1 Methods

In this chapter, we will give an overview about the recurrent network (RNN) models for exploration of the feedforward recurrent alignment hypothesis that are 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 characters of symmetric RNNs, they are often applied in models for neuroscience for a better understanding of certain dynamics. Therefore, we firstly consider the basic case of having a symmetric recurrent network, which has the 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.

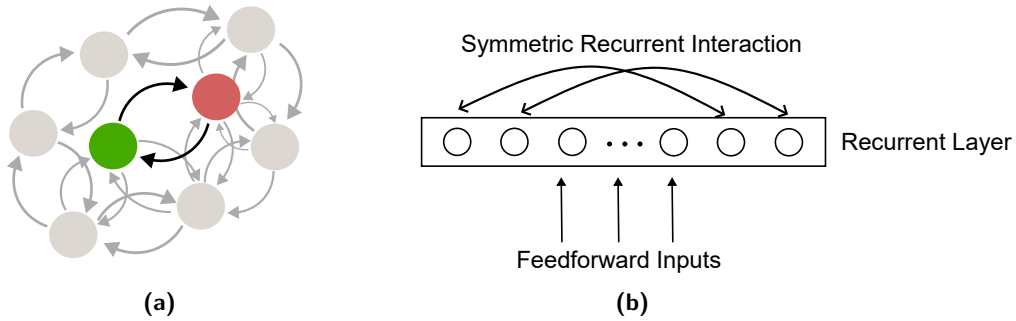

Figure 1.1 Illustration of symmetric recurrent networks (symmetric RNNs). **(a)** An example of symmetric connections between two neurons in a network with multiple neurons. If there are connections between two neurons, here for example the green and red neurons, the directed connection from the green 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).

1.1.1 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,

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

where n is the number of neurons involved in the RNN. The eigenvalues $\{\lambda_i\}_{i=1,...,n}$ of J are limited by parameter $R < 1$ through

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

$\tilde{\lambda}_i$ are the original eigenvalues of J. 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 **??** could be described as [**?**]:

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

with the vector $r \in \mathbb{R}^{n\times 1}$ describing responses of neurons in the recurrent layer, the vector $h \in \mathbb{R}^{n\times 1}$ as feedforward inputs ,and τ_r the time constant controlling the speed of dynamic. The steady state of the dynamic system 1.4 can be received by setting the ordinary differential equation to zero. For simplicity, the time constant is set to one. We then have

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

r^* the steady state for responses. $I_n \in \mathbb{R}^{n\times n}$ is 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 1.4 could also be written in an elementary expression:

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

The derivative of f_i to r_j is

$$\frac{\partial f_i}{\partial r_j} = \begin{cases} -1 + J_{ij} & \text{if } i = j \\ J_{ij} & \text{if } i \neq j \end{cases}\,. \tag{1.7}$$

The Jacobian matrix A of the dynamic system $\frac{\mathrm{d}r}{\mathrm{d}t}$ is then

$$A := \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix} = -I_n + J\,. \tag{1.8}$$

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 $E := \{e_i\}_{i=1,\dots,n}$ the matrix containing eigenvectors of J column-wise,

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

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

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

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

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 quick enough, we could apply the steady state r^* for further analysis.

1.1.3 Feedforward Recurrent Alignment for Symmetric Interactions

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

$$\nu := \frac{h^T J h}{\parallel h \parallel_2^2} \tag{1.11}$$

If the inputs are aligned to the eigenvectors e_i of the recurrent interaction J, i.e.,

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

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

$$\nu = \frac{h^T J h}{\parallel h \parallel_2^2} \propto \frac{{e_i}^T J e_i}{\parallel e_i \parallel_2^2} = \frac{\lambda_i {e_i}^T e_i}{\parallel e_i \parallel_2^2} = \lambda_i\,. \tag{1.13}$$

It was therefore observed that the maximal alignment was attained when the input was proportional to the eigenvector $e_{\max}$ with maximal eigenvalue $\lambda_{\max}$ [**?**].

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

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. Large correlation implies high reliability of the response generated by the RNN.

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

$$h \sim \mathcal{N}(\mu, \Sigma)\,. \tag{1.14}$$

Then, the steady state response $r^* = (I_n - J)^{-1} \cdot h$ from (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 of the multivariate normal distribution

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

Therefore, the moment-generating function of the random vector r^* becomes

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

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

As calculated in [**?**], the trial to trial correlation β 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.

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

To model the time dependent input $h(t) \in \mathbb{R}^{n\times 1}$ distributed as (1.14) and its evoked steady state responses $r(t) \in \mathbb{R}^{n\times 1}$, the following stochastic differential equations were formulated

$$\mathrm{d}h = \mu \mathrm{dt} + \sigma_{\text{time}} \mathrm{d}W \tag{1.19a}$$
$$\mathrm{d}r = (-r + J \cdot \mu)\mathrm{d}t + \sigma_{\text{time}} \mathrm{d}W\,, \tag{1.19b}$$

with W the Wiener process, which is a continuous-time stochastic process with independent Gaussian increments.

To approximate the evoked response $r(t)$, the equation (1.19b) is solved numerically with Euler-Maruyama scheme

$$r_{t+1} = r_t + (-r_t + J \cdot \mu)\Delta t + \sigma_{\text{time}} \sqrt{\Delta t} \Delta \tilde{W}_t\,, \tag{1.20}$$

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

$$\Delta \tilde{W}_t \sim \mathcal{N}(0_v, I_n)\,. \tag{1.21}$$

For a certain step width $\tilde{\Delta} t$, the intra-trial stability $c(\Delta \tilde{t})$ was defined by the correlation between normalized response at time t and its delayed response at time $t + \Delta \tilde{t}$

$$c(\Delta \tilde{t}) := \bar{r}(t)^T \bar{r}(t + \Delta t)\,, \tag{1.22}$$

where the normalized response is defined as

$$\bar{r}(t) := \frac{r - \langle r \rangle}{\sigma_r}\,, \tag{1.23}$$

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 $0 \leq t \leq T - \Delta \tilde{t}$

$$\begin{aligned} \bar{c}(\Delta \tilde{t}) &= \frac{1}{T - \Delta \tilde{t}} \int_0^{T - \Delta \tilde{t}} c(\Delta \tilde{t}) \mathrm{d}t \\ &= \frac{1}{T - \Delta \tilde{t}} \int_0^{T - \Delta \tilde{t}} \bar{r}(t)^T \bar{r}(t + \Delta t) \mathrm{d}t\,. \end{aligned} \tag{1.24}$$

Dimensionality The dimensionality of neuron responses reflect the complexity of the informatbetaion they encoded. A diverse response pattern corresponds with a broader distribution of the variance over principal components, leading to a higher-dimensional linear manifold. The corresponding dimensionality for more diverse

and variable response pattern is therefore higher [**?**]. Given the multivariate normal distributed inputs $h \in \mathbb{R}^{n\times 1}$

$$h \sim \mathcal{N}(0_v, \Sigma^{\mathrm{Dim}})\,, \tag{1.25}$$

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

$$r \sim \mathcal{N}(0_v, (I_n - J)^{-1}\Sigma^{\mathrm{Dim}}(I_n - J)^{-T}) \tag{1.26}$$

with

$$\Sigma^{\mathrm{Dim}} := \sum_{i=L}^{L+M} \exp\left(\frac{2(i-L)}{\beta}\right) e_i e_i^T\,, \tag{1.27}$$

in which the parameter $M := \kappa\beta$ and β 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 $\mathbb{R}^n$, they could be chosen as e_i for (1.27). Hereby, the eigenvectors are ordered according to their eigenvalues in descending order. The exponential factor in (1.27) simulates the exponential decay of variance ratio observed in prior data [**?**].

The linear effective dimensionality based on participation ratio was defined to quantify the tendency of dimensionality during visual maturation. The participation ratio is defined as

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

where λ_i the eigenvalues of a certain response pattern with covariance $\mathbf{\Sigma}$. Since as defined in (1.27), Σ^{Dim} has the same eigenvectors (aka. principal components)as J. Therefore, the eigenvalues λ_i^{Dim}(aka. variance ratio or variance explained) for Σ^{Dim} are transformed eigenvalues λ_i of J expressed as

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

The covariance of the responses share the same eigenvectors as Σ^{Dim} based on its distribution (1.26) and therefore also the same as J. The eigenvalues (aka. variance ratio) for the responses could be constructed analogously with

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

for $i = L, ..., L+M$, due to the inverse transformation.

Insert the eigenvalues of responses (1.30) in the formula for effective dimensionality to get the final formulation of dimensionality for responses

$$d_{\mathrm{eff}}^r = \frac{\left(\sum_{i=L}^{L+M} \exp\left(-2\frac{i-L}{\beta}\right)(1-\lambda_i)^{-2}\right)^2}{\sum_{i=L}^{L+M} \exp\left(-4\frac{i-L}{\beta}\right)(1-\lambda_i)^{-4}} \tag{1.31}$$

Since the vector of explained variance ratios in the principal component analysis (PCA) is the normalized vector containing eigenvales of the covariance matrix re-scaled by the largest eigenvalue in descending order, which then explains how much variance does the corresponding principal component contribute. 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 dimentionality, i.e.,

$$d_{\text{eff}} = \frac{\left(\sum_{i=L}^{L+M} \text{var}_i\right)^2}{\sum_{i=L}^{L+M} \text{var}_i^2} \tag{1.32}$$

with var_i the i-th variance ratio.

Alignment with spontaneous activity The alignment of activity patterns to spontaneous activity is in principle the dimensionality explained by the principal components of spontaneous activity. Assuming having the evoked response pattern as R and the spontaneous activity pattern as S, we then have the variance of activity pattern R explained by principal components of pattern S expressed as a vector $\mathbf{v}$ with

$$\mathbf{v}_i = \frac{\mathbf{p}_{i,S}^T \cdot \mathbf{\Sigma_R} \cdot \mathbf{p}_{i,S}}{\text{Tr}(\mathbf{\Sigma_R})}\,, \tag{1.33}$$

for $i = 1, ..., n$. $\mathbf{p}_{i,S}$ are the principal component of spontaneous activity and $\mathbf{\Sigma_R}$ the covariance of evoked activity. It was observed in [**?**] that patterned visual experience is required to cause stable alignment between visual responses and the spontaneous activities, while naive visual responses show only loose alignment with spontaneous activities. To capture this change of property as a quantity instead of a vector as (**??**), it had to firstly define the alignment between two activity patterns. The alignment of R to S was defined as the average of pattern S explained by normalized trials $r_{i,R}$. The alignment between pattern S and one trial $r_{i,R}$ from R is defined as

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

where Σ_S is the covariance of pattern S. The final alignment between S and R is then the average value of alignment between S and all trials of R.

To model the inputs and responses, we assumed that spontaneous activity was evoked by inputs from broad sources. Besides, since the spontaneous activity already exists almost a week before eye opening, we assume that they already fit to 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 (1.25), that is higher dimensionality. Since the parameter β

in (1.25) indicates the dimensionality, we could set for spontaneous activity higher $\beta_{\text{spont}} > \beta$ to generate high dimensional inputs. Therefor, we than have the broad inputs $h^{\text{spont}} \in \mathbb{R}^{n\times 1}$ and spontaneous activity $r^{\text{spont}} \in \mathbb{R}^n$, which are multivariate distributed vectors

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

and

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

The covariance matrix $\mathbf{\Sigma}^{\text{spont}}$ is constructed in the same way as $\mathbf{\Sigma}^{\text{Dim}}$ only with $L = 1$ and larger β_{spont}, that is

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

1.2 Asymmetric Recurrent Network Model

1.2.1 Asymmetric Recurrent Interaction

1.2.2 Modification of Feedforward Recurrent Alignment for Asymmetric Interactions

1.2.3 Related Modification for Evaluation

1.3 Low Rank Recurrent Network Model

1.3.1 Construction of Low Rank Interactions

1.3.2 Modification of Feedforward Recurrent Alignment for Low Rank Interactions

1.4 Black Box Recurrent Network Model

1.4.1 Approximation with White Noise Evoked Activity

1.4.2 Iterative Approximation with Low Dimensional Inputs

1.5 Hebbian Learning in Feedforward Recurrent Networks

1.5.1 Model Setting

1.5.2 Update Rules for Feedforward Network

1.5.3 Projection of the Feedforward Weights on Eigenvectors