network [**?**]. Therefore, the inputs would be explained by more directions (or eigenvectors) than stimuli-evoked responses as modeled before with eq.(1.24). The spontaneous activity then has a higher dimensionality. Since the parameter β_{dim} in eq.(1.24) indicates the dimensionality, we could set for spontaneous activity higher dimensionality with $\beta_{spont} > \beta_{dim}$ to generate high dimensional inputs. Therefor, we than have the broad inputs $h^{spont} \in R^{n\times 1}$ and spontaneous activity $r^{spont} \in R^{n\times 1}$, which are both multivariate normal distributed vectors

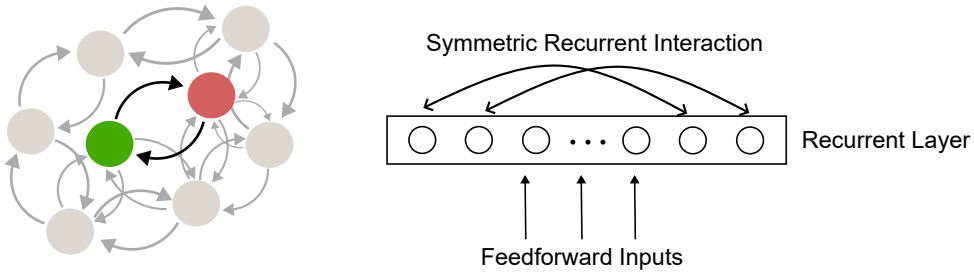
$$h^{spont} \sim \mathcal{N}(0_v, \Sigma^{spont}) \tag{1.35}$$

and

$$r^{spont} \sim \mathcal{N}\left(0_v, (I_n - J)^{-1}\Sigma^{spont}(I_n - J)^{-T}\right)\,. \tag{1.36}$$

The covariance matrix Σ^{spont} is constructed in the same way as Σ^{Dim} only with $L = 1$ and larger β_{spont} for $M_{spont} := \kappa\beta_{spont}$, that is

$$\Sigma^{spont} := \sum_{i=1}^{M_{spont}+1} exp\left(\frac{-2(i-1)}{\beta_{spont}}\right) e_i e_i^T\,. \tag{1.37}$$



(a) (b)

Figure 1.1: **Illustration of symmetric recurrent networks (symmetric RNNs).**
(a) An example of symmetric connections between two neurons in a network with multiple neurons. If there are connections between two neurons, here for example the green and red neurons, the directed connection from the green neuron to the red neuron has the same strength as the directed connection from the red to the green. **(b)** Structure of a symmetric RNN with feedforward inputs as the inputs for the recurrent layer. The connections between neurons inside the recurrent layer are symmetric, as illustrated in Figure (a).

1.2 Asymmetric Recurrent Network Model

Different from symmetric recurrent networks, asymmetric recurrent networks do not necessarily have symmetric interaction strength between two neurons. That is, if we have two neurons n_i and n_j, the interaction strength from n_i to n_j can generally differ from the interaction strength from n_j to n_i. Therefore, the network has more complex neural interactions and dynamics. Compared to a symmetric recurrent network, the asymmetric recurrent network reflects more general neural interactions and is thus more biologically realistic.

1.2.1 Asymmetric Recurrent Interaction

For the modeling, we construct the asymmetric interaction matrix J through disturbing a symmetric interaction matrix by a general asymmetric random matrix

$$J := aJ_{sym} + (1-a)J_{asym}\,. \tag{1.38}$$

The symmetric part is generated as described in section 1.1.1. The asymmetric part has Gaussian distributed entries with mean 0 and variance 1. Parameter a indicates the degree of symmetry in the network.

The recurrent network dynamics remains the same as in symmetric case, described by eq.(1.4). Therefore, the steady state also keeps its form as eq.(1.5). Stability analysis of the steady state follows the same procedure as for symmetric interaction matrix (section 1.1.2). However, here we have to keep in mind that the asymmetric interaction matrix J now have left and right eigenvectors, which are not identical to each other any more. The Jacobian matrix $A = -I_n + J$ defined in eq.(1.8) still has the same set of eigenvectors as J. The eigenvalues are found in the similar way as eq.(1.9) but with left and right eigenvectors. If E_l and E_r the matrices containing the left and right eigenvectors column-wise, it follows

$$E_l^*(-I_n+J) = -E_l^*+E_l^*J = -E_l^*+\Lambda E_l^* = (-I_n+\Lambda)E_l^*(-I_n+J)E_r = -E_r+JE_r = -E_r+E_r\Lambda = E_r(-I_n+\Lambda)\,, \tag{1.39}$$

with E_l^* the conjugate transpose matrix of E_l, I_n the identity matrix, and Λ the diagonal matrix that contains eigenvalues of J on its diagonal. The eigenvalues $\{\lambda_i\}$ for J are in general complex numbers

$$\{\lambda_i \in C \mid \lambda_i eigenvalues of$$

J$\}$.(1.40)Therefore, the eigenvalues for the Jacobian matrix A is now $-I_n+\Lambda \in C^{n\times n}$.

To determine the stability of steady state, we now have to consider the real part of Jacobian matrix $A = -I_n + \Lambda$. Criteria below are followed. If $\exists i : -1 + Re(\lambda_i) > 0 \Rightarrow the steady state will be unstable$. However, if $\forall i : -1 + Re(\lambda_i) < 0 \Rightarrow the steady state is stable$. As a result, the responses converge to the steady state after enough long time,

$$\forall i : -1 + Re(\lambda_i) < 0 \;\Rightarrow\; lim_{t\to\infty} r(t) = r^* \tag{1.41}$$

Thus, we need to limit the range of eigenvalues for J such that the real part of eigenvalues is limited by $R < 1$ and we could then apply the steady state for further analysis. The limitation could be achieved by dividing the complex eigenvalues by maximal absolute magnitude and re-scaled by R,

$$\tilde{\lambda}_i = \frac{R\lambda_i}{max\|\lambda\|} \quad \forall i\,. \tag{1.42}$$

The distribution of eigenvalues is influenced by the parameter a, which determines the proportion of symmetric interaction. In the case of only having a symmetric interaction network, all eigenvalues are real, and therefore the distribution forms a line in the complex plane. On the other hand, if J is only a general asymmetric interaction network, the distribution forms a circle [**?**]. For a between 0 and 1, the distribution is a symmetric ellipse.

1.2.2 Modifications of Feedforward Recurrent Alignment for Asymmetric Interactions

We want to have the similar quantification of feedforward recurrent alignment as for the symmetric interaction matrix (section 1.1.3). However, when aligning the feedforward input to eigenvectors of asymmetric recurrent interaction network J like in eq.(1.13), the result calculated with eq.(1.11) is a complex number and can therefore be hardly interpreted. To embed such a sore with analogical idea, we deliberate the following modifications and try to evaluate the outcomes.

Aligning feedforward input $h \in C^{n\times 1}$ to the asymmetric interaction J, we consider:

1. Only the real part of input h is relevant and calculate the feedforward recurrent alignment with $\tilde{h} := Re(h) \in R^{n\times 1}$.

$$\nu_{Re} := \frac{\tilde{h}^T J\tilde{h}}{\|\tilde{h}\|^2} = \frac{Re(h)^T JRe(h)}{\|Re(h)\|^2}\,. \tag{1.43}$$

2. Consider the magnitude of all entries in feedforward input h and calculate entries $\tilde{h_i} := |h_i| \in R\ \forall i$. Therefore $\tilde{h} \in R^{n\times 1}$, and the feedforward recurrent alignment has the formulation

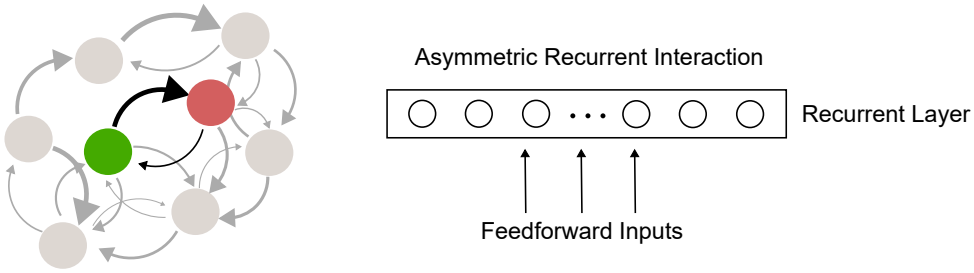
$$\nu_{mag} := \frac{\tilde{h}^T J\tilde{h}}{\|\tilde{h}\|^2} with \tilde{h}_i = |h_i| \in R\,. \tag{1.44}$$

3. Symmetrize J through

$$\tilde{J} = \frac{J + J^T}{2}\,. \tag{1.45}$$

   Instead of aligning the feedforward input directly to J, we align it indirectly to $\tilde{J}$ with its eigenvectors $\tilde{e}_i \in R^{n\times 1}$. So the modified alignment is $\tilde{h} := \tilde{e}_i$, the feedforward recurrent alignment will be calculated with eigenvectors of $\tilde{J}$:

$$\nu_{sym} = \frac{\tilde{h}^T J\tilde{h}}{\|\tilde{h}\|^2} = \frac{\tilde{e}_i^T J\tilde{e}_i}{\|\tilde{e}_i\|^2} \in R\,. \tag{1.46}$$



(a) (b)

Figure 1.2: **Illustration of asymmetric recurrent networks (asymmetric RNNs).** In general, asymmetric recurrent networks do not have symmetric interaction strength between two neurons. **(a)** An example of asymmetric connections between two neurons in a network with multiple neurons. There are connections between the green and red neurons. The connection from green to red is stronger than the connection from red to green. Some connections are only from one neuron to the other but no connection back from the other neuron. **(b)** Structure of an asymmetric RNN with feedforward inputs as the inputs for the recurrent layer. The connections between neurons inside the recurrent layer are asymmetric, as illustrated in Figure (a).

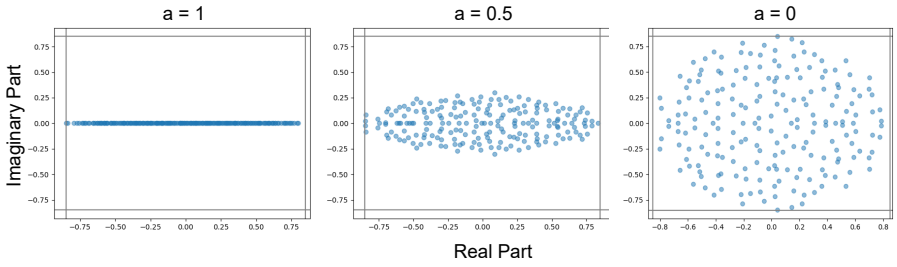


Figure 1.3: **Eigenvalue distribution in dependence of parameter a in eq.(1.38).** In general, an asymmetric matrix has eigenvalues in complex plane. The degree of symmetry determines the form of distribution from a line ($a = 1$, only real eigenvalues) to a full circle ($a = 0$) with radius $R < 1$. Between $a = 0$ and $a = 1$, the distribution is a symmetric ellipse along the line where imaginary part equals 0. Subfigures from left to right show the eigenvalue distribution in the complex plane from $a = 1, 0.5$, and 0. The gray line marks the radius $-R$ and R. Here $R = 0.85 < 1$.

1.2.3 Related Modifications for Evaluation

As for symmetric RNNs, four activity properties are taken into account to evaluate the feedforward recurrent alignment hypothesis with the measurement of the alignment score. In the case of asymmetric RNNs, the modified alignment scores (section 1.2.2) are considered in the evaluation.

More than the four response properties that are considered in the case of symmetric recurrent interaction matrix, for a new definition of feedforward recurrent alignment score, the monotony of score and corresponding eigenvalues have to be verified. Following are more details about how modifications in section 1.2.2 influence the modeling of properties that are involved in the evaluation.

Monotony For a certain eigenvector e_i, its corresponding eigenvalues λ_i can reflect the strength of the response. Dominant eigenvectors with large eigenvalues can therefore trigger strong response and suppress noise, leading to more reliable evoked activity. If the feedforward recurrent alignment score could well predict the reliability of evoked activities, inserting dominant eigenvectors should result in large alignment score values. Therefore, a large feedforward recurrent alignment score calculated with modified eigenvector $\tilde{h}$ should correspond with large eigenvalue. In other words, a monotonic positive correlation should exist between eigenvalues and feedforward recurrent alignment score inserting corresponding modified eigenvectors. When reliable responses occur, both feedforward recurrent alignment score and eigenvalue should be large.

Trial-to-trial correlation Similar to the case with symmetric RNNs, the first characteristic that is evaluated is the trial-to-trial correlation. With above modifications in section 1.2.2, the inputs are aligned with $\tilde{h}$ for all modifications. That is the mean vector for modeling the inputs by multivariate normal distribution is determined by $\tilde{h}$,

$$h \sim \mathcal{N}(\tilde{h}, \sigma_{trial} I_n)\,. \tag{1.47}$$

The steady state responses evoked by the inputs are transformed multivariate normal distribution

$$r \sim \mathcal{N}\left((1-J)^{-1}\tilde{h}, \sigma_{trial}(1-J)^{-1}(1-J)^{-T}\right)\,. \tag{1.48}$$

Trial-to-trial correlation is determined by β_s defined in eq.(1.18).

Intra-trial stability Modulation of the single trial is described by the stochastic differential equations (1.1.4). When the inputs are aligned to the modified eigenvectors, the mean value of inputs are again determined by $\tilde{h}$ defined in modifications above. As a result, the stochastic differential equations with modifications are $\mathrm{d\ h\ } = \tilde{h}\mathrm{d}t + \sigma_{time}\mathrm{d}W$
$\mathrm{d}r = (-r + J\cdot\tilde{h})\mathrm{d}t + \sigma_{time}\mathrm{d}W$, The solution of evoked activity is approximated by Euler-Maruyama scheme eq.(1.19). Intra-trial stability is then calculated by $\bar{c}(\Delta t)$ with eq.(1.23).

Dimensionality With symmetric interaction networks, the covariance matrix for the generation of inputs and responses is constructed with eigenvectors of the interaction matrix since they build up a set of basis for $R^{n\times n}$. But with an asymmetric interaction matrix, the eigenvectors are the basis for $C^{n\times n}$. If using complex eigenvectors, the inputs and responses will be complex without plausible interpretations. Therefore, the construction of the covariance matrix needs to be modified synchronously to have at least real vectors.

The same problem exists also for the analytical calculation for effective dimensionality in eq.(1.28): we now have complex eigenvalues that lead to the dimensionality also complex. To overcome this problem, we work along the same modifications as above for covariance matrix.

As a result, having $e_i \in C^{n\times 1}$ eigenvectors and eigenvalues $\lambda_i \in C$ of asymmetric interaction network J, we transfer complex eigenvectors and eigenvalues to real vectors and values based on section 1.2.2:

- For modification 1 with eq.(1.43), applying $\tilde{e}_i := Re(e_i)$ for construction covariance matrix and $\tilde{\lambda}_i := Re(\lambda_i)$ for calculation dimensionality.
- For modification 2 with eq.(1.44), applying magnitude for all entries in e_i to formulate $\tilde{e}_i$ and also magnitude of eigenvalues $\tilde{\lambda}_i := |\lambda_i|$.
- For modification 3 with eq.(1.46), applying eigenvectors $\tilde{e}_i$ and eigenvalues $\tilde{\lambda}_i$ from symmetrized interaction matrix $\tilde{J}$ by eq.(1.45).

The covariance matrix for generating inputs is constructed similar to it with symmetric interaction network defined by eq.(1.26) but with $\tilde{e}_i$ from above,

$$\Sigma^{Dim} := \sum_{i=L}^{L+M_{dim}} exp\left(\frac{2(i-L)}{\beta_{dim}}\right) \tilde{e}_i \tilde{e}_i^T . \tag{1.49}$$

Analogously, calculating the effective dimensionality analytically defined by eq.(1.31) but with $\tilde{\lambda}_i$,

$$d_{eff,ana} = \frac{\left(\sum_{i=L}^{L+M_{dim}} exp\left(-2\frac{i-L}{\beta_{dim}}\right)(1-\tilde{\lambda}_i)^{-2}\right)^2}{\sum_{i=L}^{L+M_{dim}} exp\left(-4\frac{i-L}{\beta_{dim}}\right)(1-\tilde{\lambda}_i)^{-4}} \tag{1.50}$$

Alignment to spontaneous activity The same formulation of covariance matrix with a higher dimensionality β_{spont} is used for generation of broader endogenous inputs for spontaneous activity, similar to section 1.1.4. The same modifications above for dimensionality in eq.(1.49) can be taken over into covariance matrix for endogenous inputs. The formulation of covariance matrix for spontaneous input is defined for symmetric RNNs in eq.(1.37). Applying the modifications above, insert therefore $\tilde{e}_i$ and receive

$$\Sigma^{spont} := \sum_{i=1}^{M_{spont}+1} exp\left(\frac{2(i-1)}{\beta_{spont}}\right) \tilde{e}_i \tilde{e}_i^T . \tag{1.51}$$

1.3 Low Rank Recurrent Network Model

Until now, we considered full-rank RNNs in both symmetric and asymmetric cases. Fully recurrent connectivity structure is one of the most popular and best-studied classes of network models. However, randomly connected networks display only very stereotyped responses to external inputs and can implement only a limited range of input-output computations [**?**].

Experimental large-scale neural recordings have established that the transformation of sensory stimuli into motor outputs relies on low-dimensional dynamics at the population level [**?**]. Besides, actual cortical connectivity appears to be neither fully random nor fully structured [**?**]. Understanding how low-dimensional computations on mixed, distributed representations emerge from the structure of the recurrent connectivity and inputs to cortical networks is still a major challenge with large potential [**?**]. It could help to understand the neural computations at the level of dynamical systems that govern low-dimensional trajectories of collective neural activity [**?**].

Low-rank RNNs rely on connectivity matrices that are restricted to be low rank, which directly generates low-dimensional dynamics. The rank of the network determines the number of collective variables needed to provide a full description of the collective dynamics [**?**].

1.3.1 Construction of Low Rank Interactions

Low-rank networks have a rank smaller than the number of neurons. We found two possible constructions of low-rank RNNs, which differ if the network is disturbed with Gaussian distributed random noise.

Low-rank RNNs without random noise [?, ?]. Here, neurons in low-rank RNNs are organized in distinct populations that correspond to clusters in the space of low-rank connectivity patterns. Each population is defined by its statistics of connectivity, described by a multivariate Gaussian distribution, so that the full network is specified by a mixture of Gaussians [**?**]. The connection matrix is constructed as a $n \times n$ dimensional matrix J where n is the number of neurons. Using singular value decomposition, the connectivity matrix with rank $G \ll n$ can be expressed as the sum of G unit rank terms

$$J_{ij} = \frac{1}{n}\sum_{g=1}^{G} l_i^{(g)} r_j^{(g)} \text{ or } J = \frac{1}{n}\sum_{g=1}^{G} l^{(g)} r^{(g)T} . \tag{1.52}$$

The connectivity is therefore characterized by a set of G n-dimensional vectors, denoted as connectivity patterns $l^{(g)} = \left\{ l_i^{(g)} \right\}_{i=1,...,n} \in R^{n\times 1}$ and $r^{(g)} = \left\{ r_i^{(g)} \right\}_{i=1,...,n} \in R^{n\times 1}$ for $g = 1, ..., G$. $\left\{ l^{(g)} \right\}$ are the left singular vectors of the connectivity matrix and $\left\{ r^{(g)} \right\}$ the right. The vectors $\left\{ l^{(g)} \right\}$ and $\left\{ r^{(g)} \right\}$ are mutually orthogonal and randomly multivariate Gaussian distributed [**?**].

Low-rank RNNs with random noise [?]. Similar to the construction of low-rank connection matrix above (figure 1.4 defined by eq.(1.52)), [**?**] suggested that the connectivity matrix can be constructed with a part P to be fixed and known and a random uncorrelated part. The fixed part P has the same construction as the network without noise in eq.(1.52). The uncontrolled random matrix can be generally constructed in a complex way with certain current-to-rate transfer function to lend more complexity [**?**]. We denote the random noise part with J_{rand}. The low-rank RNN can be formulated as

$$J = \frac{1}{n}\sum_{g=1}^{G} l^{(g)} r^{(g)T} + J_{rand}\,, \tag{1.53}$$

with $J_{rand} \in R^{n\times n}$ a random matrix.

1.3.2 Feedforward Recurrent Alignment Hypothesis with Low-rank RNNs

To test the feedforward recurrent alignment hypothesis applying the low-rank RNNs, we first consider the simple case of having symmetric low-rank connectivity with more easier interpretable dynamic. Then we tried the asymmetric low-rank RNNs and additionally discover the influence of rank on response properties.

Symmetric low-rank RNN This could be achieved through choosing the left connectivity vectors $\left\{l^{(g)}\right\}$ equal the right connectivity vectors $\left\{r^{(g)}\right\}$. As a result, the low-rank RNN without noise can be formulated by sum of symmetric matrices and therefore also symmetric:

$$J = \frac{1}{n}\sum_{g=1}^{G} l^{(g)} l^{(g)T} \;\; or \;\; J = \frac{1}{n}\sum_{g=1}^{G} r^{(g)} r^{(g)T}\,. \tag{1.54}$$

If considering the connectivity matrix with noise, for simplicity, we add a random $n \times n$ symmetrized Gaussian distributed matrix as J_{rand} to J in eq.(1.54).

The dynamics and conditions for stable stability of the response in the symmetric low-rank RNNs should be kept the same as with symmetric full rank RNNs eq.(1.4, 1.10). To keep the steady state of responses stable, the eigenvalues $\lambda_i \in R$ of the symmetric low-rank RNNs J in eq.(1.54) are limited by $R < 1$ through normalizing with the maximal eigenvalue as done by eq.(1.3).

Asymmetric low-rank RNN When the set of left connectivity vector $\left\{l^{(g)}\right\}$ is different from the set of right connectivity vector $\left\{r^{(g)}\right\}$, we receive the asymmetric low-rank RNN with eq.(1.52) for the case without random noise. If considering the random noise, we again add a simple $n \times n$ Gaussian distributed full rank asymmetric matrix as Figure 1.5 illustrated with definition eq.(1.53).

To enable the stable steady state of responses, as shown in asymmetric RNNs in eq.(1.41), the real part of eigenvalues $Re(\lambda_i)$ of an asymmetric RNN with

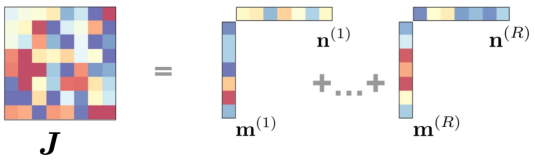


Figure 1.4: **Low-rank recurrent networks (RNNs) constructed with distinct Gaussian distribution** [?]. A low-rank matrix could be written as a sum of outer products of vectors that are Gaussian distributed. As a result, the connectivity matrix is a mixture of Gaussians.

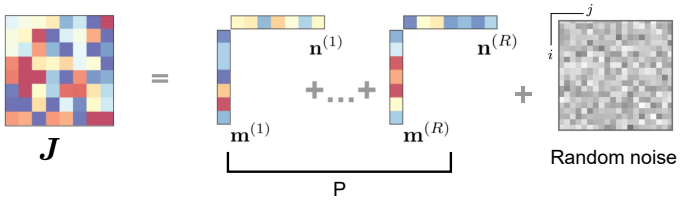


Figure 1.5: **Low-rank recurrent networks (RNNs) constructed with fixed part and random noise** [?]**.** The connectivity matrix is given by the sum of a structured, controlled matrix P and an uncontrolled, random matrix. Except for the fixed and known part P, which is a mixture of uncorrelated Gaussians, the RNN is disturbed by random noise.

interaction matrix J defined in eq.(1.53) are limited by $R < 1$ through normalization with the maximal magnitude of eigenvalues as done in eq.(1.42).

1.3.3 Evaluation of Feedforward Recurrent Alignment Hypothesis Based on Response Properties

Similar to the analysis on full-rank RNNs, we also want to evaluate the application of feedforward recurrent alignment with low-rank RNNs based on the four response properties, especially with the focus on the correlation between response properties and feedforward recurrent alignment score. Those four properties are:

- Trial-to-trial correlation β_s defined in eq.(1.18).
- Intra-trial stability $\bar{c}$ defined in eq.(1.23).
- Dimensionality in both analytical and empirical formulations $d_{eff,ana}$ and $d_{eff,emp}$ defined in eq.(1.31) and eq.(1.32).
- Alignment of evoked activity pattern to spontaneous activity pattern γ defined in eq.(1.34).

Symmetric low-rank RNNs For symmetric low-rank RNNs, the methods for modeling and evaluations are traced back to the full-rank symmetric RNNs before in section 1.1.4. The feedforward recurrent alignment is defined by eq.(1.11). Since the eigenvectors and eigenvalues for symmetric low-rank RNNs are real vectors and real numbers, the definitions and methods for symmetric full-rank RNNs can be directly applied in the same way.

Asymmetric low-rank RNNs Since for asymmetric low-rank RNNs, the problem of complex eigenvectors and eigenvalues still exists, the modifications for dealing with this problem at asymmetric full-rank RNNs (section 1.2.2 and 1.2.3) can therefore be directly applied here when aligning the inputs to asymmetric low-rank RNNs. Generally, the modifications that are considered can be roughly described as 1) only consider the real part of complex eigenvalues and eigenvectors, 2) consider the magnitude of complex eigenvectors and eigenvalues, 3) align the inputs to symmetrized network to approximate. Apply the eigenvectors and eigenvalues from the symmetrized interaction matrix.

1.4 Black Box Recurrent Network Model

Until now, we assume that we already know the structure of recurrent networks and evaluate the feedforward recurrent alignment hypothesis on the networks. However, in reality during the experiments, mostly only a small part of the network is known and the total structure of the network is mostly like a black box. As a result, dominant directions for generating reliable neural activity with networks are generally unknown. The feedforward recurrent alignment cannot use the dominant modes to characterize the development of feedforward recurrent network systems any more.

It was pointed out that the reliability of evoked dynamics in recurrent networks is dependent on the stimulus used. As a consequence, a recurrent network would correspond to a set of stimuli that are more efficiently transmitted than others [**?**]. Especially the stimulus inputs that align with the structure of endogenous sub-networks would be recurrently amplified, leading to more reliable evoked responses [**?**].

Besides, the similarity between spontaneous and evoked activity in sensory cortical areas could be a signature of efficient transmission and propagation across cortical networks. Based on a better recall caused by a match between spontaneous activity and input statistics, it was hypothesized that the recurrent connectivity could have been shaped by a learning process so that the spontaneous activity matches the natural input statistics to increase the efficient transmission [**?**].

Therefore, we wonder if we could apply spontaneous-like activity to characterize the feedforward recurrent alignment. Without knowing the eigenvectors of the recurrent interactions, align feedforward inputs to the spontaneous-like activity patterns instead. This can be realized for general asymmetric RNNs with a further modification at modeling feedforward recurrent alignment.

To model the spontaneous-like activity, the lab of H.Mulholland [**?**] suggested on possible way with white noise. Thus, for aligning inputs to spontaneous-like activity, we model it with alignment to white-noise-evoked activity.

Furthermore, we consider repeatedly applying the recurrently amplified spontaneous stimuli as inputs to discover the effect of repetitions of white-noise-evoked activity on feedforward recurrent alignment.

1.4.1 Approximation with White Noise Evoked Activity

With an unknown recurrent structure, which is in general asymmetric, it is then difficult to find the stimuli pattern such that the trial-to-trial correlation, intra-trial stability, and alignment between evoked activity to spontaneous activity are high while keeping dimensionality low. In other words, we cannot apply the eigenvectors of the recurrent interaction to align with inputs and then characterize the development of feedforward inputs leading to stable response properties.

Alternatively, we align the inputs to white-noise-evoked activity and explore the response properties in correlation with feedforward recurrent alignment.

White noise is modeled by multivariate normal distribution with mean vector the zero vector $0_v \in R^{n\times 1}$ and covariance matrix the identity matrix $I_n \in R^{n\times n}$,

$$h_{white} \sim \mathcal{N}(0_v, I_n)\,. \tag{1.55}$$

The white-noise-evoked activity is then modeled by the transformed steady-state response $r \in R^{n\times 1}$, which is also multivariate normal distributed with transformed covariance matrix (section 1.1.4),

$$r_{white} \sim \mathcal{N}\left(0_v, (1-J)^{-1}(1-J)^{-T}\right)\,. \tag{1.56}$$

The form of response pattern is determined by the covariance matrix. If aligning the inputs to response patterns, the eigenvectors of covariance matrices are aligned. The eigenvectors of a covariance matrix are also known as principal components for the distribution.

To model the feedforward recurrent alignment hypothesis, the inputs are now aligned to principal components of white-noise-evoked activity and the feedforward alignment score is formulated with principal components instead of with modified eigenvectors of asymmetric recurrent network as in section 1.2.2. Since covariance matrix is symmetric, its eigenvectors and eigenvalues are of real numbers. For an input h aligned to a principal component p, the feedforward recurrent alignment is constructed with

$$\nu := \frac{p^T J p}{\|p\|^2}\,. \tag{1.57}$$

In the newly defined feedforward recurrent alignment eq.(1.57), the original recurrent network is now approximated by the spontaneous-like response patterns and the inputs are aligned to the principal components of the spontaneous-like response pattern. The same as prior in the work, some properties of the newly formulated feedforward recurrent alignment score are going to be evaluated, and especially the expected correlations between response properties and feedforward recurrent alignment eq.(1.57) should be fulfilled. The perspectives that we take into account are:

- Monotonously positive correlation between feedforward recurrent alignment score and eigenvalues of covariance matrix from the white-noise-evoked pattern.
- Positive correlation between feedforward recurrent alignment score and trial-to-trial correlation.
- Positive correlation between feedforward recurrent alignment score and intra-trial stability.
- Negative correlation between feedfroward recurrent alignment score and dimensionality.
- Feedforward recurrent alignment score is positively correlated with alignment of evoked activity to spontaneous activity.

Monotony The feedforward recurrent alignment should reflect how well the input pattern is aligned with the considered recurrent network. The more the input pattern is aligned with the dominant projection direction in activity space spanned by eigenvectors of recurrent interaction, the stronger should be the evoked response.

The response strength is determined by the corresponding eigenvalue of the aligned direction. Large eigenvalues result in stronger evoked activity [**?**]. Since the inputs are aligned to the white-noise-evoked activity pattern, the feedforward recurrent alignment should be monotonously positively correlated with the eigenvalues of the white-noise-evoked activity pattern.

Trial-to-trial correlation We consider the case of general asymmetric recurrent network from section 1.2.1 with formulation considering different grade of symmetry in eq.(1.38) for theoretical exploration. Thus, the modification is similar to section 1.2.3. The input pattern h is aligned to the principal component p of the covariance matrix from white-noise-evoked activity pattern and therefore modeled by

$$h \sim \mathcal{N}(p, \sigma_{trial} I_n)\,. \tag{1.58}$$

The steady-state response evoked by the inputs are transformed multivariate normal distribution

$$r \sim \mathcal{N}\left((1-J)^{-1}p, \sigma_{trial}(1-J)^{-1}(1-J)^{-T}\right)\,. \tag{1.59}$$

The trial-to-trial correlation reflects the variation between different trials from one stimulus. It is the average of pairwise Pearson correlation between response trials as defined by β_s with eq.(1.18).

Intra-trial stability Intra-trial stability quantifies the variation inside one response trial evoked by input. One time-dependent input and evoked steady-state response trial is approximated by Euler-Maruyama scheme described by (1.19). The input pattern h is aligned to principal components p of white-noise-evoked activity pattern. Therefore, the mean vector for input distribution is the aligned principal component p. The input and evoked response are therefore approximated as $\mathrm{d\ h\ = p\ dt} + \sigma_{time}\mathrm{d}W$
$\mathrm{d}r = (-r + J \cdot p)\mathrm{d}t + \sigma_{time}\mathrm{d}W$. The intra-trial stability is the time average of delayed-response correlation defined by eq.(1.24).

Dimensionality For modeling the change in dimensionality against alignment score, the covariance matrix for input distribution is constructed as eq.(1.26) but with principal components p_i of white-noise-evoked activity as an approximation to the eigenvectors of the original recurrent network,

$$\Sigma^{Dim} := \sum_{i=L}^{L+M_{dim}} exp\left(\frac{-2(i-L)}{\beta_{dim}}\right) p_i p_i^T\,. \tag{1.60}$$

The input and the evoked activity are modeled by multivariate normal distribution $\mathrm{h} \sim \mathcal{N}(0_v, \Sigma^{Dim})$
$r \sim \mathcal{N}(0_v, (I_n - J)^{-1}\Sigma^{Dim}(I_n - J)^{-T})$. The effective dimensionality from is approximated by the eigenvalues of covariance matrix from white-noise-evoked activity pattern based on eq.(1.32).

Alignment between evoked activity and spontaneous activity To grantee the spontaneous has a broader input than the evoked activity, the spontaneous activity for alignment to evoked activity is constructed similarly to eq.(1.60) with a higher dimensioality $\beta_{spont} > \beta_{dim}$. For the formulation of covariance matrix for spontaneous activity, the principal components from white-noise-evoked activity pattern is use,

$$\Sigma^{spont} := \sum_{i=L}^{M_{spont}+1} exp\left(\frac{-2(i-1)}{\beta_{spont}}\right) p_i p_i^T . \tag{1.61}$$

The spontaneous activity is then modeled by

$$r \sim \mathcal{N}(0_v, (I_n - J)^{-1}\Sigma^{spont}(I_n - J)^{-T}) . \tag{1.62}$$

The amount of overlap between evoked activity pattern generated with eq.(1.60b) and the principal components of spontaneous activity from eq.(1.62) quantifies the alignment between them. The average alignment over all N evoked-response trials is the final alignment to spontaneous activity.

$$\gamma = \frac{1}{N}\left(\frac{r_i^T \cdot \Sigma^{spont} \cdot r_i}{\|r_i\|^2 Tr(\Sigma^{spont})}\right) \tag{1.63}$$

1.4.2 Iterative Approximation with Low Dimensional Inputs

Low dimensional inputs can be generated experimentally easier than high dimensional inputs. So, we wonder if the feedforward recurrent alignment can also be adapted to represent the development and better alignment under the settings that, 1) only low dimensional inputs are offered and 2) the original recurrent network is asymmetric but unknown.

To model the low dimensional input, random orthonormal basis vectors e_i for construction of covariance matrix are obtained through Gram-Schmidt process. The same scheme as for the construction of covariance Σ^{Dim} from eq.(1.26) is applied,

$$\Sigma_{Low} := \sum_{i=0}^{M} exp\left(\frac{-2(i-1)}{\beta_{Low}}\right) e_i e_i^T . \tag{1.64}$$

Low dimensionality is realized by parameter β_{Low}, which should be smaller than it in Σ^{Dim} and Σ^{spont}.

The low dimensional inputs are then modeled by multivariate normal distribution with zero vector $0_v \in R^{n\times 1}$ as mean vector and Σ_{Low} as covariance matrix,

$$h_{Low} \sim \mathcal{N}(0_v, \Sigma_{Low}) . \tag{1.65}$$

The response evoked by low dimensional input from eq.(1.65) is modeled by the linearly transformed multivariate normal distribution

$$r_0 \sim \mathcal{N}\left(0_v, (1-J)^{-1}\Sigma_{Low}(1-J)^{-T}\right) . \tag{1.66}$$

At this step, the feedforward recurrent alignment ν can be calculated with low dimensional input h

$$\nu_0 = \frac{h^T J h}{\|h\|^2} . \tag{1.67}$$

Prior knowledge predicts stimulus inputs that align to spontaneous activity can be recurrently amplified and lead to more reliable responses [**?**]. We, therefore, wonder if the responses also align better with the recurrent network and if the repeated application of responses leads to a development of input alignment to the recurrent network. In other words, if the alignment of iterative responses could also capture the neural development process in a certain way.

For that, we apply repeatedly the prior response as the input for the network and at the same time update step-wise the corresponding feedforward recurrent alignment.

If r_0 is the input for the recurrent network, the evoked response r_1 is due to the linear transformation of normal distribution also a multivariate normal distribution with linearly transformed variance,

$$r_1 \sim \mathcal{N}\left(0_v, \left((1-J)^{-1}\right)^2 \Sigma_{Low} \left((1-J)^{-T}\right)^2\right) . \tag{1.68}$$

The corresponding feedforward recurrent alignment, noted as ν_1, is determined by the input r_0,

$$\nu_1 = \frac{r_0^T J r_0}{\|r_0\|^2} . \tag{1.69}$$

Iteratively, at the n-th time of applying prior response r_{n-1} as input, the evoked response r_n has the general formulation of a multivariate normal distribution with linearly transformed covariance matrix,

$$r_n \sim \mathcal{N}\left(0_v, \left((1-J)^{-1}\right)^{n+1} \Sigma_{Low} \left((1-J)^{-T}\right)^{n+1}\right) . \tag{1.70}$$

The corresponding n-th feedforward recurrent ν_n alignment has the general formulation with its input r_{n-1},

$$\nu_n = \frac{r_{n-1}^T J r_{n-1}}{\|r_{n-1}\|^2} . \tag{1.71}$$

1.5 Hebbian Learning in Feedforward Recurrent Networks

In the neocortex, which forms the convoluted outer surface of the human brain, neurons lie in six vertical layers highly coupled within cylindrical columns. There are multiple types of connections. Feedforward connections bring input to a given region from another region located at an earlier stage along a particular processing pathway. Recurrent synapses interconnect neurons within a particular region that are considered to be at the same stage along the processing pathway [**?**].

Until now, we only considered the recurrent layer under given feedforward inputs (Figure 1.1b and Figure 1.2b). However, the feedforward inputs are also the outputs from the feedforward network. In this part of the work, we will expand the network structure to a feedforward recurrent network containing an input layer, feedforward interaction, and output layer connected by recurrent interactions.

Activity-dependent synaptic plasticity is widely believed to be the basic phenomenon underlying learning and memory, and it is also thought to play a crucial role in the development of neural circuits [**?**]. To count on this essential characteristic of networks during their development, we integrate the plasticity in the network dynamic.

The fundamental rule for synaptic plasticity in learning and memory is called the Hebbian rule, raised by Donald Hebb in 1949. Hebb suggested that if input from neuron A often contributes to the firing of neuron B, then the synapse from A to B should be strengthened. In other words, neurons that fire together, wire together. The activity-dependent synaptic plasticity of the Hebbian type refers to the plasticity that is based on the correlation of pre-and postsynaptic firing [**?**].

There are different types of training procedures for Hebbian-type plasticity, including unsupervised learning, supervised learning, reinforcement learning, and so on. we mainly focus on unsupervised learning. Unsupervised learning provides a model for the effects of experience on mature networks [**?**]. Based on our network model (Figure **??**), we consider the case in which there are multiple postsynaptic neurons.

1.5.1 Model Setting

To relieve the understanding of dynamics during the modeling, we start with the simplified assumption of random symmetric recurrent interaction J and linear feedforward recurrent networks.

Two basic cases are considered for our start-up modeling: 1) only one input rate u, and 2) only Hebbian learning update of feedforward interaction W. The output layer is occupied by a number of n output neurons.

The feedforward interaction W can be described as a $R^{n\times 1}$ dimensional vector containing W_i the strength of connection between i-th output neuron and input neuron. Output rates $v \in R^{n\times 1}$ describes the activity firing rate for output neurons. The random symmetric interaction $J \in R^{n\times n}$ is constructed in

the same way as for feedforward recurrent alignment hypothesis for symmetric RNNs in section 1.1.1.

The output rates $v \in R^{n\times 1}$ in this linear case is determined by

$$\tau \frac{dv}{dt} = -v + W \cdot u + J \cdot v \,. \tag{1.72}$$

For simplicity, the time scale constant is set to be 1. $h := W \cdot u \in R^{n\times 1}$ summarizes the feedforward input for recurrent interaction part.

Provided that the real parts of the eigenvalues of J are less than 1, this equation has a stable fixed point with a steady-state output activity vector determined by (shown in section 1.1.2 and 1.2.1)

$$v = W \cdot u + J \cdot v \,. \tag{1.73}$$

Solving the above equation (1.73), the steady state response of output layer for the feedforward recurrent network is

$$v^* = (I_n - J)^{-1} \cdot W \cdot u \,. \tag{1.74}$$

1.5.2 Update Rules for Feedforward Network

For multiple postsynaptic neurons with fixed recurrent weights $J \in R^{n\times n}$ and plastic feedforward weights $W \in R^{n\times 1}$, the basic Hebbian modification over the training input $u \in R$ is

$$\tau_w \frac{dW}{dt} = vu \,, \tag{1.75}$$

where τ_w is a time constant that controls the rate at which the weights change and for simplicity is set to be 1. If and only if both presynaptic rate u and postsynaptic rate v have the same signs in rates, meaning that both pre-and postsynaptic activities are both suppressed or activated at the same time, the connection weight between those two neurons increases. Therefore, the eq.(1.75) implies that simultaneous pre-and postsynaptic activity increases the feedforward-synaptic strength.

To compute the weight changes induced by a series of input patterns u, a convenient alternative is to average over all of the different input patterns and compute the weight change induced by this average, leading to the average Hebbian rule:

$$\frac{dW}{dt} = \langle vu \rangle \,, \tag{1.76}$$

with angle brackets $\langle\,\rangle$ denoting averages over the ensemble of input patterns presented during training.

Replace the output rate v in eq.(1.76) with the steady state formulation eq.(1.74), the average plasticity rule can be rewrote as

$$\frac{dW}{dt} = \langle (I_n - J)^{-1} \cdot W \cdot uu \rangle = (I_n - J)^{-1} \cdot W \cdot \langle uu \rangle \,. \tag{1.77}$$

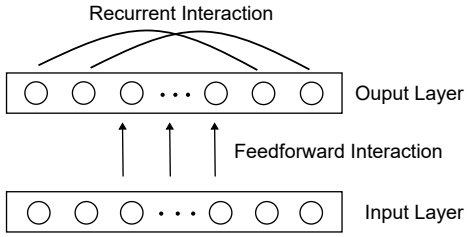


Figure 1.6: **Illustration of a general feedforward recurrent network construction.** Shown in the figure is a feedforward recurrent network with an input layer, an output layer, a feedforward synaptic weight matrix for feedforward interactions, and a recurrent synaptic weight matrix for recurrent interactions. Firing rate models are applied in both the input and output layer for modeling.

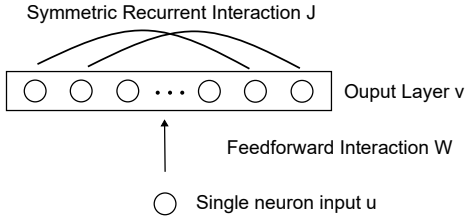


Figure 1.7: **Illustration of feedforward recurrent network model with single input neuron.** For start-up of feedforward recurrent network dynamic analysis, the simple case of only one input neuron with fixed random symmetric recurrent interaction J. The output layer has a number of n neurons. The feedforward interaction matrix is updated with the Hebbian rule.

Defining $Q := \langle uu \rangle$ the input autocorrelation, the final formulation of the average rule is

$$\frac{dW}{dt} = (I_n - J)^{-1} \cdot W \cdot Q \,. \tag{1.78}$$

Since for the simple case here that $u \in R$, the input autocorrelation $Q = 1$ for all input patterns. Therefore, specifically for the simple network (Figure 1.7) we consider, the average rule can be further simplified as

$$\frac{dW}{dt} = (I_n - J)^{-1} \cdot W \,. \tag{1.79}$$

1.5.3 Projection of the Feedforward Weights on Eigenvectors

With the average Hebbian rule, the dynamic of the feedforward weight W can be approximated with the help of Euler scheme, resulting the iterative update for W at time point $t + 1$

$$W_{t+1} = W_t + \Delta t (I_n - J)^{-1} \cdot W_t \,, \tag{1.80}$$

where Δt is a enough small time distance during the total time period T_{Hebb}. The initial weight W_0 for iteration is random Gaussian distributed.

As a result, the change of connections between each output neuron and the single input neuron over time can be approximated through eq.(1.80). The distribution of the weights can be important information about which connections are strengthened over time by unsupervised learning of feedforward interaction.

Moreover, the output from feedforward interaction $h := W \cdot u$ is exactly the feedforward input for the recurrent interaction. Continuing the idea from the feedforward recurrent alignment hypothesis, aligning the feedforward input h well to the recurrent network could predict the reliability of evoked activity. Since $u \in R$, the feedforward input for the recurrent network is proportional to the feedforward weight vector W. Therefore, we align directly the feedforward weight vector W to the recurrent network through the projection of vector W to space spanned by eigenvectors of the recurrent network.

Recurrent network has symmetric interaction J, thus the eigenvectors $\{e_i\}_{i=1,...,n}$ of J is a set of basis vectors that span the vector space R^n. The feedforward weight vector $W \in R^{n \times 1}$ can be thus expressed as linear combination of eigenvectors $\{e_i\}$. For better interpretation, $\{e_i\}$ are ordered so that their corresponding eigenvalues $\{\lambda_i\}$ are in descending order:

$$e_{max}, ..., e_i, e_j, ..., e_{min} \; suchthat \; \lambda_{max} > ... > \lambda_i > \lambda_j > ... > \lambda_{min} \,. \tag{1.81}$$

The feedforward weight vector W_t at time point t can be formulated as the linear combination

$$W_t = \sum_{i=1}^{n} \phi_i e_i = A\phi \,, \tag{1.82}$$

where $\phi \in R^{n \times 1}$ is the vector containing all projection coefficients ϕ_i and A is the matrix containing the eigenvectors e_i column-wise.

As a result, the projection coefficient can be gained through equivalent reformulation of eq.(1.82),

$$\phi = A^{-1} \cdot W_t \,. \tag{1.83}$$

Matrix A is invertible because all columns of A are linearly independent.

The distribution of ϕ_i reflects to which eigenvector-directions the feedforward weight vector W aligns to. Since the feedforward input h is proportional to W, the distribution also determines the directions to which the input h is aligned. According to the feedforward recurrent alignment hypothesis, in an experienced feedforward recurrent network, the feedforward input aligns to dominant eigenvectors of the recurrent interaction matrix, those are the eigenvectors with large eigenvalues. The strength of alignment to dominant eigenvectors can be reflected by the projection coefficients for them.

To more directly quantify the change of alignment of W to dominant eigenvectors, we define the projection ratio ρ as the percentage that the coefficients for the first twenty eigenvectors [2] take over all coefficients. Since the sign of coefficients only reflects the direction of alignment, we consider the absolute value of coefficients for the calculation of projection ratio ρ.

$$\rho := \frac{\sum_{i=1}^{20} |\phi_i|}{\sum_{i=1}^{n} |\phi_i|} \,, \tag{1.84}$$

where n is the total number of output neurons.

1.5.4 Dynamics of Feedforward Recurrent Alignment

An alternative to discover the feedforward recurrent alignment hypothesis is to directly observe the dynamic of feedforward recurrent alignment score during Hebbian learning. With the time-dependent update of feedforward weight vector W by Euler scheme in eq.(1.80), the feedforward input for recurrent network h can be updated simultaneously through multiplying input rate u,

$$h_t = W_t \cdot u_t \,, \tag{1.85}$$

with h_t, W_t, u_t the feedforward input for recurrent network, feedforward interaction, and input firing rate at time point t.

The development of feedforward recurrent alignment score over time can be then calculated with its definition inserting the updated feedforward input for recurrent network h_t. The feedforward recurrent at time point would be ν_t with

$$\nu_t = \frac{h_t^T J h_t}{\|h_t\|^2} = \frac{u_t W_t^T J W_t u_t}{u_t^2 \|W_t\|^2} = \frac{W_t^T J W_t}{\|W_t\|^2} \,. \tag{1.86}$$

Because $u_t \in R$ for all t, the feedforward recurrent alignment ν_t is directly determined by feedforward weight vector W_t.

The derivative of feedforward recurrent alignment ν_t can illustrate the change over time more intuitively. For further calculation, firstly the derivative of the

[2] Under the initial condition that the number of neurons n is larger than 20.

numerator and denominator of ν_t in eq.(1.86). Defining thereby the numerator as $g(t)$ and denominator $h(t)$. Without loss of generality, the Euclidean norm is applied for denominator. That is $\mathrm{g(t)} := \mathrm{W}_t^T J W_t$.
$h(t) := \|W_t\|_2^2$. Applying product rule for derivative of $g(t)$, it results in

$$\frac{\mathrm{d}g(t)}{\mathrm{d}t} = \frac{\mathrm{d}W_t^T}{\mathrm{d}t} J W_t + W_t^T J \frac{\mathrm{d}W_t}{dt} \,. \tag{1.87}$$

Inserting the Hebbian rule eq.(1.79) for feedforward weights leads to

$$\frac{\mathrm{d}g(t)}{\mathrm{d}t} = W_t^T (I_n - J)^{-T} J W_t + W_t^T J (I_n - J)^{-1} W_t \,. \tag{1.88}$$

Since the recurrent network interaction J is symmetric, the matrix $(I_n - J)$ and its inverse matrix is also symmetric. Therefore, $(I_n - J)^{-T} = (I_n - J)^{-1}$. So, the final derivative of numerator $g(t)$ is

$$\frac{\mathrm{d}g(t)}{\mathrm{d}t} = W_t^T \left((I_n - J)^{-1} J + J (I_n - J)^{-1} \right) W_t \,. \tag{1.89}$$

The derivative of denominator $h(t)$ can be obtained after inserting the Euclidean norm and applying chain rule for its derivative,

$$\frac{\mathrm{d}h(t)}{\mathrm{d}t} = \frac{\mathrm{d}\|W_t\|_2^2}{\mathrm{d}t} = \frac{\mathrm{d}\sum_{i=1}^n W_{t,i}^2}{\mathrm{d}t} = \sum_{i=1}^n \frac{\mathrm{d}W_{t,i}^2}{\mathrm{d}t} = 2 W_t^T \frac{\mathrm{d}W_t}{\mathrm{d}t} = 2 W_t^T (I_n - J)^{-1} W_t \,. \tag{1.90}$$

The last equation is due to the Hebbian rule from eq.(1.79).

Following the quotient rule with derivatives of numerator $g(t)$ and denominator $h(t)$, the derivative for feedforward recurrent alignment score ν_t is

$$\frac{\mathrm{d}\nu_t}{\mathrm{d}t} = \frac{\frac{\mathrm{d}g(t)}{\mathrm{d}t} h(t) - g(t) \frac{\mathrm{d}h(t)}{\mathrm{d}t}}{h(t)^2} = \frac{W_t^T \left((I_n - J)^{-1} J + J (I_n - J)^{-1} \right) W_t \|W_t\|_2^2 - 2 W_t^T J W_t W_t^T (I_n - J)^{-1} W_t}{\|W_t\|_2^4} = \frac{V}{\|V\|} \tag{1.91}$$

To simplify the notation, define the normalized feedforward weights as $\bar{W}_t$. Thus, the final formulation of the derivative for feedforward recurrent alignment score ν_t at time point t is

$$\frac{\mathrm{d}\nu_t}{\mathrm{d}t} = \bar{W}_t{}^T (I_n - J)^{-1} J \bar{W}_t + \bar{W}_t{}^T J (I_n - J)^{-1} \bar{W}_t - 2 \bar{W}_t{}^T J \bar{W}_t \bar{W}_t{}^T (I_n - J)^{-1} \bar{W}_t \,. \tag{1.92}$$