Theoretical Exploration of the Feedforward Recurrent Alignment Hypothesis

by

Rutian Zhou

at

Frankfurt Institute for Advanced Studies

A Master's Thesis

Submitted to the Institute for Informatics of the Department 12-Informatics and Mathematics of the Goethe University Frankfurt am Main

For the Degree of *Master of Science* in *Bioinformatics*

April 11, 2024

First Supervisor: Prof. Dr. Matthias Kaschube

Second Supervisor: Prof. Dr. Franziska Matthäus

Erklärung zur Abschlussarbeit

gemäß § 35, Abs. 16 der Ordnung für den Masterstudiengang Bioinformatik vom 17. Juni 2019:

Hiermit erkläre ich

_____

(Nachname, Vorname)

Die vorliegende Arbeit habe ich selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel verfasst.

Ebenso bestätige ich, dass diese Arbeit nicht, auch nicht auszugsweise, für eine andere Prüfung oder Studienleistung verwendet wurde.

Zudem versichere ich, dass die von mir eingereichten schriftlichen gebundenen Versionen meiner Masterarbeit mit der in elektronischer Form eingereichten Version dieser Masterarbeit übereinstimmen.

Frankfurt am Main, den

_____

Unterschrift der/des Studierenden

Zusammenfassung

Die theoretische Neurowissenschaft trägt dazu bei, die zugrundeliegenden Mechanismen neuronaler Systeme durch theoretische Modellierung und Analyse zu verstehen. Die Dynamik der kortikalen Netzwerke lässt sich mittels feedforward-rekurrenter Netzwerke modellieren. Wenn neuartige visuelle Eingaben durch feedforward-Eingaben repräsentiert werden, kann die Koordination zwischen dem feedforward- und dem rekurrenten Netzwerk die Zuverlässigkeit der finalen Repräsentationen der Antwort bestimmen. Dies kann dazu beitragen, die Mechanismen zu erforschen, die es endogen generierten Netzwerken ermöglichen, reife Repräsentationen mit dem Einsetzen sensorischer Erfahrungen zu bilden. Bisher beschränkt sich die konzeptuelle Modellierung jedoch auf den Fall idealisierter symmetrischer neuronaler Interaktionen. Hier erweitern wir die bisherige Modellierung um allgemeinere Interaktionsstrukturen der Netzwerke und untersuchen die Leistung der Modellierung unter komplexeren Bedingungen. Für asymmetrische neuronale Interaktionen kann die Symmetrisierung die meisten Musterinformationen bewahren und eine Anordnung in der Realzahlenebene ermöglichen. Das durch Weißrauschen hervorgerufene Aktivitätsmuster ist ein möglicher Kandidat für eine Anordnung, um unbekannte rekurrente Interaktionen anzunähern. Die Einbindung des Hebb'schen Lernens in das Modell zeigt das Potenzial der Plastizität zur Erforschung der Mechanismen. Die Arbeit demonstriert die verschiedenen Modifikationen für verschiedene Netzwerkstrukturbedingungen, die die feedforward-rekurrente Koordination ermöglichen. Die theoretischen Untersuchung der Koordination zwischen feedforward- und rekurrenten Netzwerken unter verschiedenen Interaktionsbedingungen können die bisherigen Modellierungsansätze verallgemeinern. Darüber hinaus können die Ergebnisse neue Perspektiven bieten, um das Verständnis der Mechanismen für durch Erfahrung gesteuerte Entwicklungen in kortikalen Netzwerken zu vertiefen.

Abstract

Theoretical neuroscience helps to understand the underlying mechanisms of neural systems by theoretical modeling and analysis. The dynamics of the cortical networks can be modeled by feedforward recurrent networks. If modeling novel visual input by feedforward input, the alignment between the feedforward network and recurrent network can determine the reliability of final response representations. This can help to explore the mechanisms that enable endogenously generated networks to form mature representations with the onset of sensory experience. However, the conceptual modeling before only cover the case of idealized symmetric neural interactions. Here we extend the previous modeling with more general network interaction structures and explore the performance of modeling under more complex circumstances. For asymmetric neural interactions, symmetrization can keep most of the pattern information and allow alignment in the real-number plane. White-noise evoked activity pattern is a possible candidate for alignment to approximate unknown recurrent interaction. Embedding Hebbian learning in the model shows the potential of plasticity for exploring the mechanisms. The work demonstrates the different modifications for different network structure conditions that enable the feedforward recurrent alignment. The theoretical explorations of alignment between feedforward and recurrent networks under different interaction conditions can help the prior modeling to become more general. Besides, the results can provide new perspectives to deepen the understanding of mechanisms for experience-driven developments in cortical networks.

Contents

1 Introduction

1.1 Introduction of Feedforward Recurrent Hypothesis

Cortical circuits embody remarkably reliable neural representations of sensory stimuli that are critical for perception and action [TWFK23]. Cortical circuits were thought to emerge from a developmental sequence that includes two distinct phases: an early period before the onset of experience during which endogenous mechanisms are thought to formulate the initial framework of cortical networks [AC14, FO10, Goo16, HFC08], and a subsequent period during which these early networks are refined under the influence of experience [AG18, Bar75, ES12, FI84, WF07].

The fundamental structure of cortical network representations is thought to arise early in development before the onset of sensory experience. However, how these endogenously generated networks respond to the onset of sensory experience, and the extent to which they reorganize with experience remains unclear [TWFK23].

In earlier work from the lab of M. Kaschube and D. Fitzpatrick [DA05], they focused on the problem of "nature-nurture transform". They tried to clarify the understanding of the capacity of the endogenous cortical network to reliably represent stimulus orientation at the onset of visual experience and the degree to which visual experience alters endogenous network structure to achieve mature stimulus representations.

To explore these early developmental dynamics, they applied data from the visual cortex of the newborn ferrets obtained through chronic *in vivo* calcium imaging [TWFK23]. Visually evoked activity in the visual cortex of postnatal ferrets before and following the onset of visual experience was employed. The visual cortex of higher mammals has served as a powerful model for exploring the contributions of these different phases to the development of mature cortical networks [TWFK23]. Before the onset of visual experience, activity-independent mechanisms combine with activity-dependent mechanisms driven by patterns of endogenous activity derived from the retina and the lateral geniculate nucleus (LGN) [FWS$^+$96, MWBS91, PRFS98] (Figure 1.1) to generate a robust modular network structure in visual cortex that is evident in patterns of spontaneous activity [CW01, SHW$^+$18]. This endogenously generated functional network is thought to form the initial framework for the emergent cortical representation of stimulus orientation since visual stimulation at or before eye-opening drives weakly orientation-selective responses at the cellular and modular scale [CWF20, CSB96, CGS98, GKBS97, SGS99] and spontaneous activity before eye-opening is predictive of the representations of stimulus orientations at eye opening [SHW$^+$18].

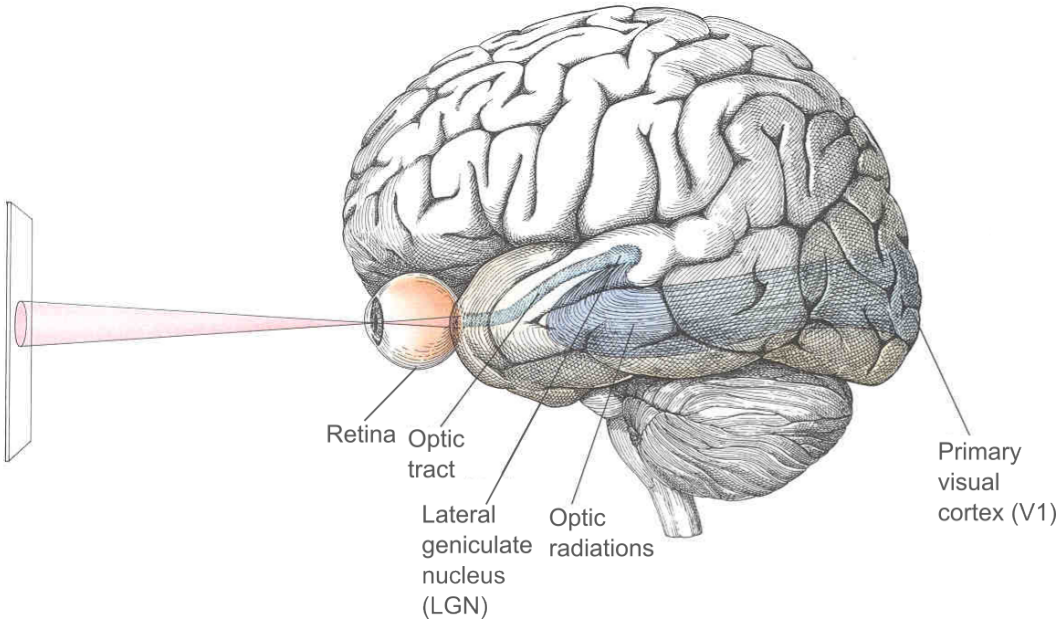


Figure 1.1 Central visual pathway in primates. The major pathway that visual information goes through from the eye to the primary visual cortex is shown. Signals are produced by receptors in the retina and are then transferred to a major relay station, the lateral geniculate nucleus (LGN) via the optic tract. Signals then travel through optic radiations to selected areas of the primary visual cortex (V1). From then on, signals are sent to higher areas of the cortex. [Oh04, Kar, Mon97]

The ferret is a species with a well-defined modular network of orientation-selective responses. Newborn ferrets open their eyes around thirty days after they are born. More than two days before eye-opening, the visual cortex was assumed to be visually naive. At least four days after eye-opening, the cortical network can gather information environment to become experienced (Figure 1.2). The data was collected at different time points from visually naive and to experienced visual cortex [TWFK23, Figure 1a]. In visually naive animals, their network responses are strong but highly variable, while in visually experienced animals, the diversity of responses was reduced and responses became more reliable.

To explore the underlying mechanism that builds reliable network responses, authors in the work from the lab of M. Kaschube and D. Fitzpatrick [TWFK23] developed the "feedforward recurrent alignment hypothesis". The hypothesis proposed that 1) the initial evoked activity pattern reflects novel visual input that is only poorly aligned with the endogenous networks and that 2) highly reliable visual representations emerge from a realignment of feedforward and recurrent networks that is optimal for amplifying these novel patterns of visually driven activity.

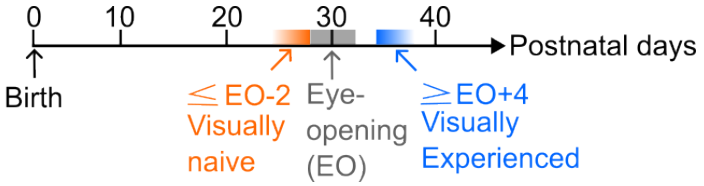


Figure 1.2 Timeline for ferret visual cortex development. Ferrets open their eyes around thirty days after birth. The visual cortex that is more than two days (inclusive of two days) before the eye-opening (EO) is considered to be visually naive. After four and more than four days from eye-opening, the visual cortex is expected to get enough training information from the environment and is therefore considered to be visually experienced [TWFK23].

1.2 Theoretical Exploration and Extensions

Ample computational work suggests that recurrent connections can give rise to amplification within subnetworks of coactive network units [Abb94, BYBOS95, DKM$^+$95, Mil16, CMVH17, PPV$^+$20]. Input that aligns more with such a subnetwork is expected to elicit a more robust response. Therefore, the "feedforward recurrent alignment hypothesis" was developed based on a conceptual computational network model of the early cortex and its response to visually evoked input using a minimal linear recurrent network [TWFK23]. In ~~the~~ this computational model, each unit represents the pooled activity in a local group of neurons. Connections between units describe the net interactions between local pools. For simplicity, it was assumed that the net interactions are symmetric, resulting symmetric interaction matrix for the recurrent network. The lab of M. Kaschube and D. Fitzpatrick [TWFK23] found out that the differences in the degree of feedforward recurrent alignment could reproduce the characteristics of network behavior from experimental observations that distinguish naive and experienced visual cortex-evoked responses (Method section 2.1 and Results section 3.1).

To explore the potential of the feedforward recurrent alignment hypothesis, we develop some theoretical explorations and extensions in this work on the existing feedforward recurrent alignment model. The explorations cover the perspectives of different recurrent network structures, experimental usability, and plasticity.

The first part of the exploration is to adapt the feedforward recurrent alignment modeling on asymmetric net interactions. The previous model considered for simplicity symmetric net interaction, which however loses the biological generality of cortical network structure. Asymmetric networks represent a more general cortical network structure but can ~~raise more complicated dynamics and~~ test result in patterns in complex planes. To solve this problem, we ~~try out~~ different modifications to adapt the prior feedforward recurrent alignment modeling. At the same time, the modified model should still reflect the experimental observations from data gathered in the

earlier work from the lab of M. Kaschube and D. Fitzpatrick [TWFK23]. The detailed method for this part is introduced in section 2.2 and the corresponding results in section 3.2.

The second part focuses on a different recurrent network structure suggested by [DVB+22, BDV+21, MO18]. The authors mentioned that large-scale neural recordings have established that the transformation of sensory stimuli into motor outputs relies on low-dimensional dynamics at the population level, while individual neurons exhibit complex selectivity. Prior experiments in behaving animals have found that trajectories of neural activity are typically restricted to low-dimensional manifolds in that space [MRB10, MSSN13, RBW+13, GG15, GPN+18, CSFW17, WNHJ18, SNMJ19]. Furthermore, they introduced the class of models, low-rank recurrent networks, directly embodies the idea of low-dimensional collective dynamics, opens the door to relating connectivity and dynamics, and provides a framework that unifies a number of specific RNN classes [MO18]. Low-rank RNNs rely on connectivity matrices that are restricted to be low-rank, which directly generate low-dimensional dynamics. Therefore, we are also interested in the adaptation of feedforward recurrent alignment on the promising low-rank recurrent network. Methods for the construction of low-rank recurrent networks are introduced in section 2.3 and its corresponding results in section 3.3.

The third part is related to the perspective of experimental usability of feedforward recurrent alignment. Since the whole cortical network structure is difficult to access during laboratory experiments, the degree of alignment between feedforward inputs and recurrent network structure is out of accessibility. Therefore, it could be helpful if the feedforward recurrent alignment model could be reformulated only demanding experimental measurable information. The work [MYDF09] pointed out that the reliability of evoked dynamics in recurrent networks is dependent on the stimulus used. [HM23] further predicted that stimulus inputs that aligned with the structure of endogenous subnetworks would be recurrently amplified, leading to more reliable evoked responses and constraining the potential outputs of the network. Based on an experimental method introduced by Mulholland et al. [HM], white-noise evoked activity can be used to generate spontaneous-like activity patterns. Inspired by the series of works, we explore the possibility of modifying the feedforward recurrent alignment with white-noise evoked activity. Since the original recurrent network structure is assumed to be unknown, we call it the black box recurrent network model. The method for this part is described in section 2.4 and the results in section 3.4.

Finally, the last part regards the feedforward recurrent alignment hypothesis from the perspective of network learning and plasticity. Connection strengths can be modified by learning from experience, and the degree of learning from each experience is a parameter that can be modified. The simplest kind of learning is Hebbian learning (Hebb, 1949), where the weight between a sending and a receiving node in-

creases if the two nodes are active at the same time. In other words "Nodes that fire together, wire together". This enables learning the correlational structure of the environment [RM21]. Thus, including the learning rule in the feedforward recurrent network could also be a potential perspective to explore the mechanism for the experience-driven change of response reliability. The methods for the exploration of feedforward recurrent alignment considering simple Hebbian rule are introduced in section 2.5 and the results for it in section 3.5.

The Figure 1.3 illustrates the summarized perspectives of this work from paragraphs above.

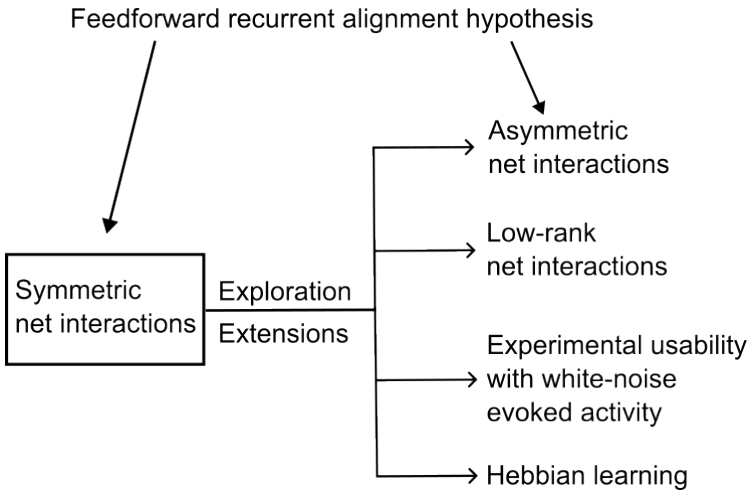


Figure 1.3 Included extensions and explorations in the work. Based on the prior work from the lab of M. Kaschube and D. Fitzpatrick [TWFK23], the feedforward recurrent alignment hypothesis was developed with a conceptual network model. For simplicity, the symmetric net interactions were considered. In this work, theoretical explorations and further extensions of the prior network model are carried out in multiple perspectives. It includes adapting the model to more biologically realistic net interaction structures, for example, asymmetric and low-rank. Moreover, due to the difficult accessibility of whole net interaction during experiments, white-noise evoked activity is chosen to modify the modeling. Last but not least, Hebbian learning as the basic learning rule is taken into account to explore the dynamic of feedforward recurrent alignment.

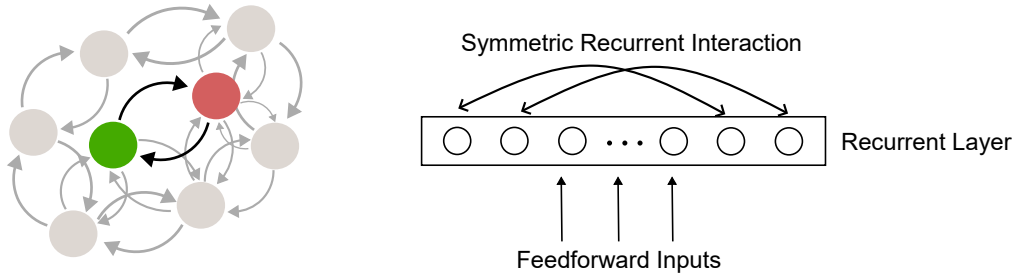
2 Methods

In this chapter, we will give an overview of the recurrent neural network (RNN) models for exploration in this work of the feedforward recurrent alignment hypothesis, which states that optimized alignment between feedforward and recurrent network can lead to reliable visual representations during development [TWFK23]. First, we introduce the symmetric network model as the basis for modifications and extensions in further models. Next, we modify these symmetric networks by applying other network structures and conditions. Finally, we consider the potential role learning could play in the feedforward recurrent alignment hypothesis.

2.1 Symmetric Recurrent Network Model

recurrent neural network (RNN).

We first consider the basic case of having a full-rank symmetric recurrent network. 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.



recurrent neural networks

Figure 2.1 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).

2.1.1 Symmetric Recurrent Interaction

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

$$J_{ij} \sim \mathcal{N}(0, 1). \tag{2.1}$$

Since the interaction matrix is symmetric, its transpose equals to itself,

$$J_{ij} = J_{ji} \text{ or } J^T = J. \tag{2.2}$$

As J is a full-rank matrix, its rank equals the number of neurons n involved in the RNN,

$$\operatorname{rank}(J) = n\,. \tag{2.3}$$

The number of eigenvectors therefore equals the number of neurons and all of their entries are real numbers. The eigenvalues λ_i are scaled such that the maximal eigenvalue $\lambda_{\max} = R$ with

$$\tilde{\lambda}_i = \frac{R\lambda_i}{\lambda_{\max}} \quad \forall i\,, \tag{2.4}$$

where the parameter $R < 1$ and $\{\lambda_i\}_{i=1,\dots,n}$ are the original eigenvalues of J and $\{\tilde{\lambda}_i\}_{i=1,\dots,n}$ the re-scaled eigenvalues. The re-scaling of eigenvalues guarantees the existence of a stable steady state for the network dynamics, which will be explained more in detail in the following section.

2.1.2 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 2.1 could be described by the ordinary differential equation [DA05]:

$$\tau_r \frac{\mathrm{d}r}{\mathrm{d}t} = -r + J \cdot r + h \overset{\tau_r=1}{\Rightarrow} \frac{\mathrm{d}r}{\mathrm{d}t} = -r + J \cdot r + h\,, \tag{2.5}$$

with the vector $r \in \mathbb{R}^{n\times 1}$ describing firing rate 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 eq.(2.5) can be received by setting the ordinary differential equation to zero. For simplicity, the time constant τ_r is set to 1. We then have the formulation for steady-state response r^*:

$$\frac{\mathrm{d}r}{\mathrm{d}t} = -r + J \cdot r + h = 0 \Rightarrow r = (I_n - J)^{-1} \cdot h =: r^*\,, \tag{2.6}$$

where $I_n \in \mathbb{R}^{n\times n}$ 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 eq.(2.5) could also be written in an element wise expression:

$$f_i(r_1, ..., r_n) := \frac{\mathrm{d}r_i}{\mathrm{d}t} = -r_i + \sum_{j=1}^{n} J_{ij} r_j + h_i \text{ for } i = 1, ..., n\,. \tag{2.7}$$

The partial differentiation of f_i to r_j is

$$\frac{\partial f_i}{\partial r_j} = \begin{cases} -1 + J_{ij} & \text{if } i = j \\ J_{ij} & \text{if } i \neq j \end{cases}\,. \tag{2.8}$$

The Jacobian matrix A of the dynamic system eq.(2.5) is then

$$A := \begin{pmatrix} \frac{\partial f_1}{\partial r_1} & \cdots & \frac{\partial f_1}{\partial r_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial r_1} & \cdots & \frac{\partial f_n}{\partial r_n} \end{pmatrix} = -I_n + J\,. \tag{2.9}$$

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, it follows the reformulation

$$(-I_n + J)E = -I_n \cdot E + J \cdot E = -I_n \cdot E + \Lambda \cdot E = (-I_n + \Lambda)E\,, \tag{2.10}$$

where Λ is 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 determine 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 are limited by the parameter $R < 1$, defined in eq.(2.4), we have

$$-1 + \lambda_i \overset{(2.4)}{<} -1 + 1 = 0\,. \quad \text{for all i = 1, ..., n} \tag{2.11}$$

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.

2.1.3 Feedforward Recurrent Alignment for Symmetric Interactions

Here, we introduce the definition of the feedforward recurrent alignment score between the feedforward input and recurrent network.

We model the neural synaptic firing rate with Gaussian distribution. The feedforward inputs are multivariate normal distributed vectors with a certain mean vector $h \in \mathbb{R}^{n\times 1}$ and a covariance matrix. In this work, if mentioning feedforward inputs without further definition, we refer to their mean vectors. If aligning the feedforward inputs to certain activity patterns described by firing rate vectors, the mean vector of inputs h is proportional to those vectors.

The alignment of a feedforward input $h \in \mathbb{R}^{n\times 1}$ with the recurrent network J is defined as [TWFK23]

$$\nu := \frac{h^T J h}{\| h \|^2}\,, \tag{2.12}$$

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

where we consider the Euclidean norm without loss of generality. h^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,

$$h \propto e_i \,, \tag{2.13}$$

the feedforward recurrent alignment ν is proportional to the eigenvalues λ_i, because inserting the proportionality eq.(2.13) in eq.(2.12) leads to

$$\nu = \frac{h^T J h}{\|h\|^2} \propto \frac{{e_i}^T J e_i}{\|e_i\|^2} = \frac{\lambda_i {e_i}^T e_i}{\|e_i\|^2} = \lambda_i \,. \tag{2.14}$$

Therefore, the alignment is maximal when the input is aligned to the eigenvector $e_{\max}$ with the maximal eigenvalue $\lambda_{\max}$ [TWFK23].

2.1.4 Response Properties for Evaluation

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

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

$$h \sim \mathcal{N}(\mu, \Sigma) \;\text{ with } \Sigma := \sigma_{\text{trial}} I_n. \tag{2.15}$$

Then, the steady state response $r^* = (I_n - J)^{-1} \cdot h$ from eq.(2.6) has the linearly transformed normal distribution

$$r^* \sim \mathcal{N}\left((I_n - J)^{-1}\mu, (I_n - J)^{-1}\Sigma(I_n - J)^{-T}\right) \,, \tag{2.16}$$

where the mean vector and covariance matrix are linearly transformed. The property could be proved analogously as in [Soc19] with the moment-generating function M_h of the multivariate normal distributed input h as following

$$M_h(t) = \mathbb{E}\left[\exp(t^T h)\right] = \exp\left[t^T\mu + \frac{1}{2}t^T\Sigma t\right] \,. \tag{2.17}$$

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

$$\begin{aligned} M_{r^*}(t) &= M_h\left((I_n - J)^{-T}t\right) \\ &= \exp\left[t^T\left((I_n - J)^{-1}\mu\right) + \frac{1}{2}t^T(I_n - J)^{-1}\Sigma(I_n - J)^{-T}t\right] \,, \end{aligned} \tag{2.18}$$

which indicates the linearly transformed distribution of r^* as in eq.(2.16).

As defined in [TWFK23], 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

$$\beta_s = \frac{2}{N(N-1)} \sum_{i=1,j=i+1}^{N} \mathrm{corr}(r_i^s, r_j^s)\,, \tag{2.19}$$

where r_i^s is the i-th response trial that evoked by stimulus s and $\mathrm{corr}(r_i^s, r_j^s)$ 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 after the visual experience. To reflect the variation of responses during the stimulation period, the quantity of "intra-trial stability" was defined [TWFK23].

To model the time-dependent input $h(t) \in \mathbb{R}^{n\times 1}$ with mean vector μ and its evoked steady-state responses $r(t) \in \mathbb{R}^{n\times 1}$, the following stochastic differential equations are formulated

$$\mathrm{d}h = \mu \mathrm{dt} + \sigma_{\mathrm{time}} \mathrm{d}W \tag{2.20a}$$

$$\mathrm{d}r = (-r + J \cdot \mu)\mathrm{d}t + \sigma_{\mathrm{time}} \mathrm{d}W\,, \tag{2.20b}$$

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

To approximate the evoked response $r(t)$, eq.(2.20b) is solved numerically with Euler-Maruyama scheme

$$r_{t+1} = r_t + (-r_t + J \cdot \mu)\Delta t + \sigma_{\mathrm{time}} \sqrt{\Delta t} \Delta \tilde{W}_t\,, \tag{2.21}$$

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 by the multivariate normal distribution with zero vector 0_v as mean vector and identity matrix I_n as covariance matrix,

$$\Delta \tilde{W}_t \sim \mathcal{N}(0_v, I_n)\,. \tag{2.22}$$

For 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 $t + \Delta\tilde{t}$

$$c(t, \Delta\tilde{t}) := \bar{r}(t)^T \bar{r}(t + \Delta\tilde{t})\,, \tag{2.23}$$

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

$$\bar{r}(t) := \frac{r - \langle r \rangle}{\sigma_r}\,, \tag{2.24}$$

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

$$\begin{aligned} \bar{c}(\Delta\tilde{t}) &:= \frac{1}{T - \Delta\tilde{t}} \int_0^{T-\Delta\tilde{t}} c(t, \Delta\tilde{t}) \mathrm{d}t \\ &= \frac{1}{T - \Delta\tilde{t}} \int_0^{T-\Delta\tilde{t}} \bar{r}(t)^T \bar{r}(t + \Delta t) \mathrm{d}t \,. \end{aligned} \tag{2.25}$$

defined as

tilde

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 [TWFK23].

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

$$h \sim \mathcal{N}(0_v, \Sigma^{\mathrm{Dim}}) \,, \tag{2.26}$$

the linear transformed responses (analogously as (2.16)) are

$$r \sim \mathcal{N}(0_v, (I_n - J)^{-1} \Sigma^{\mathrm{Dim}} (I_n - J)^{-T}) \tag{2.27}$$

with

$$\Sigma^{\mathrm{Dim}} := \sum_{i=L}^{L+M_{\mathrm{dim}}} \exp\left(\frac{-2(i-L)}{\beta_{\mathrm{dim}}}\right) e_i e_i^T \,, \tag{2.28}$$

in which the parameter M_{dim} determined the number of eigenvectors that are taken into account to construct the matrix. The parameter β_{dim} controls the dimensionality of the pattern, since it determines how flat the explained ratio curve will be [TWFK23]. The index L determines the first vector for the chosen set and starts from index 1 to the number of half of the neuron number $\frac{n}{2}$. Since the eigenvectors of J build a set of basis for $\mathbb{R}^n$, they could be chosen as basis vectors e_i for covariance matrix Σ^{Dim} in eq.(2.28). Hereby, the eigenvectors are ordered according to their eigenvalues in descending order,

$$e_{\max}, ..., e_i, e_j, ..., e_{\min} \text{ such that } \lambda_{\max} > ... > \lambda_i > \lambda_j > ... > \lambda_{\min} \,. \tag{2.29}$$

The exponential factor in eq.(2.28) simulates the exponential decay of variance ratio observed in data [TWFK23].

The linear effective dimensionality for quantifying the change of pattern diversity during visual maturation is defined based on participation ratio citation [TWFK23]

$$d_{\text{eff}} := \frac{(\sum_i \lambda_i)^2}{(\sum_i \lambda_i^2)}\,, \tag{2.30}$$

where λ_i are the eigenvalues of a covariance matrix Σ from an activity pattern distribution. Since as defined in eq.(2.28), 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

$$\lambda_i^{\text{Dim}} = \exp\left(\frac{-2(i-L)}{\beta_{\text{dim}}}\right)\lambda_i\,. \tag{2.31}$$

The covariance of the responses shares the same eigenvectors as Σ^{Dim} based on its distribution eq.(2.27) and therefore also the same as J. The eigenvalues λ_i^{Act} for the responses can be obtained through re-scaling eq.(2.31)

$$\lambda_i^{\text{Act}} = \exp\left(\frac{-2(i-L)}{\beta_{\text{dim}}}\right)\frac{1}{(1-\lambda_i)^2}\,, \tag{2.32}$$

for $i = L, ..., L + M_{\text{dim}}$. ~~Ans~~ L the index from 1 to $\frac{n}{2}$

Insert the eigenvalues of responses eq.(2.32) in the eq.(2.30) for effective dimensionality to get the final formulation of dimensionality for responses with eigenvalues of J,

$$d_{\text{eff, ana}} = \frac{\left(\sum_{i=L}^{L+M_{\text{dim}}} \exp\left(-2\frac{i-L}{\beta_{\text{dim}}}\right)(1-\lambda_i)^{-2}\right)^2}{\sum_{i=L}^{L+M_{\text{dim}}} \exp\left(-4\frac{i-L}{\beta_{\text{dim}}}\right)(1-\lambda_i)^{-4}}\,. \tag{2.33}$$

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.,

$$d_{\text{eff, emp}} = \frac{\left(\sum_{i=L}^{L+M_{\text{dim}}} \text{var}_i\right)^2}{\sum_{i=L}^{L+M_{\text{dim}}} \text{var}_i^2} \tag{2.34}$$

with var_i the i-th variance ratio. The index L determines the first vector for the chosen set and starts from index 1 to the number of half of the neuron number.

Alignment with spontaneous activity If projecting the evoked activity patterns into the space spanned by principal components of spontaneous activity patterns, the alignment of activity patterns to spontaneous activity reflects the size of overlaps between explained variance curves from evoked and spontaneous activity patterns. Let V be the evoked response pattern and S the spontaneous activity pattern, the projection of V to S could be quantified as the covariance matrix of V explained by the principal components of S, which results in a variance ratio vector with elements v_i calculated as

$$v_i = \frac{p_{i,S}^T \cdot \Sigma_V \cdot p_{i,S}}{\mathrm{Tr}(\Sigma_V)}\,, \tag{2.35}$$

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

$$\gamma_i = \frac{r_{i,V}^T \cdot \Sigma_S \cdot r_{i,V}}{\|r_{i,V}\|^2 \mathrm{Tr}(\Sigma_S)}\,, \tag{2.36}$$

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.

Since the spontaneous activity is proposed to already exist almost a week before eye-opening for ferrets, we assume that they already fit into the activity space generated by the recurrent network. The covariance matrix can therefore be constructed with eigenvectors of the recurrent network. Besides, spontaneous activity is assumed to be evoked by inputs from broad sources. As a result, a larger number of eigenvectors are applied for the covariance matrix than in eq.(2.28). Finally, the spontaneous activity is more variable than evoked activity and therefore has a higher dimensionality. Because the parameter β_{dim} in eq.(2.26) indicates the dimensionality, we can set for spontaneous activity higher dimensionality with $\beta_{\mathrm{spont}} > \beta_{\mathrm{dim}}$ to generate high dimensional inputs [TWFK23]. Therefore, we than have the broad inputs $h^{\mathrm{spont}} \in \mathbb{R}^{n\times 1}$ and spontaneous activity $r^{\mathrm{spont}} \in \mathbb{R}^{n\times 1}$ generated from it, which are both multivariate normal distributed vectors

$$h^{\mathrm{spont}} \sim \mathcal{N}(0_v, \Sigma^{\mathrm{spont}}) \tag{2.37}$$

and

$$r^{\mathrm{spont}} \sim \mathcal{N}\left(0_v, (I_n - J)^{-1}\Sigma^{\mathrm{spont}}(I_n - J)^{-T}\right)\,. \tag{2.38}$$

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

$$\Sigma^{\text{spont}} := \sum_{i=1}^{M_{\text{spont}}+1} \exp\left(\frac{-2(i-1)}{\beta_{\text{spont}}}\right) e_i e_i^T \,. \tag{2.39}$$

M_{spont} determines the number of eigenvectors that are taken into account for the construction of the covariance matrix for spontaneous activity, which is also larger than M_{dim}.

2.2 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 RNNs, the asymmetric RNNs ~~is~~ are more biologically realistic because they allow asymmetric connections between neurons.

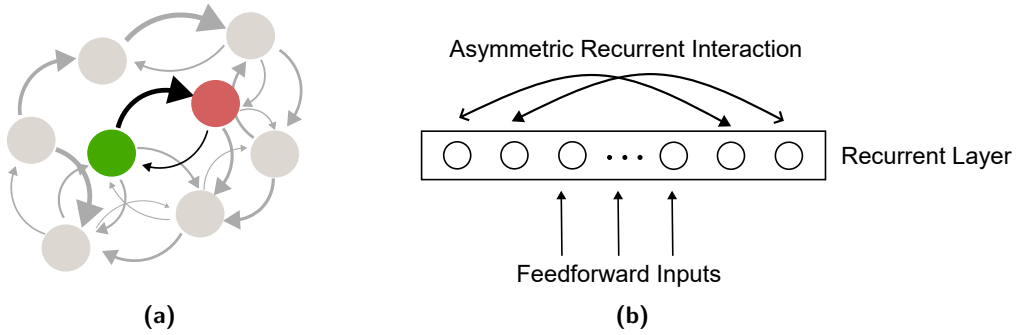


Figure 2.2 Illustration of asymmetric recurrent networks (asymmetric RNNs). In general, asymmetric recurrent networks do not have symmetric interaction strength between two neurons. **(a)** An example of asymmetric connections between two neurons in a network with multiple neurons. There are connections 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).

2.2.1 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

full-rank full-rank J_sym full-rank J_asym

$$J := aJ_{\text{sym}} + (1-a)J_{\text{asym}}\,. \tag{2.40}$$

The symmetric part is generated as described in section 2.1.1. 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.(2.5). Therefore, the steady state also keeps its form as eq.(2.6). Stability analysis of the steady state follows the same procedure as for symmetric interaction matrix (section 2.1.2). However, here ~~we have to keep in mind that~~ the

asymmetric interaction matrix J ~~now have~~ has left and right eigenvectors, which are not identical to each other any more. The Jacobian matrix $A = -I_n + J$ defined in eq.(2.9) still has the same set of eigenvectors as J. The eigenvalues are found in the similar way as eq.(2.10) but with left and right eigenvectors. If E_l and E_r the matrices containing the left and right eigenvectors column-wise, it follows

$$\begin{aligned} E_l^*(-I_n + J) &= -E_l^* + E_l^* J = -E_l^* + \Lambda E_l^* = (-I_n + \Lambda)E_l^* \\ (-I_n + J)E_r &= -E_r + JE_r = -E_r + E_r\Lambda = E_r(-I_n + \Lambda)\,, \end{aligned} \tag{2.41}$$

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

$$\{\lambda_i \in \mathbb{C} \mid \lambda_i \text{ eigenvalues of } J\}\,. \tag{2.42}$$

Therefore, the eigenvalues for the Jacobian matrix A ~~is~~ are now $-I_n + \Lambda \in \mathbb{C}^{n\times n}$ on the diagonal of matrix.

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: eigenvalues of

$$\exists i : -1 + \mathrm{Re}(\lambda_i) > 0 \Rightarrow \text{the steady state will be unstable.}$$

However, if

$$\forall i : -1 + \mathrm{Re}(\lambda_i) < 0 \Rightarrow \text{the steady state is stable.}$$

As a result, the responses converge to the steady state after enough long time, if

$$\forall i : -1 + \mathrm{Re}(\lambda_i) < 0 \;\Rightarrow\; \lim_{t\to\infty} r(t) = r^*\,. \tag{2.43}$$

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, original

$$\lambda_i = \frac{R\tilde{\lambda}_i}{\max\|\tilde{\lambda}\|} \quad \forall i\,. \tag{2.44}$$

with \tilde{\lambda} the original eigenvalues of J and \lambda the re-scaled eigenvalues. As a result, final eigenvalues of J are smaller than R.

Indent! 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 [RA06]. For a between 0 and 1, the distribution is a symmetric ellipse ~~2.3~~. (Figure 2.3)

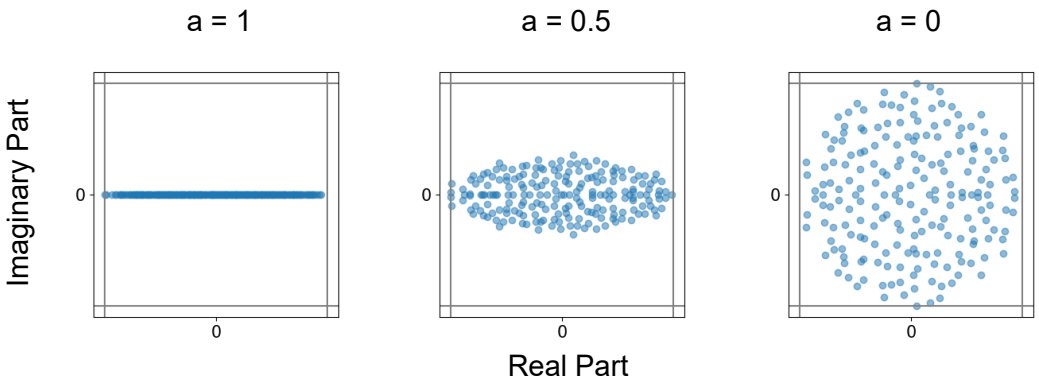


Figure 2.3 Eigenvalue distribution in dependence of parameter a in eq.(2.40). In general, an asymmetric matrix has eigenvalues in complex plane. The degree of symmetry determines the form of distribution from a line ($a = 1$, only real eigenvalues) to a full circle ($a = 0$) with radius $R < 1$. Between $a = 0$ 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 $R = 0.85 < 1$.

2.2.2 Modifications of Feedforward Recurrent Alignment for Asymmetric Interactions

The goal is to have the similar quantification of feedforward recurrent alignment as for the symmetric interaction matrix (section 2.1.3). However, when aligning the feedforward input to eigenvectors of asymmetric recurrent interaction network J like in eq.(2.14), the result calculated with eq.(2.12) 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 $h \in \mathbb{C}^{n\times 1}$ to the asymmetric interaction J, we consider:

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

$$\nu_{\mathrm{Re}} := \frac{\tilde{h}^T J \tilde{h}}{\|\tilde{h}\|^2} = \frac{\mathrm{Re}(h)^T J \mathrm{Re}(h)}{\|\mathrm{Re}(h)\|^2}\,. \tag{2.45}$$

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

$$\nu_{\mathrm{mag}} := \frac{\tilde{h}^T J \tilde{h}}{\|\tilde{h}\|^2} \text{ with } \tilde{h}_i = |h_i| \in \mathbb{R}\,. \tag{2.46}$$

3. Symmetrize J through

$$\tilde{J} = \frac{J + J^T}{2} \,. \tag{2.47}$$

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

$$\nu_{\text{sym}} = \frac{\tilde{h}^T J \tilde{h}}{\|\tilde{h}\|^2} = \frac{\tilde{e}_i^T J \tilde{e}_i}{\|\tilde{e}_i\|^2} \in \mathbb{R} \,. \tag{2.48}$$

2.2.3 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 2.2.2) are considered in the evaluation.

Except for the four response properties that are considered in the case of symmetric recurrent interaction matrix, for the modified forms 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 2.2.2 influence the modeling of properties that are involved in the evaluation.

citation [DA05]

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 $\tilde{h}$ from above section 2.2.2 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 2.2.2, 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 $\tilde{h}$,

$$h \sim \mathcal{N}(\tilde{h}, \sigma_{\text{trial}} I_n) \,. \tag{2.49}$$

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

$$r \sim \mathcal{N}\left((1-J)^{-1}\tilde{h}, \sigma_{\text{trial}}(1-J)^{-1}(1-J)^{-T}\right) . \tag{2.50}$$

Trial-to-trial correlation is determined by β_s defined in eq.(2.19).

Intra-trial stability Modulation of the single trial is described by the stochastic differential equations (2.20). 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

$$\mathrm{d}h = \tilde{h}\mathrm{d}t + \sigma_{\text{time}}\mathrm{d}W \tag{2.51a}$$

$$\mathrm{d}r = (-r + J \cdot \tilde{h})\mathrm{d}t + \sigma_{\text{time}}\mathrm{d}W\,, \tag{2.51b}$$

The solution of evoked activity is approximated by Euler-Maruyama scheme eq.(2.21). Intra-trial stability is then calculated by $\bar{c}(\Delta t)$ with eq.(2.25).

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 $\mathbb{R}^{n\times n}$. But with an asymmetric interaction matrix, the eigenvectors are the basis for $\mathbb{C}^{n\times n}$. 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.(2.30): 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 the covariance matrix.

As a result, having $e_i \in \mathbb{C}^{n\times 1}$ eigenvectors and eigenvalues $\lambda_i \in \mathbb{C}$ of asymmetric interaction network J, we transfer complex eigenvectors and eigenvalues to real vectors and values based on section 2.2.2:

- For modification 1 with eq.(2.45), applying $\tilde{e}_i := \mathrm{Re}(e_i)$ for construction covariance matrix and $\tilde{\lambda}_i := \mathrm{Re}(\lambda_i)$ for calculating dimensionality.
- For modification 2 with eq.(2.46), 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.(2.48), applying eigenvectors $\tilde{e}_i$ and eigenvalues $\tilde{\lambda}_i$ from symmetrized interaction matrix $\tilde{J}$ by eq.(2.47).

The covariance matrix for generating inputs is constructed ~~similar~~ similarily to it with symmetric interaction network defined by eq.(2.28) but with $\tilde{e}_i$ from above,

$$\Sigma^{\text{Dim}} := \sum_{i=L}^{L+M_{\text{dim}}} \exp\left(\frac{2(i-L)}{\beta_{\text{dim}}}\right) \tilde{e}_i \tilde{e}_i^T \,. \tag{2.52}$$

Analogously, calculating the effective dimensionality analytically defined by eq.(2.33) but with $\tilde{\lambda}_i$,

$$d_{\text{eff, ana}} = \frac{\left(\sum_{i=L}^{L+M_{\text{dim}}} \exp\left(-2\frac{i-L}{\beta_{\text{dim}}}\right)(1-\tilde{\lambda}_i)^{-2}\right)^2}{\sum_{i=L}^{L+M_{\text{dim}}} \exp\left(-4\frac{i-L}{\beta_{\text{dim}}}\right)(1-\tilde{\lambda}_i)^{-4}} \,. \tag{2.53}$$

The index L determines the first vector for the chosen set and starts from index 1 to the number of half of the neuron number.

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 2.1.4. The same modifications above for dimensionality in eq.(2.52) 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.(2.39). Applying the modifications above, insert therefore $\tilde{e}_i$ and receive

$$\Sigma^{\text{spont}} := \sum_{i=1}^{M_{\text{spont}}+1} \exp\left(\frac{2(i-1)}{\beta_{\text{spont}}}\right) \tilde{e}_i \tilde{e}_i^T \,. \tag{2.54}$$

2.3 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 fully-connected networks display only very stereotyped responses to external inputs and can implement only a limited range of input-output computations [MO18].

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 [MO18]. 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 [MO18]. Modeling with low-rank RNNs could help to understand the neural computations at the level of dynamical systems that govern low-dimensional trajectories of collective neural activity [BDV+21]. Therefore, we also explore the low-rank RNNs in extened feedforward recurrent alignment hypothesis.

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 [BDV+21].

2.3.1 Construction of Low Rank Interactions

Low-rank networks have a rank significantly 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 [BDV+21, DVB+22]. 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 [BDV+21].

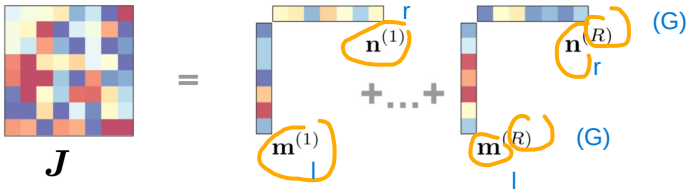


Figure 2.4 Low-rank recurrent networks (RNNs) constructed with distinct Gaussian distribution without random noise. A low-rank matrix could be written as a sum of outer products of vectors that are Gaussian distributed. As a result, the connectivity matrix is a mixture of Gaussians [BMS+23].

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

$$J_{ij} = \frac{1}{n} \sum_{g=1}^{G} l_i^{(g)} r_j^{(g)} \text{ or } J = \frac{1}{n} \sum_{g=1}^{G} l^{(g)} r^{(g)T} . \tag{2.55}$$

The connectivity is therefore characterized by a set of G n-dimensional vectors, denoted as connectivity patterns $l^{(g)} = \left\{ l_i^{(g)} \right\}_{i=1,...,n} \in \mathbb{R}^{n \times 1}$ and $r^{(g)} = \left\{ r_i^{(g)} \right\}_{i=1,...,n} \in \mathbb{R}^{n \times 1}$ 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 [BDV+21].

Figure

Low-rank RNNs with random noise [MO18]. Similar to the construction of low-rank connection matrix above (figure 2.4 defined by eq.(2.55)), [MO18] suggested that the connectivity matrix can be constructed with a part P to be fixed and known and a random uncorrelated part.

as

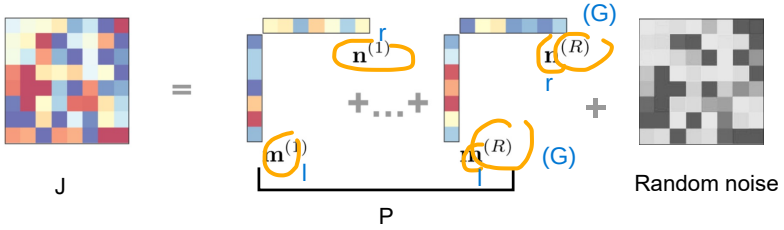


Figure 2.5 Low-rank recurrent networks (RNNs) constructed with fixed part and random noise. The connectivity matrix is given by the sum of a structured, controlled matrix P and an uncontrolled, random matrix. Except for the fixed and known part P, which is a mixture of uncorrelated Gaussians, the RNN is disturbed by random noise [BMS+23, MO18].

in general

The fixed part P has the same construction as the network without noise in eq.(2.55). The uncontrolled random matrix can be generally constructed in a complex way with certain current-to-rate transfer function to lend more complexity [MO18]. We denote the random noise part with J_{rand}. The low-rank RNN can be formulated as

$$J = \frac{1}{n} \sum_{g=1}^{G} l^{(g)} r^{(g)T} + J_{\text{rand}} , \tag{2.56}$$

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

full-rank

2.3.2 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. with adapted modifications from full-rank asymmetric RNNs for the evaluations..

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:

$$J = \frac{1}{n}\sum_{g=1}^{G} l^{(g)} l^{(g)T} \text{ or } J = \frac{1}{n}\sum_{g=1}^{G} r^{(g)} r^{(g)T} . \tag{2.57}$$

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.(2.57).

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.(2.5, 2.11). To keep the steady state of responses stable, the eigenvalues $\lambda_i \in \mathbb{R}$ of the symmetric low-rank RNNs J in eq.(2.57) are limited by $R < 1$ through normalizing with the maximal eigenvalue as done by eq.(2.4).

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.(2.55) 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 2.5 illustrated with definition eq.(2.56).

To enable the stable steady state of responses, as shown in asymmetric RNNs in eq.(2.43), the real part of eigenvalues $\text{Re}(\lambda_i)$ of an asymmetric RNN with interaction matrix J defined in eq.(2.56) are limited by $R < 1$ through normalization with the maximal magnitude of eigenvalues as done in eq.(2.44).

2.3.3 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.(2.19).
- Intra-trial stability $\bar{c}$ defined in eq.(2.25).
- Dimensionality in both analytical and empirical formulations $d_{\text{eff, ana}}$ and $d_{\text{eff, emp}}$ defined in eq.(2.33) and eq.(2.34).
- Alignment of evoked activity pattern to spontaneous activity pattern γ defined in eq.(2.36).

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 2.1.4. The feedforward recurrent alignment is defined by eq.(2.12). 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 2.2.2 and 2.2.3) 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.

modification
can consider in list
modification
modification

checkpoint 2

2.4 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. So, we can use their eigenvectors for the alignment. However, in reality during the experiments, the total interaction structure of the network is difficult to access. As a result, dominant directions for generating reliable neural activity with networks are generally unknown. The feedforward recurrent alignment cannot use eigenvectors of interaction matrices to characterize the development of feedforward recurrent network systems anymore. So, we consider possible ways of approximating the dominant eigenvectors to still enable the modeling of feedforward recurrent alignment under the "black box" condition.

dominant

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 [MYDF09]. Especially the stimulus inputs that align with the structure of endogenous sub-networks would be recurrently amplified, leading to more reliable evoked responses [HM23].

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 [MYDF09].

Therefore, we ~~wonder~~ test if we could apply spontaneous-like activity to characterize the feedforward recurrent alignment. Without knowing the eigenvectors of the recurrent interactions, ~~align~~ we 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, Mulholland et al. [HM] suggested a 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.

repeated amplification

2.4.1 Approximation with White Noise Evoked Activity

With an unknown recurrent structure, which is in general assume to be 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~~ reliable responses properties as before in section 2.1, 2.2 and 2.3. So, we have to find an alternative to approximate the original dominant eigenvectors for feedforward recurrent inputs to align.

The inputs are aligned to white-noise-evoked activity and ~~explore~~ we explore the response properties in correlation with feedforward recurrent alignment.

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

$$h_{\text{white}} \sim \mathcal{N}(0_v, I_n)\,. \tag{2.58}$$

The white-noise-evoked activity is then modeled by the transformed steady-state response $r \in \mathbb{R}^{n\times 1}$, which is also multivariate normal distributed with transformed covariance matrix (section 2.1.4),

$$r_{\text{white}} \sim \mathcal{N}\left(0_v, (1-J)^{-1}(1-J)^{-T}\right)\,. \tag{2.59}$$

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 2.2.2~~. Since the 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

$$\nu := \frac{p^T J p}{\|p\|^2}\,. \tag{2.60}$$

In the newly defined feedforward recurrent alignment eq.(2.60), the original recurrent network is now approximated by the spontaneous-like response patterns and the inputs are aligned to ~~the~~ their 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.(2.60) 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 [DA05]. 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 2.2.1 with formulation considering different grade of symmetry in eq.(2.40) for theoretical exploration. ~~Thus~~, the modification is similar to section 2.2.3. 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

$$h \sim \mathcal{N}(p, \sigma_{\text{trial}} I_n)\,. \tag{2.61}$$

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

$$r \sim \mathcal{N}\left((1-J)^{-1}p, \sigma_{\text{trial}}(1-J)^{-1}(1-J)^{-T}\right)\,. \tag{2.62}$$

The trial-to-trial correlation reflects the variation between ~~different~~ trials ~~from one stimulus~~ under the same alignment. It is the average of pairwise Pearson correlations between response trials as defined by β_s with eq.(2.19).

Intra-trial stability Intra-trial stability quantifies the variation inside one response trial evoked by input. ~~One time-dependent input and~~ The evoked steady-state response ~~trial is~~ approximated by Euler-Maruyama scheme described by (2.21). 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

$$\mathrm{d}h = p\mathrm{d}t + \sigma_{\text{time}}\mathrm{d}W \tag{2.63a}$$

$$\mathrm{d}r = (-r + J \cdot p)\mathrm{d}t + \sigma_{\text{time}}\mathrm{d}W\,. \tag{2.63b}$$

The intra-trial stability is the time average of delayed-response correlation defined by eq.(2.26).

Dimensionality For modeling the change in dimensionality against alignment score, the covariance matrix for input distribution is constructed as eq.(2.28) but with principal components p_i of white-noise-evoked activity as an approximation to the eigenvectors of the original recurrent network,

$$\Sigma^{\text{Dim}} := \sum_{i=L}^{L+M_{\text{dim}}} \exp\left(\frac{-2(i-L)}{\beta_{\text{dim}}}\right) p_i p_i^T\,. \tag{2.64}$$

The input and the evoked activity are modeled by multivariate normal distribution

$$h \sim \mathcal{N}(0_v, \Sigma^{\text{Dim}}) \tag{2.65a}$$

$$r \sim \mathcal{N}(0_v, (I_n - J)^{-1}\Sigma^{\text{Dim}}(I_n - J)^{-T})\,. \tag{2.65b}$$

The effective dimensionality is approximated by the eigenvalues of covariance matrix from white-noise-evoked activity pattern based on eq.(2.34).

checkpoint 3

Alignment between evoked activity and spontaneous activity To guarantee the spontaneous has a broader input than the evoked activity, the spontaneous activity for alignment to evoked activity is constructed similarly to eq.(2.64) with a higher dimensionality $\beta_{\text{spont}} > \beta_{\text{dim}}$. For the formulation of covariance matrix for spontaneous activity, the principal components from white-noise-evoked activity pattern is,

$$\Sigma^{\text{spont}} := \sum_{i=L}^{M_{\text{spont}}+1} \exp\left(\frac{-2(i-1)}{\beta_{\text{spont}}}\right) p_i p_i^T\,. \tag{2.66}$$

The spontaneous activity is then modeled by

$$r \sim \mathcal{N}(0_v, (I_n - J)^{-1}\Sigma^{\text{spont}}(I_n - J)^{-T})\,. \tag{2.67}$$

The amount of overlap between evoked activity pattern generated with eq.(2.64b) and the principal components of spontaneous activity from eq.(2.67) quantifies the alignment between them. The average alignment over all N evoked-response trials is the final alignment to spontaneous activity.

$$\gamma = \frac{1}{N}\left(\frac{r_i^T \cdot \Sigma^{\text{spont}} \cdot r_i}{\|r_i\|^2 \mathrm{Tr}(\Sigma^{\text{spont}})}\right) \tag{2.68}$$

2.4.2 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.(2.28) is applied,

$$\Sigma_{\text{Low}} := \sum_{i=0}^{M} \exp\left(\frac{-2(i-1)}{\beta_{\text{Low}}}\right) e_i e_i^T \,. \tag{2.69}$$

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 $0_v \in \mathbb{R}^{n\times 1}$ as mean vector and Σ_{Low} as covariance matrix,

$$h_{\text{Low}} \sim \mathcal{N}(0_v, \Sigma_{\text{Low}}) \,. \tag{2.70}$$

The response evoked by low dimensional input from eq.(2.70) is modeled by the linearly transformed multivariate normal distribution

$$r_0 \sim \mathcal{N}\left(0_v, (1-J)^{-1}\Sigma_{\text{Low}}(1-J)^{-T}\right) \,. \tag{2.71}$$

At this step, the feedforward recurrent alignment ν_0 can be calculated with low dimensional input h

$$\nu_0 = \frac{h^T J h}{\|h\|^2} \,. \tag{2.72}$$

Prior knowledge predicts stimulus inputs that align to spontaneous activity can be recurrently amplified and lead to more reliable responses [HM23]. 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,

$$r_1 \sim \mathcal{N}\left(0_v, \left((1-J)^{-1}\right)^2 \Sigma_{\text{Low}} \left((1-J)^{-T}\right)^2\right) \,. \tag{2.73}$$

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

$$\nu_1 = \frac{r_0^T J r_0}{\|r_0\|^2} \,. \tag{2.74}$$

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,

$$r_n \sim \mathcal{N}\left(0_v, \left((1-J)^{-1}\right)^{n+1} \Sigma_{\text{Low}} \left((1-J)^{-T}\right)^{n+1}\right) \,. \tag{2.75}$$

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

$$\nu_n = \frac{r_{n-1}^T J r_{n-1}}{\|r_{n-1}\|^2} \,. \tag{2.76}$$

2.5 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 between and inside those layers. 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 [DA05].

Until now, we only considered the recurrent layer under given feedforward inputs (Figure 2.1b and Figure 2.2b). 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.

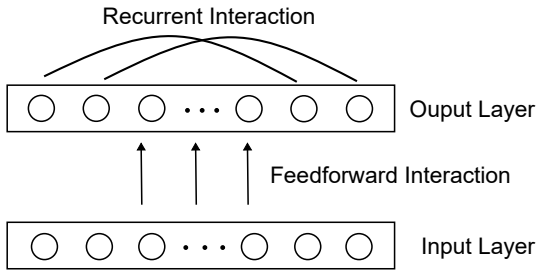


Figure 2.6 Illustration of a general feedforward recurrent network construction. Shown in the figure is a feedforward recurrent network with an input layer, an output layer, a feedforward synaptic weight matrix for feedforward interactions, and a recurrent synaptic weight matrix for recurrent interactions. Firing rate models are applied in both the input and output layer for modeling.

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 [DA05]. 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. Besides, neurons that fire together, should 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 [DA05].

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 [DA05]. Based on our network model (Figure 2.6), we consider the case in which there are multiple postsynaptic neurons.

2.5.1 Model Setting

To relieve the understanding of dynamics during the modeling, we start with the simplified assumption of a 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 $\mathbb{R}^{n\times 1}$ dimensional vector containing W_i the strength of connection between i-th output neuron and input neuron. Output rates $v \in \mathbb{R}^{n\times 1}$ describes the activity firing rate for output neurons. The random symmetric interaction $J \in \mathbb{R}^{n\times n}$ is constructed in the same way as for feedforward recurrent alignment hypothesis for symmetric RNNs in section 2.1.1.

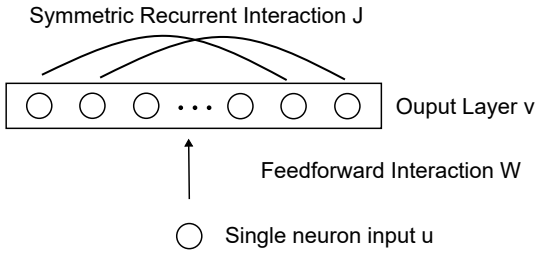


Figure 2.7 Illustration of feedforward recurrent network model with single input neuron. For start-up of feedforward recurrent network dynamic analysis, the simple case of only one input neuron with fixed random symmetric recurrent interaction J. The output layer has a number of n neurons. The feedforward interaction matrix is updated with the Hebbian rule.

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

$$\tau \frac{dv}{dt} = -v + W \cdot u + J \cdot v\,. \tag{2.77}$$

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

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 2.1.2 and 2.2.1)

$$v = W \cdot u + J \cdot v\,. \tag{2.78}$$

Solving the above equation (2.78), the steady state response of output layer for the feedforward recurrent network is

$$v^* = (I_n - J)^{-1} \cdot W \cdot u\,. \tag{2.79}$$

where I_n is the identity matrix.

2.5.2 Update Rules for Feedforward Network

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

$$\tau_w \frac{dW}{dt} = vu\,, \tag{2.80}$$

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~~ either suppressed or activated at the same time, the connection weight between those two neurons increases. Therefore, the eq.(2.80) 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:

$$\frac{dW}{dt} = \langle vu \rangle\,, \tag{2.81}$$

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

Replace the output rate v in eq.(2.81) with the steady state ~~formulation~~ eq.(2.79), the average plasticity rule can be rewrote as

$$\frac{dW}{dt} = \langle (I_n - J)^{-1} \cdot W \cdot uu \rangle = (I_n - J)^{-1} \cdot W \cdot \langle uu \rangle\,. \tag{2.82}$$

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

$$\frac{dW}{dt} = (I_n - J)^{-1} \cdot W \cdot Q\,. \tag{2.83}$$

Since for the simple case here that $u \in \mathbb{R}$, the input autocorrelation $Q = 1$ for all input patterns. Therefore, specifically for the simple network (Figure 2.7) ~~we consider~~, the average rule can be further simplified as

$$\frac{dW}{dt} = (I_n - J)^{-1} \cdot W\,. \tag{2.84}$$

2.5.3 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 $t+1$

$$W_{t+1} = W_t + \Delta t(I_n - J)^{-1} \cdot W_t \,, \tag{2.85}$$

where Δt is a enough small time distance during the total time period T_{Hebb}. The initial weight W_0 for iteration is ~~random~~ randomly 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.(2.85). 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~~ increase the reliability of evoked activity. Since $u \in \mathbb{R}$, 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 the space spanned by eigenvectors of the recurrent ~~network~~ interaction matrix.

Recurrent network has symmetric interaction J, thus the eigenvectors $\{e_i\}_{i=1,\dots,n}$ of J is a set of basis vectors that span the vector space $\mathbb{R}^n$. The feedforward weight vector $W \in \mathbb{R}^{n\times 1}$ 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:

$$e_{\max}, ..., e_i, e_j, ..., e_{\min} \text{ such that } \lambda_{\max} > ... > \lambda_i > \lambda_j > ... > \lambda_{\min} \,. \tag{2.86}$$

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

$$W_t = \sum_{i=1}^{n} \phi_i e_i = A\phi \,, \tag{2.87}$$

where $\phi \in \mathbb{R}^{n\times 1}$ 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.(2.87),

$$\phi = A^{-1} \cdot W_t \,. \tag{2.88}$$

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.

$$\rho := \frac{\sum_{i=1}^{20} |\phi_i|}{\sum_{i=1}^{n} |\phi_i|}\,, \tag{2.89}$$

where n is the total number of output neurons.

2.5.4 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.(2.85), the feedforward input for recurrent network h can be updated simultaneously through multiplying input rate u,

$$h_t = W_t \cdot u_t\,, \tag{2.90}$$

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~~

$$\nu_t = \frac{h_t^T J h_t}{\|h_t\|^2} = \frac{u_t W_t^T J W_t u_t}{u_t^2 \|W_t\|^2} = \frac{W_t^T J W_t}{\|W_t\|^2}\,. \tag{2.91}$$

Because $u_t \in \mathbb{R}$ 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 numerator and denominator of ν_t in eq.(2.91). Defining thereby the numerator as

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

$g(t)$ and denominator $h(t)$. Without loss of generality, the Euclidean norm is applied for denominator. That is

$$g(t) := W_t^T J W_t \tag{2.92a}$$
$$h(t) := \|W_t\|_2^2 \,. \tag{2.92b}$$

Applying product rule for derivative of $g(t)$, it results in

$$\frac{\mathrm{d}g(t)}{\mathrm{d}t} = \frac{\mathrm{d}W_t^T}{\mathrm{d}t} J W_t + W_t^T J \frac{\mathrm{d}W_t}{dt} \,. \tag{2.93}$$

Inserting the Hebbian rule eq.(2.84) for feedforward weights leads to

$$\frac{\mathrm{d}g(t)}{\mathrm{d}t} = W_t^T (I_n - J)^{-T} J W_t + W_t^T J (I_n - J)^{-1} W_t \,. \tag{2.94}$$

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

$$\frac{\mathrm{d}g(t)}{\mathrm{d}t} = W_t^T \left((I_n - J)^{-1} J + J (I_n - J)^{-1} \right) W_t \,. \tag{2.95}$$

Without the general non-commutativity of matric produc, the eq.(2.95) cannot be furthermore simplified.

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

$$\begin{aligned} \frac{\mathrm{d}h(t)}{\mathrm{d}t} &= \frac{\mathrm{d}\|W_t\|_2^2}{\mathrm{d}t} = \frac{\mathrm{d}\sum_{i=1}^n W_{t,i}^2}{\mathrm{d}t} = \sum_{i=1}^n \frac{\mathrm{d}W_{t,i}^2}{\mathrm{d}t} = 2 W_t^T \frac{\mathrm{d}W_t}{\mathrm{d}t} \\ &= 2 W_t^T (I_n - J)^{-1} W_t \,. \end{aligned} \tag{2.96}$$

The last equation is due to the Hebbian rule from eq.(2.84).

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

$$\begin{aligned} \frac{\mathrm{d}\nu_t}{\mathrm{d}t} &= \frac{\frac{\mathrm{d}g(t)}{\mathrm{d}t} h(t) - g(t) \frac{\mathrm{d}h(t)}{\mathrm{d}t}}{h(t)^2} \\ &= \frac{W_t^T \left((I_n - J)^{-1} J + J (I_n - J)^{-1} \right) W_t \|W_t\|_2^2 - 2 W_t^T J W_t W_t^T (I_n - J)^{-1} W_t}{\|W_t\|_2^4} \\ &= \frac{W_t^T}{\|W_t\|_2} (I_n - J)^{-1} J \frac{W_t}{\|W_t\|_2} + \frac{W_t^T}{\|W_t\|_2} J (I_n - J)^{-1} \frac{W_t}{\|W_t\|_2} \\ &\quad - 2 \frac{W_t^T}{\|W_t\|_2} J \frac{W_t}{\|W_t\|_2} \frac{W_t^T}{\|W_t\|_2} (I_n - J)^{-1} \frac{W_t}{\|W_t\|_2} \,. \end{aligned} \tag{2.97}$$

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

$$\frac{\mathrm{d}\nu_t}{\mathrm{d}t} = \bar{W}_t{}^T(I_n - J)^{-1}J\bar{W}_t + \bar{W}_t{}^T J(I_n - J)^{-1}\bar{W}_t - 2\bar{W}_t{}^T J\bar{W}_t\bar{W}_t{}^T(I_n - J)^{-1}\bar{W}_t\,. \quad (2.98)$$

3 Results

3.1 The Correlation between Response Properties from symmetrical Recurrent Interaction Networks and Feedforward Recurrent Alignment

The feedforward recurrent alignment, defined in the method under section 2.1.3 quantifies the degree of how much the feedforward input is aligned with the direction that is spanned by eigenvectors of the recurrent network. If the input is well aligned with the eigenvector corresponding to the maximal eigenvalue of the recurrent interaction J, the random noise in the response would be suppressed by the evoked response due to selective amplification. Response amplification determines the steady-state response,

$$r^* = \sum_{i=1}^{n} \frac{(e_i \cdot h)e_i}{1 - \lambda_i}, \tag{3.1}$$

where $\cdot$ denotes the dot product between two vectors.

The steady state response is then dominated by the projection of the input vector h along the axis defined by the eigenvector $e_{\max}$ whose eigenvalue $\lambda_{\max}$ is maximal and near one [DA05],

$$r^* \approx \frac{(e_{\max} \cdot h)e_{\max}}{1 - \lambda_{\max}}. \tag{3.2}$$

The projection of input h on eigenvector $e_{\max}$ reaches its maximal when h is approximately $e_{\max}$ itself. Therefore, the steady state responses reaches its maximum and the feedforward recurrent alignment equals the maximal eigenvalue $\lambda_{\max}$ of J because of eq.(2.14).

On the other hand, if the input is not well aligned with the dominant eigenvectors, the random noise is large relative to the response since the response is very weak. For an extreme example, if the input is aligned with the eigenvector $e_{\min}$ with minimal eigenvalue $\lambda_{\min}$, the response will almost not contribute to the steady state response at all due to the response amplification eq.(3.1). This could also be reflected by the feedforward recurrent alignment, which equals $\lambda_{\min}$ in this case because of eq.(2.14). As a result, the response is very weak respect to the constant noise and is therefore unreliable.

We can thus conclude that the feedforward recurrent alignment eq.(2.12) reflects the alignment between the feedforward input h and the dominant eigenvectors of the recurrent interaction network J. The better the input h is aligned to the dominant eigenvectors of J, the stronger and more reliable the response, and at the same time the higher the feedforward recurrent alignment score.

In the following sections, the four properties introduced at section 2.1.4 will be evaluated in the model and compare to the experimental tendency observed in [TWFK23].

3.1.1 Trial-to-Trial Correlation increases with larger Alignment

As defined under section 2.1.4 in paragraph "Trial-to-trial correlation", the inputs are constructed by multivariate normal distribution eq.(2.15). When aligning the input h with eigenvectors e_i, e_i are the the mean vector for input distribution as defined in eq.(2.15),

$$h \sim \mathcal{N}(e_i, \sigma^2_{\text{trial}} I_n)\,. \tag{3.3}$$

The feedforward recurrent alignment score is then determined by corresponding eigenvalue λ_i due to eq.(2.14), and reaches its maximal when align input h to $e_{\max}$.

We want to find out the correlation between the feedforward recurrent and the trial-to-trial correlation β_s defined by eq.(2.19). If sorting the eigenvectors in the order such that their corresponding eigenvalues are in ascending order,

$$e_{\min}, ..., e_i, e_j, ..., e_{\max} \text{ such that } \lambda_{\min} < ... < \lambda_i < \lambda_j < ... < \lambda_{\max}\,, \tag{3.4}$$

with λ_i the corresponding eigenvalue for eigenvector e_i. The inputs that aligned with eigenvectors in this order should have monotonously increasing feedforward recurrent alignments.

Generating the results with eq.(2.16) for N trials. The trial-to-trial correlation can be calculated with eq.(2.19).

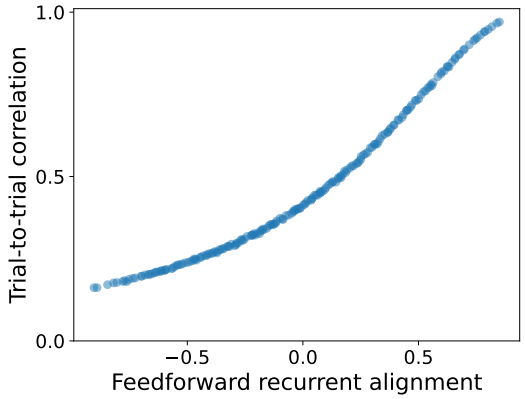


Figure 3.1 Correlation between feedforward recurrent alignment and trial-to-trial correlation for symmetric RNNs. Inputs aligned to eigenvectors e_i of interaction matrix J in the ascending order of eigenvalues eq.(3.4), resulting the feedforward recurrent alignment varies approximately between $\lambda_{\min}$ and $\lambda_{\max}$. For each input alignment to an eigenvector, $N = 100$ trials of evoked responses were generated for calculation of the trial-to-trial correlation calculated with eq.(2.19).

We assume feedforward recurrent alignment for visually naive cortex equals zero, which could be interpreted as responses evoked by random inputs. The trial-to-trial correlation with random inputs is smaller than it in the case, when the input is aligned to $e_{\max}$. This coincides with the experimental observations of responses from visually naive and experienced primary visual cortex of ferrets [TWFK23].

Moreover, the modeling result Figure 3.1 suggests a positive correlation between the feedforward recurrent alignment and the trial-to-trial correlation over the whole alignment range. The result confirms the idea that with the feedforward inputs

more and more aligned with the dominant eigenvectors of the recurrent network, the stability between trials increases also simultaneously. The process of reaching higher trial-to-trial stability is therefore a process of becoming more aligned with the dominant eigenvector. This positive correlation also coincides with the experimental results from ferrets [TWFK23].

feedforward input

3.1.2 Intra-Trial Stability increases with larger Alignment

Now we want to see if our modeling could capture the change in intra-trial stability during the development observed in the primary visual cortex of ferrets [TWFK23]. The intra-trial stability increased after the eye-opening and a couple of days. The feedforward recurrent alignment hypothesis suggests that the visually experienced cortex should have a better alignment between feedforward inputs and the dominant modes in the recurrent network. To confirm this idea, we would expect the intra-trial stability would be larger with a higher feedforward recurrent alignment score.

Analogous to trial-to-trial correlation, the eigenvectors can be sorted in descending order according to the eigenvalues eq.(3.4). The intra-trial stability is calculated with eq.(2.25).

ascending

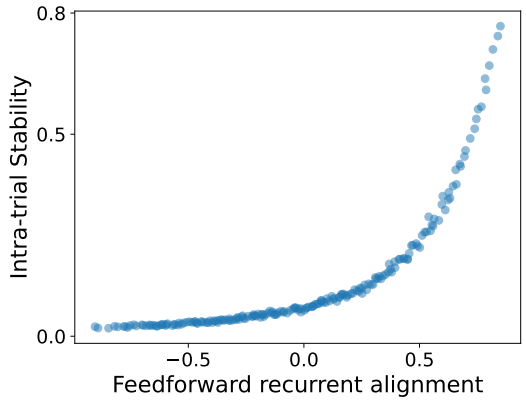


Figure 3.2 Correlation between feedforward recurrent alignment and intra-trial stability for symmetric RNNs. Inputs aligned to eigenvectors e_i of interaction matrix J sorted according to the ascending order of eigenvalues eq.(3.4), resulting the feedforward recurrent alignment varies approximately between $\lambda_{\min}$ and $\lambda_{\max}$. For one input aligned to an eigenvector, the intra-trial stability is calculated with the evoked steady-state response eq.(2.25).

The result (Figure 3.2) indicates a positive correlation between the feedforward recurrent alignment and the intra-trial stability. With random feedforward inputs, the feedforward recurrent alignment takes the value near zero. So the eye-opening happens somewhere between feedforward recurrent equals zero and reaches maximal its maximum. Thus, before the eye-opening, there is already a certain alignment that leads to a certain degree of intra-trial correlation.

Furthermore, the correlation is almost exponential. So, the enhancement of the input alignment to the dominant eigenvector is more rapid after the eye-opening than before. One assumption for this phenomenon could be that after the eye-opening, the environment provides more training data for the network so that the alignment

between inputs and the dominant eigenvector could be improved more efficiently. As a result, the responses get more intense and drive a better alignment forward. A positive loop could arise and speed up until the optimum is reached.

3.1.3 Dimensionality decreases with larger Alignment

Dimensionality is a generally important property of neural representations and could help to understand processes for example in learning and controlling [BSMA20, BBKK21]. A low-dimensional representation will encode a diverse range of inputs into a small set of common, orthogonal activity patterns. In other words, low-dimensional activity patterns require a small number of basis vectors from the response space to represent themselves. On the contrary, a high-dimensional representation will separate even similar inputs into orthogonal activity patterns. Compared to low-dimensional representation, a high-dimensional activity pattern is represented through a large set of basis vectors [BBKK21].

Therefore, we want to take a look at the change of dimensionality during the increase of alignment in the modeling. In ferrets' primary visual cortex, the dimensionality decreased from days before eye opening until days after eye-opening [TWFK23]. We would then expect that the model should also suggest a decrease in dimensionality with an increase in alignment between inputs and dominant eigenvectors of recurrent networks.

For a certain alignment, the principal component analysis reflects directly the dimensionality of the evoked activity pattern under this alignment. Because the principal components are the eigenvectors of response covariance, they also build up a set of basis vectors for the activity pattern space. Variance ratios reflect the weight that each principal component takes to represent the activity pattern. Thus, if only a small number of principal components contribute the most, the activity pattern is then low dimensional. If a broad set of principal components are similarly important, the activity pattern is highly dimensional.

With the idea of obtaining the linear dimensionality defined as participation ratio based on the principal component analysis, the eigenvectors here are ordered in descending order,

$$e_{\max}, ..., e_i, e_j, ..., e_{\min} \text{ such that } \lambda_{\max} > ... > \lambda_i > \lambda_j > ... > \lambda_{\min} . \tag{3.5}$$

For the generation of inputs with covariance matrix Σ^{Dim} defined in eq.(2.28), a subset of eigenvectors $\{e_i\}_{i=L,...,L+M_{\text{dim}}}$ will be chosen for $L = 1, ..., \frac{n}{2}$. M_{dim} then determines how many eigenvectors will contribute to generating inputs and evoked activity. In each such subset of eigenvectors, the leading eigenvector is e_L. Approximate here the feedforward recurrent alignment with the leading eigenvector only. Since L is considered only in the range of the first half of eigenvectors ordered as eq.(3.5), the range of feedforward recurrent alignment is between around 0 and $\lambda_{\max}$.

The linear dimensionality analytically and empirically will be calculated according to eq.(2.33) and eq.(2.34).

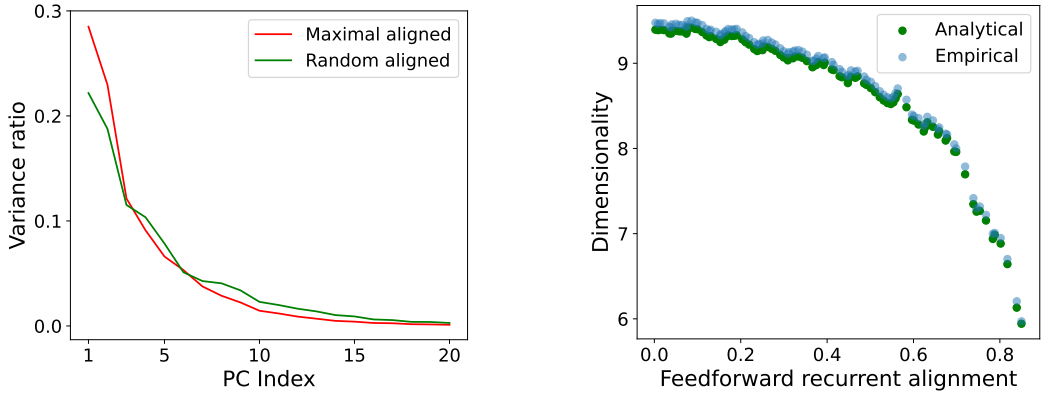


Figure 3.3 The correlation between dimensionality and feedforward recurrent alignment for symmetric RNNs. Prior experimental observations suggested that the dimensionality decreases from prior until post eye-opening [TWFK23]. With the feedforward recurrent alignment hypothesis, the dimensionality property of the neural representation during development could be captured. **(a)** Principal component analysis for the evoked activity under input aligned to $e_{\max}$ and spontaneous random activity. The red line is for maximal alignment and the green line is for spontaneous alignment. **(b)** Obtain the correlation between dimensionality and feedforward recurrent alignment with analytically eq.(2.33) and empirically eq.(2.34). The green line displays the analytical approximation for dimensionality and the blue dots for empirical.

As explained above, the principal component analysis helps to decode the dimensionality. The curve in Figure 3.3a of the variance ratio reflects the dimensionality. Spontaneous random alignment has a flatter variance ratio curve, indicating a broader range of eigenvector contributions. Therefore, the spontaneous random alignment has a higher dimensionality than under alignment with $e_{\max}$.

In total, a negative correlation between the dimensionality and feedforward recurrent alignment is shown in Figure 3.3b. The correlation forms nearly a flipped logarithmic function. The error between analytical and empirical results is small, which confirmed the good approximation formulated analytically by eq.(2.33).

Besides, the flipped logarithmic correlation suggests the a similar principle for intra-trial stability (Figure 3.3). After eye-opening, the reduction of dimensionality becomes larger when the feedforward inputs align better with the dominant eigenvector. It could be the case, that after the eye-opening, the recurrent network tries to encode the environment information with a large number of eigenvectors, which is costly for the system. After some time of getting used to the stimuli and the

evoked activity becomes more stable, the information is more determined, and fewer eigenvectors are needed to efficiently to encode it.

3.1.4 Alignment to Spontaneous Activity Increases with larger Alignment

Spontaneous activity in neural systems is defined as neural activity that is not driven by an external stimulus. The activity patterns of spontaneous activity are not completely random and have often unique spatiotemporal patterns that instruct neural circuit development in the developing brain. Moreover, normal and aberrant patterns of spontaneous activity underline behavioral states and diseased conditions in adult brains. Therefore, spontaneous activity is essential for the understanding of brain development [ILMR18]. The alignment between activity patterns and spontaneous activity patterns could show the structural relation between patterns.

In newborn ferrets' brains, visual responses that are loosely aligned with spontaneous activity in the cortex before eye-opening transformed to reliable and well-aligned responses several days after eye-opening [TWFK23]. Thus, we expect that the feedforward recurrent hypothesis can reflect the tendency that evoked activity aligned better with spontaneous activity during the development of the brain.

Under the assumption that at eye-opening, the patterns of feedforward inputs are aligned to random activity patterns, thus not as well aligned to the recurrent network as the spontaneous activity. The evoked and spontaneous pattern overlaps only a little (Figure 3.4(a), visually comparing the green line to the red line.). It could be observed here that only the last few principal components have a similar variance ratio, while the first few dominant eigenvectors differ a lot. There is not much overlap between green and red curves. On the contrary, experience-driven changes that optimize the feedforward-recurrent alignment to $e_{\max}$ results in a stronger overlap between distributions of evoked and spontaneous activity patterns (Figure 3.4(a), visually comparing red line to blue line.). Most of the principal components have a similar explained variance ratio. The overlap between the red and blue line is significantly large. In both cases, the theoretical modeling hypothesis matches experimental observations in baby ferrets' brains [TWFK23].

To visualize and quantify the overlaps between activity patterns from evoked and endogenous patterns, we considered the summarized alignment score eq.(2.36). An exponential correlation between alignment to spontaneous activity and feedforward recurrent alignment (Figure 3.4b) is suggested by the modeling. The strong growth of overlaps between evoked and endogenous activity starts only shortly before the optimal experience-driven alignment, indicating that the alignment could require a large amount of experience and training. The costly process of optimal alignment to spontaneous activity could on the other hand reflect the importance of the connection between evoked and endogenous activity patterns.

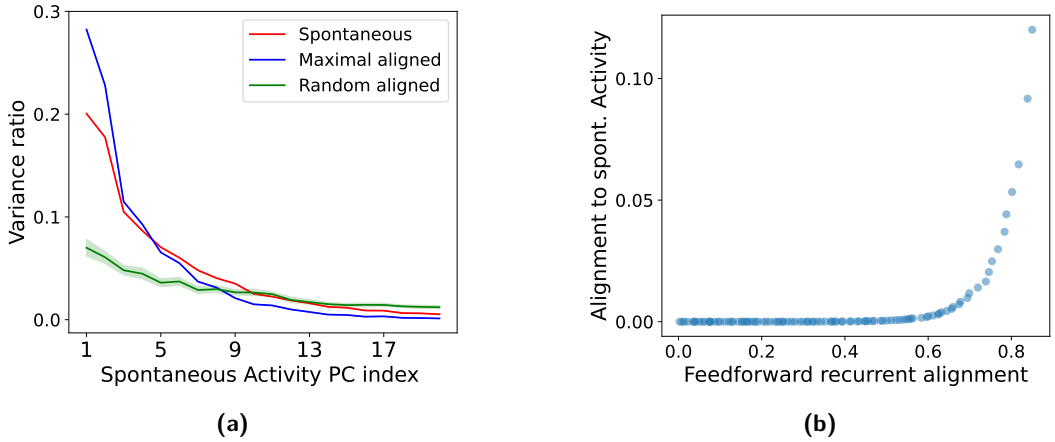


Figure 3.4 Correlation between alignment to spontaneous activity and feedforward recurrent alignment score in symmetric RNNs. The spontaneous activity reflects inputs from a wide range of different sources and is considered to be already aligned to the recurrent network[TWFK23]. Aligning activity patterns to spontaneous activity is in principle to explain the activity pattern by the principal components of spontaneous activity. **(a)** Variance ratio eq.(2.35) of spontaneous activity, evoked activity by feedforward input maximally aligned to recurrent network, and evoked activity by randomly aligned to recurrent network explained by principal components of spontaneous activity. The red line illustrates the variance ratio of spontaneous activity, the blue line the maximal alignment, and the green line the random alignment. Shadow shows the 95% confidence interval for 50 symmetric RNNs. **(b)** The correlation between final alignment to spontaneous activity and feedforward recurrent alignment eq.(2.36). The eigenvalues are ordered in descending order as eq.(3.5). Only the first half of eigenvectors are taken into account to determine the correlation.

3.2 Evaluation of Feedforward Recurrent Alignment Modulations for asymmetric Recurrent Interaction Networks

For symmetric interaction networks, the feedforward recurrent alignment hypothesis and the modeling based on it in section 3.1 can demonstrate the response properties observed in ferrets [TWFK23]. With better alignment between inputs and recurrent network, key results of the response properties were

- The trial-to-trial correlation increases.
- The intra-trial stability increases.
- The Dimensionality decreases.
- The alignment of evoked activity to spontaneous activity increases.

However, since the symmetric interaction matrices simplify the neural connection dramatically, we try to embed the more biology-realistic asymmetric interaction matrices. Considering the modifications listed in methods in section 2.2.2, we want to evaluate the modifications of the feedforward recurrent alignment score based on the key results we got with symmetric RNNs.

Firstly, we check if the feedforward recurrent alignment score keeps the proportionality to eigenvalues. We expect that a suitable modification could keep the monotonously positive correlation with eigenvalues. Then, we go through the four response properties to verify if the tendency above is still kept with increased alignment between inputs and recurrent network.

3.2.1 Monotony of Feedforward Recurrent Alignment Score in dependence of Eigenvalues

During the development, suggested by the feedforward recurrent alignment hypothesis for symmetric interactions, the inputs align better to the RNNs through aligning to dominant eigenvectors. Feedforward recurrent alignment is proportional to the corresponding eigenvalue of the eigenvector that is aligned with input eq.(2.14).

When considering the asymmetric interaction network, keeping the hypothesis that the inputs align more and more to the eigenvector with maximal eigenvalue, we examine if the modified feedforward recurrent alignment score could still keep the proportionality to eigenvalues. If a monotonously positive correlation between the feedforward recurrent alignment and eigenvalues could be kept, the alignment score could at least definitely quantify how well inputs are aligned to the dominant eigenvector.

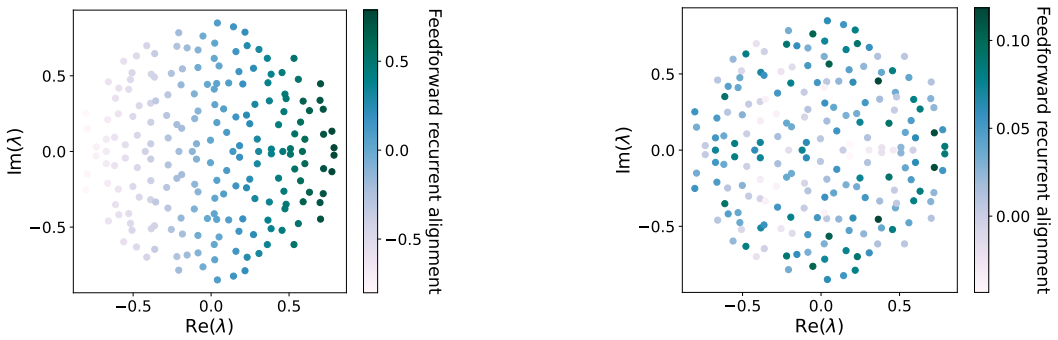


Figure 3.5 Correlation between eigenvalues and feedforward recurrent alignment of modifications for full-rank asymmetric RNNs. To verify the considered modifications from section 2.2.2, a monotonously positive correlation between eigenvalues and feedforward recurrent alignment is necessary. Represent the correlation in the complex plane, since eigenvalues are complex. The color-bar indicates the score of feedforward recurrent alignment. The darker the color, the larger the alignment score. **(a)** Representation of alignment score calculated with modification 1 by eq.(2.45) in complex plane. **(b)** Representation of alignment score calculated with modification 2 by eq.(2.46) in complex plane.

Modification 1 : Apply the real part of complex inputs with eq.(2.45).

With the feedforward recurrent alignment, we mainly consider the correlation between alignment score and corresponding eigenvalues. Inserting inputs h aligned to eigenvectors, for simplicity $h := e_i \in \mathbb{C}^{n\times 1}$ into eq.(2.45), it leads to

$$\nu_{\mathrm{Re}} = \frac{\mathrm{Re}(e_i)^T J \mathrm{Re}(e_i)}{\|\mathrm{Re}(e_i)\|^2}\,. \tag{3.6}$$

Because $\|e_i\| = c\|\mathrm{Re}(e_i)\|$ with $c \in \mathbb{R}_+$ a general positive constant, we could get the proportionality between alignment score ν_{Re} and the real part of eigenvalues $\mathrm{Re}(\lambda_i)$ of J :

$$\begin{aligned}\nu_{\mathrm{Re}} &= c\mathrm{Re}\left(\frac{e_i}{\|e_i\|}\right)^T J\,\mathrm{Re}\left(\frac{e_i}{\|e_i\|}\right)\\ &= c\mathrm{Re}\left(\frac{e_i^T}{\|e_i\|} J \frac{e_i}{\|e_i\|}\right)\\ &= c\mathrm{Re}(\lambda_i)\,,\end{aligned} \tag{3.7}$$

with c a general positive constant.

The positive correlation between alignment and the real part of eigenvalues λ_i could also be observed in representation in the complex plane (Figure 3.5a). With

increasing the real part of the eigenvalue, the size of the alignment score also becomes larger. A better alignment between inputs and a certain dominant activity pattern is thus the only reason for the increase in the feedforward recurrent alignment score. Since the alignment score only depends on the real part, there is no correlation found in the direction of the imaginary part of eigenvalues.

Modification 2 : Apply the magnitude for each neuron input eq.(2.46).

When the inputs are aligned to eigenvectors e of interaction matrix J, the feedforward recurrent alignment could be modified with $|e| := (|e_i|)_{i=1,\dots,n} \in \mathbb{R}^{n\times 1}$. Since the euclidean norm of vector $|e|$ is the same as the norm directly on the complex eigenvector $\|e\|_2$, the final feedforward recurrent alignment score from eq.(2.46) can be formulated as following:

$$\begin{aligned} \nu_{\text{mag}} &= \frac{|e|^T J |e|}{\|e\|^2} \\ &= \frac{\sum_{j=1}^{n} |e_j| \left(\sum_{i=1}^{n} |e_i| J_{ij}\right)}{\|e\|^2}, \end{aligned} \tag{3.8}$$

where e_i, e_j are the i-th and j-th element of the vector e, and J_{ij} the matrix element at i-th row and j-th column.

No direct proportionality between the alignment score and corresponding eigenvalues λ can be established. The lack of correlation is also represented in the complex plane (Figure 3.5b).

Modification 3 : Align the inputs to eigenvectors of symmetrized network eq.(2.48).

Instead of aligning the inputs to eigenvectors of original asymmetric RNNs and thinking about modifications of complex eigenvectors to calculate the feedforward recurrent alignment score, we now align the inputs to real eigenvectors $\tilde{e}$ from symmetrized interaction matrix $\tilde{J}$ as an approximation. However, the alignment score is still being projected to the original asymmetric RNNs defined in eq.(2.48). Despite of original asymmetric interaction matrix for the feedforward recurrent alignment score, there is still a proportionality between the eigenvalues of the symmetrized network and alignment score, shown in Figure 3.6.

The kept proportionality can also be obtained analytically. With the formulation of symmetrized interaction matrix through eq.(2.47), it follows with the help of definition for ν_{sym} from eq.(2.48),

$$
\begin{aligned}
&\frac{\tilde{e}^T \tilde{J} \tilde{e}}{\|\tilde{e}\|^2} = \frac{\tilde{e}^T \frac{J+J^T}{2} \tilde{e}}{\|\tilde{e}\|^2} = \frac{1}{2}\left(\frac{\tilde{e}^T J \tilde{e}}{\|\tilde{e}\|^2} + \frac{\tilde{e}^T J^T \tilde{e}}{\|\tilde{e}\|^2}\right) \\
\Rightarrow & 2\frac{\tilde{e}^T \tilde{J} \tilde{e}}{\|\tilde{e}\|^2} - \frac{\tilde{e}^T J^T \tilde{e}}{\|\tilde{e}\|^2} = \frac{\tilde{e}^T J \tilde{e}}{\|\tilde{e}\|^2} \\
\Rightarrow & 2\tilde{\lambda} - c = \frac{\tilde{e}^T J \tilde{e}}{\|\tilde{e}\|^2} \text{ with } c := \frac{\tilde{e}^T J^T \tilde{e}}{\|\tilde{e}\|^2} \in \mathbb{R} \text{ and } \tilde{\lambda} \text{ the corresponding eigenvalue of } \tilde{e} \\
\Rightarrow & \nu_{\text{sym}} = \frac{\tilde{e}^T J \tilde{e}}{\|\tilde{e}\|^2} \propto \tilde{\lambda}\,.
\end{aligned}
\tag{3.9}
$$

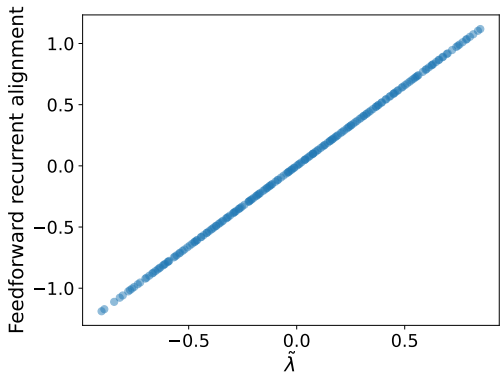


Figure 3.6 Positive correlation between feedforward recurrent alignment score and eigenvalues of symmetrized network as modification for asymmetric RNNs. Align the inputs to the eigenvectors of the symmetrized network while keeping feedforward recurrent alignment obtained by the original asymmetric interaction matrix with eq.(2.48). The correlation between the alignment score (y-axis) to the corresponding eigenvalues of the symmetrized network (x-axis) remains positive.

Section conclusion After the analysis above for all modifications, our expectation of a contained positive correlation is kept when considering the real part (modification 1) and alignment with the symmetrized interaction matrix (modification 3). The variant with magnitude (modification 2) fails to fulfill the expected correlation. Thus, we will leave modification 2 out without further analysis of response properties in the following section.

3.2.2 Verifying Response Properties with modified Feedforward Recurrent Alignment

After verifying the expected correlation between the feedforward recurrent alignment score and eigenvalues, we next test if the remaining two modifications could still reflect the four experimental observations listed at the beginning of the section 3.2.

For this, we align the inputs to eigenvectors of the asymmetric recurrent network to compute the four response properties and compare them with experimental ob-

servations from [TWFK23]. The following modifications for the modeling are taken into account:

- modification 1: consider only the real part of eigenvectors eq.(2.45).
- modification 3: align the inputs to eigenvectors of the symmetrized interaction matrix but calculates the alignment score still with the original asymmetric network eq.(2.48).

For the modeling, the asymmetric RNNs are constructed with a certain degree of symmetry with eq.(2.40). We therefore also controlled how much the degree of symmetry influences the results.

Trial-to-trial Correlation Trial-to-trial correlation is the averaged correlations between single trial responses defined by eq.(2.19). From the experimental results in [TWFK23], a positive correlation between feedforward recurrent alignment and trial-to-trial correlation is expected. Under various degree of symmetry, the positive correlation should still be kept.

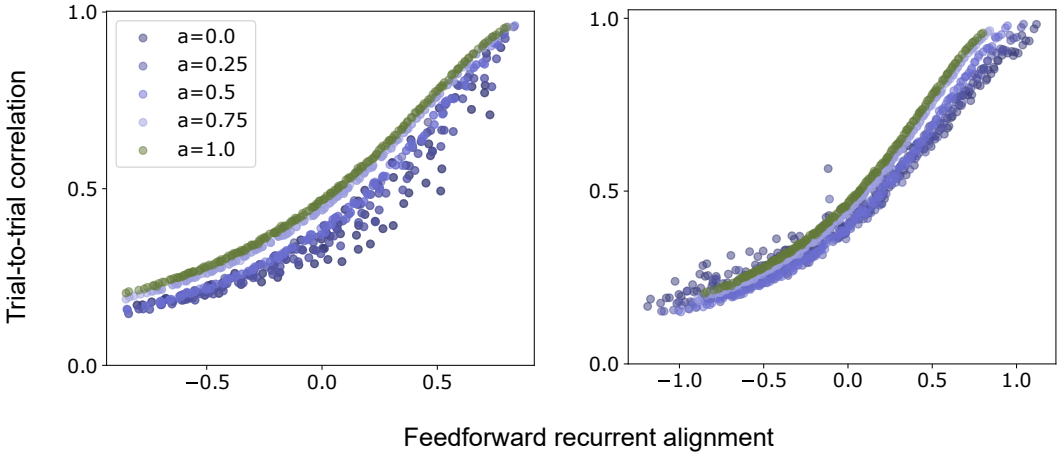


Figure 3.7 Trial-to-trial correlation concerning feedforward alignment and degree of symmetry in full-rank asymmetric RNNs. Color-coded is the degree of symmetry a (see legend). The green dots with $a = 1.0$ illustrate the case of symmetric RNNs. The darkest purple dots $a = 0.0$ represent the asymmetric network without any degree of symmetry. The x-axis shows the alignment score calculated with considered modifications 1 and 3, and the y-axis indicates the trial-to-trial correlation calculated with aligned inputs through eq.(2.19) **Left**: Results generated with modification 1 (only consider real part). **Right**: Result with modification 3 (symmetrized interaction matrix).

The results in Figure 3.7 show that the trend of positive correlation between trial-to-trial correlation and feedforward recurrent alignment score keeps while the network increases its asymmetry for both modifications 1 and 3.

However, when the network is fully asymmetric, the correlation has a larger dispersion with modification 1. We suspect that the reason is that in the case of full asymmetry, more information got lost if only considering information from the real part of the aligned inputs.

If there is a certain degree of symmetry, a part of the information originates from the symmetric structure. For this part, no information gets lost when only taking the real part of aligned inputs. On the other hand, aligning to fully asymmetric RNNs ($a = 0.0$ in Figure 3.8 left panel) and only taking the real part of aligned inputs, all neurons receive then incomplete information. Therefore, the correlation between evoked patterns will be disturbed mostly.

In the case of modification 3, the correlation is almost maintained the overall degree of symmetry as expected, shown in Figure 3.8 right panel.

Intra-trial Stability Intra-trial stability is the averaged time-delayed activity correlation inside one single trial quantified by eq.(2.25). As the name indicates, it shows how stable the information is represented inside one trial. According to the result from symmetric RNNs (Figure 3.2), we expect a similar exponential positive correlation also with full-rank asymmetric RNNs. The degree of symmetry should also not influence the result significantly under the assumption that the hypothesis works well with asymmetric RNNs generally.

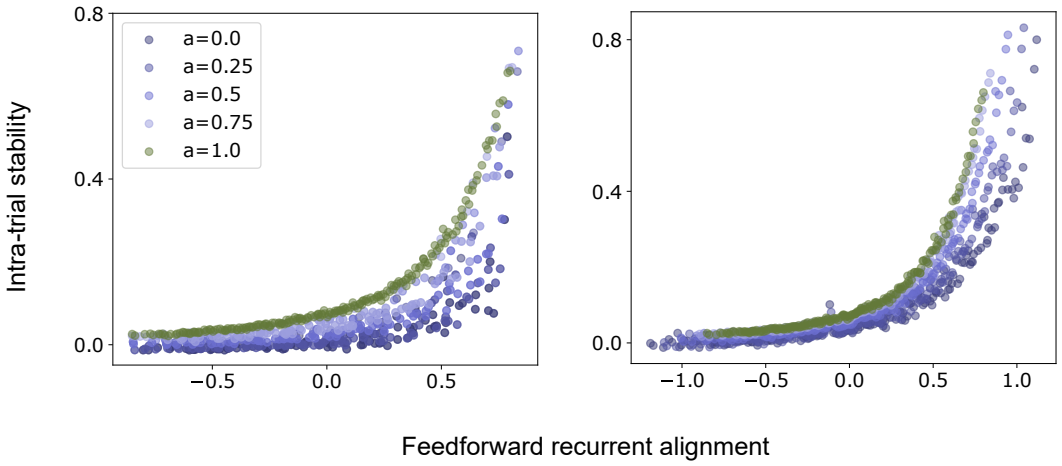


Figure 3.8 Intra-trial stability with respect to feedforward recurrent alignment and the influence from the degree of symmetry in full-rank asymmetric RNNs. For multiple degrees of symmetry a, different color dots are applied shown in the legend. From complete symmetric ($a = 1.0$ as the control group) to full asymmetric ($a = 0.0$) RNNs, the corresponding feedforward recurrent alignment (x-axis) is plotted against the intra-trial stability (y-axis). **Left**: Results with modification 1 (only real part of aligned inputs). **Right**: Results with modification 3 (align input to symmetrized network).

In the left panel of Figure 3.8, the dispersion increased when the network become more symmetric. We suspect with a similar reason as in trial-to-trial correlation that for modification 1, which is the loss of imaginary part information of aligned inputs.

For modification 3, the expected positive correlation between intra-trial stability and feedforward recurrent alignment score is kept. However, we also observe that the influence of the degree of symmetry influences here is more pronounced than in trial-to-trial correlation (Figure 3.7 right panel). It is perhaps because the intra-trial stability is more sensitive to the information lost than trial-to-trial correlation. With the increased asymmetry, more information got lost during symmetrization. Imagine if having a total symmetric network, the result of symmetrization is the network itself, and therefore no information is lost during the symmetrization. However, if having a fully asymmetric RNN, after symmetrization, complex eigenvectors are transformed into real-number eigenvectors. This could lead to the loss of information in those eigenvectors that characterize the asymmetric RNN. Thus, when aligning inputs to the symmetrized interactions, transformed information could be the reason for the influence of the degree of symmetry.

Dimensionality Dimensionality reflects the complexity of the information encoded by activity patterns. A high dimensional activity pattern needs more orthogonal activity patterns for representation than a low dimensional activity pattern and thus indicates larger variability and higher complexity of contained information [TWFK23, BSMA20, BBKK21]. We continuing for modeling the dimensionality with modifications 1 (consider real part) and 3 (symmetrization of asymmetric RNNs), which change the eigenvectors for the construction of input covariance and the eigenvalues for analytical calculation of effective dimensionality (section 2.2.3 eq.(2.52), (2.53)). If both modifications work well, the results should be similar to results from symmetric RNNs (Figure 3.2bb).

Different from the results in Figure 3.2bb, with modification 1 (Figure 3.9 left panel), the empirical dimensionality (eq.(2.34)) differs a lot from the analytical calculation (eq.(2.33)). Moreover, there is no significant correlation between dimensionality and feedforward recurrent alignment in the empirical modeling as long as asymmetric structure is included. There is a substantial mismatch between the analytical results and the empirical approximations. The reason for those phenomena could be the loss of orthogonality between the real part of eigenvectors, that is

$$e_i \perp e_j \;\not\Rightarrow \mathrm{Re}(e_i) \perp \mathrm{Re}(e_j)\ \forall i, j = 1, ..., n\,. \tag{3.10}$$

As a result, the covariance matrix Σ^{Dim} from eq.(2.52) is not necessarily constructed with orthogonal vectors anymore, which contradicts the definition of covariance matrix.

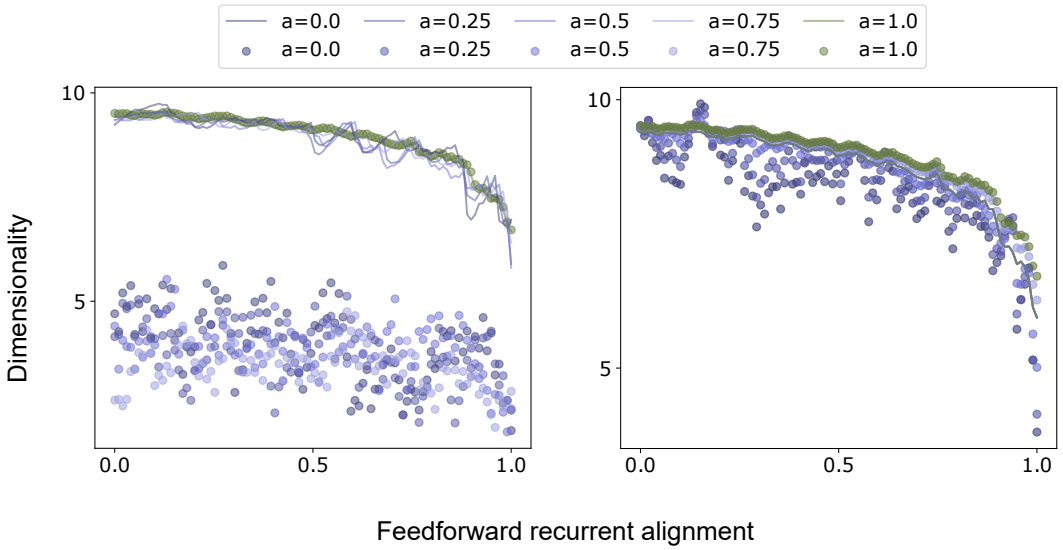


Figure 3.9 Analytical and empirical effective dimensionality with respect to the feedforward recurrent alignment score and the degree of symmetry in asymmetric RNNs. The full symmetric recurrent network ($a = 1.0$) is a control for other cases. When $a = 0.0$, the RNN is fully asymmetric. The dots represent the empirical approximation for effective dimensionality eq.(2.34). The lines are for the analytical calculation for dimensionality adjusted to modification 1 and 3 (section 2.2.3 eq.(2.52), (2.53)). **Left**: Results with modification 1 (only real part of aligned inputs and eigenvalues for analytical dimensionality). **Right**: Results with modification 3 (align inputs to symmetrized interaction matrix).

With modification 3, the results fulfill largely our expectations: the negative correlation between dimensionality and feedforward recurrent alignment is kept in both analytical and empirical approximations. Little dispersion could be due to the structural information lost during symmetrization.

Thus, until this step, we would also drop modification 1 and only further consider modification 3.

Alignment to spontaneous activity The alignment of the evoked activity pattern to the endogenous pattern measures the overlap between their variance ratio curves explained by principal components of endogenous pattern eq.(2.35). For example the overlaps between curves in Figure 3.4a. Only modification 3 is now left for the evaluation.

Similar to prior cases, if the modification works, we expect the correlation between alignment to spontaneous activity and feedforward recurrent alignment score to be similar to symmetric RNNs in Figure 3.4b. The alignment to spontaneous activity is modified with eq.(2.54) and calculated with eq.(2.36).

As shown in Figure 3.10, the dispersion increases with the degree of asymmetry due to the increased information lost. But the general tendency that the alignment of evoked activity pattern to spontaneous activity becomes larger when the inputs are aligned to more dominant eigenvectors of symmetrized interaction matrix is conserved. With a high degree of symmetry, for example, $a = 0.75$, the correlation is very similar to it with total symmetric recurrent network.

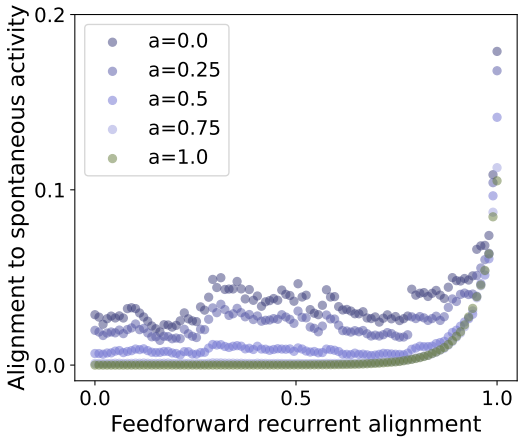


Figure 3.10 Alignment to spontaneous activity with respect to the feedforward recurrent alignment and influence from the degree of symmetry for asymmetric RNNs with symmetrized interactions. As a control group,a fully symmetric RNN ($a = 1.0$) is represented with dark green dots. For different degrees of symmetry from $a = 0.75$ to 0, the darker the dots' color, the more asymmetry is in the network. Only modification 3 is evaluated for alignment to spontaneous activity.

Section conclusion After evaluating the rest of modifications 1 (consider real part of complex eigenvectors) and 3 (symmetrization of asymmetric RNNs) with four perspectives of response properties, modification 1 is filtered out because of the lack of orthogonality between eigenvectors after modification. This could be the reason for the large inconsistency between analytical and empirical approximations of effective dimensionality.

Modification 3 performs at the end the best through all four response properties, and could generally fulfill the expectations of correlations between the response properties and feedforward recurrent alignment.

Therefore, modification 3 which aligns the inputs to asymmetric RNNs with a symmetrized interaction matrix is considered to be a good candidate for modeling feedforward recurrent alignment hypothesis in asymmetric recurrent networks.

3.3 Modeling Feedforward Recurrent Alignment Hypothesis on Low-rank Recurrent Neural Networks (Low-rank RNNs)

Although fully recurrent connectivity structure is one of the most popular network models for theoretical neuroscience, there are experimental recordings suggesting that the transformation of sensory stimuli into motor outputs relies on low-dimensional dynamics at the population level [MO18]. Therefore, the low-rank connectivity structure can be a good candidate for understanding the neural mechanism from another perspective.

We hence also try to model the feedforward recurrent alignment hypothesis on the low-rank RNNs to discover if the hypothesis modeling adapted from full-rank RNNs also could work with low-rank RNNs. Hereby, we consider both symmetric and asymmetric RNNs with constructions with and without noise described in section 2.3.1 by eq.(2.55) and eq.(2.56). Under symmetric or asymmetric conditions, we go through both constructions with the four response properties in correlation with the feedforward recurrent alignment score as for full-rank RNNs before. Those four response properties are trial-to-trial correlation, intra-trial stability, dimensionality, and alignment to spontaneous activity.

3.3.1 Evaluation of Feedforward Recurrent Alignment in symmetric Low-rank RNNs based on response properties

We first consider symmetric low-rank networks in constructions with and without noise. For each case, the results of response property analysis based on the modeling of the feedforward recurrent alignment hypothesis for symmetric networks are evaluated with correlations between them.

Low-rank RNNs without random noise The formulation of low-rank RNNs is followed by eq.2.57 with the rank G significantly smaller than the number of neurons n. The eigenvalues of symmetric low-rank RNNs are real numbers.

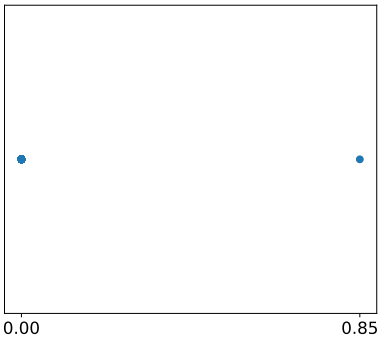


Figure 3.11 Eigenvalue distribution of symmetric low-rank RNNs without random noise. The eigenvalues of symmetric low-rank RNNs are real numbers (x-axis). With rank $G = 1 \ll n = 200$ the number of neurons, 1 eigenvalue takes the value of normalization factor $R = 0.85$, and the rest $n - G = 199$ eigenvalues equal 0.

As shown in Figure 3.11, only two real eigenvalues are seen. A number of G eigenvalues equal the normalization factor $R < 1$ that limits the value range by eq.(2.4). The rest of $n - G$ eigenvalues take the value 0.

The phenomenon of bi-polarized eigenvalue distribution is due to the construction of low-rank RNNs here. If considering the case of having the left connectivity vectors as orthonormal basis for construction of RNNs from eq.(2.57), which is

$$J = \frac{1}{n} \sum_{g=1}^{G} l^{(g)} l^{(g)T} = \sum_{g=1}^{G} \frac{1}{n} l^{(g)} l^{(g)T} . \tag{3.11}$$

If the rank equals the number of neurons, $G = n$, the formulation of low-rank matrix eq.(3.11) is at the same time a symmetric full rank matrix with eigenvectors $\left\{ l^{(g)} \right\}$ and all eigenvectors correspond to the same eigenvalue $\frac{1}{n}$. Since the eigenvalues are re-scaled by parameter $R < 1$ to enable the stable steady state, the eigenvalues are equal to R based on eq.(2.4).

Now, if the rank G is smaller than the number of neurons n, the formulation of low-rank RNN eq.(3.11) can be rewritten as

$$J = \sum_{g=1}^{G} \frac{1}{n} l^{(g)} l^{(g)T} + 0 = \sum_{g=1}^{G} \frac{1}{n} l^{(g)} l^{(g)T} + \sum_{g=G+1}^{n} 0 l^{(g)} l^{(g)T} . \tag{3.12}$$

So, there are G basis vectors that have eigenvalue $\frac{1}{n}$, which is further re-scaled to $R < 1$. The rest of $n - G$ eigenvectors have eigenvalue 0. This then results the eigenvalue distribution shown in Figure 3.11.

We first look at the trial-to-trial correlation and expect a positive correlation with the feedforward recurrent alignment. When aligning the inputs to the symmetric recurrent alignment, the inputs-distribution has the mean vector matched to the eigenvectors of the low-rank interaction matrix J. Due to the feedforward recurrent alignment formulation, the alignment scores are equal to the corresponding eigenvalues because of eq.(2.14).

Since there are only two eigenvalues R and 0 for the rank G smaller than number of neurons n, we assume that there is no continuous correlation but two groups of trial-to-trial correlation values. However, low feedforward recurrent alignment should still correlate with a small trial-to-trial correlation value and a large alignment score with a big trial-to-trial correlation. We expect therefore here a discontinuous positive correlation between trial-to-trial correlation and feedforward recurrent alignment.

As the result in Figure 3.12a shows, the trial-to-trial correlation distributes separately into two groups due to the distribution of eigenvalues. Also as expected, there is still a positive correlation between feedforward recurrent alignment score and trial-to-trial correlation.

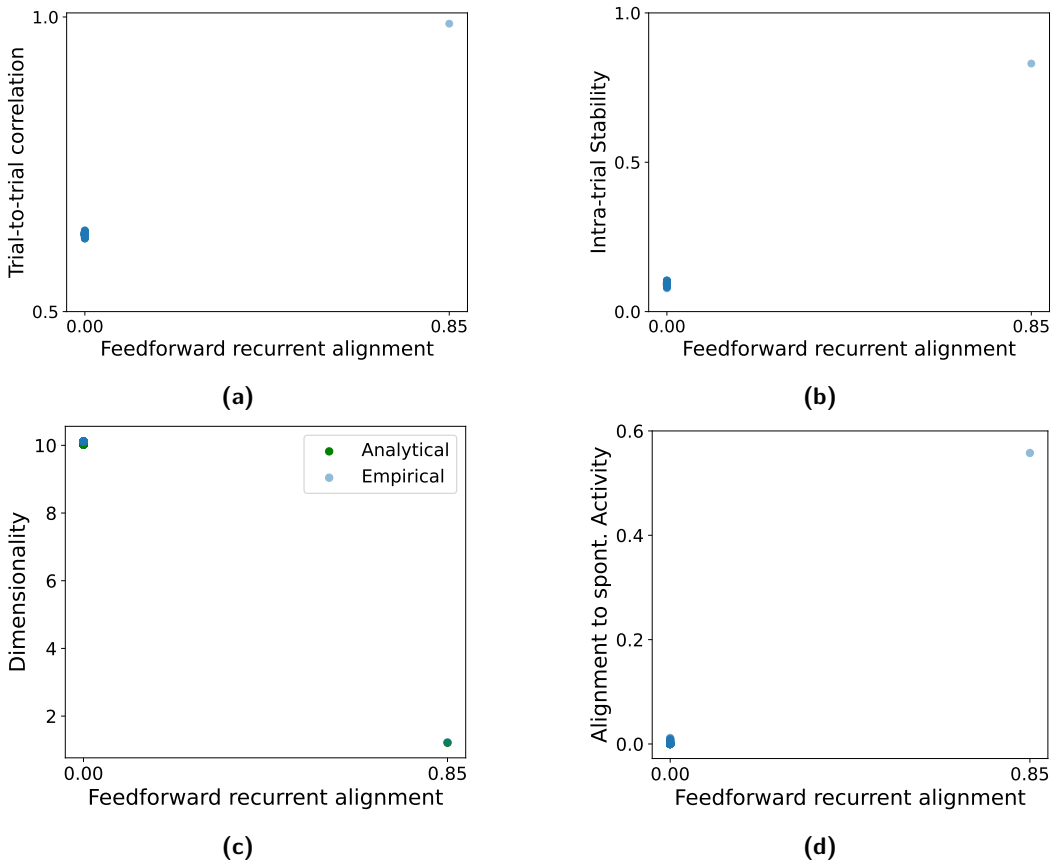


Figure 3.12 Correlation between response properties and feedforward recurrent alignment considering symmetric low-rank RNNs without random noise. Construct symmetric low-rank RNNs under the assumption of having left connectivity vectors equal to right connectivity vectors eq.(3.11). Rank G is the number of connectivity vectors and is significantly smaller than the number of neurons n. Here $G = 1$ and $n = 200$. To evaluate the feedforward recurrent alignment hypothesis, the correlations between response properties and the modeled feedforward recurrent alignment score are considered.
(a) Trial-to-trial correlation (y-axis) in correlation with feedforward recurrent alignment score (x-axis).
(b) Intra-trial stability (y-axis) in correlation with feedforward recurrent alignment score (x-axis).
(c) Dimensionality (y-axis) calculated analytically (green dots, eq.(2.33)) and empirically (blue dots, eq.(2.34)) in relationship with feedforward recurrent alignment score (x-axis).
(d) Correlation between alignment to spontaneous activity (y-axis) and feedforward recurrent alignment score (x-axis).

Analogous to the trial-to-trial correlation, we receive the results for intra-trial stability (Figure 3.12b) and alignment to spontaneous activity (Figure 3.12c) also be a discontinuous positive correlation to feedforward recurrent alignment, meanwhile a discontinuous negative correlation for dimensionality (Figure 3.12d). Due to the same reason of existing only two groups of eigenvalues at R and 0, the intra-trial stability value, alignment to spontaneous activity, and dimensionality distribute also separately into two groups while keeping the expected correlations with feedforward recurrent alignment.

Low-rank RNNs with random noise Low-rank symmetric RNNs with random noise include an extra part of symmetrized Gaussian distributed matrix defined as eq.(2.56) in section 2.3.2. If analog to eq.(3.11) considering the right connectivity vectors equal to the left connectivity vectors, the symmetric low-rank RNNs with random noise can be formulated with a part of symmetric low-rank RNN and a part of symmetrized Gaussian distributed matrix J_{rand}.

$$J = \frac{1}{n} \sum_{g=1}^{G} l^{(g)} l^{(g)T} + J_{\text{rand}} \,. \tag{3.13}$$

The symmetric random part J_{rand} should provide more dynamics in the network.

Since J_{rand} is a full-rank symmetric matrix, the final recurrent network J is also a full-rank symmetric matrix despite the low-rank symmetric part with rank G. Therefore, the correlations between the response properties and feedforward recurrent alignment score are similar to the case of full-rank symmetric RNNs (Figure 3.1, 3.2, 3.2b, and 3.4b).

As a result, the feedforward recurrent alignment hypothesis can be modeled. The expected relationships between feedforward recurrent alignment and response properties from section 3.1) are fulfilled, as shown in Figure 3.13. However, the influence of low-rank construction is therefore overwritten by the random noise, which leads to the final dynamics not much different from the case with general random symmetric RNNs.

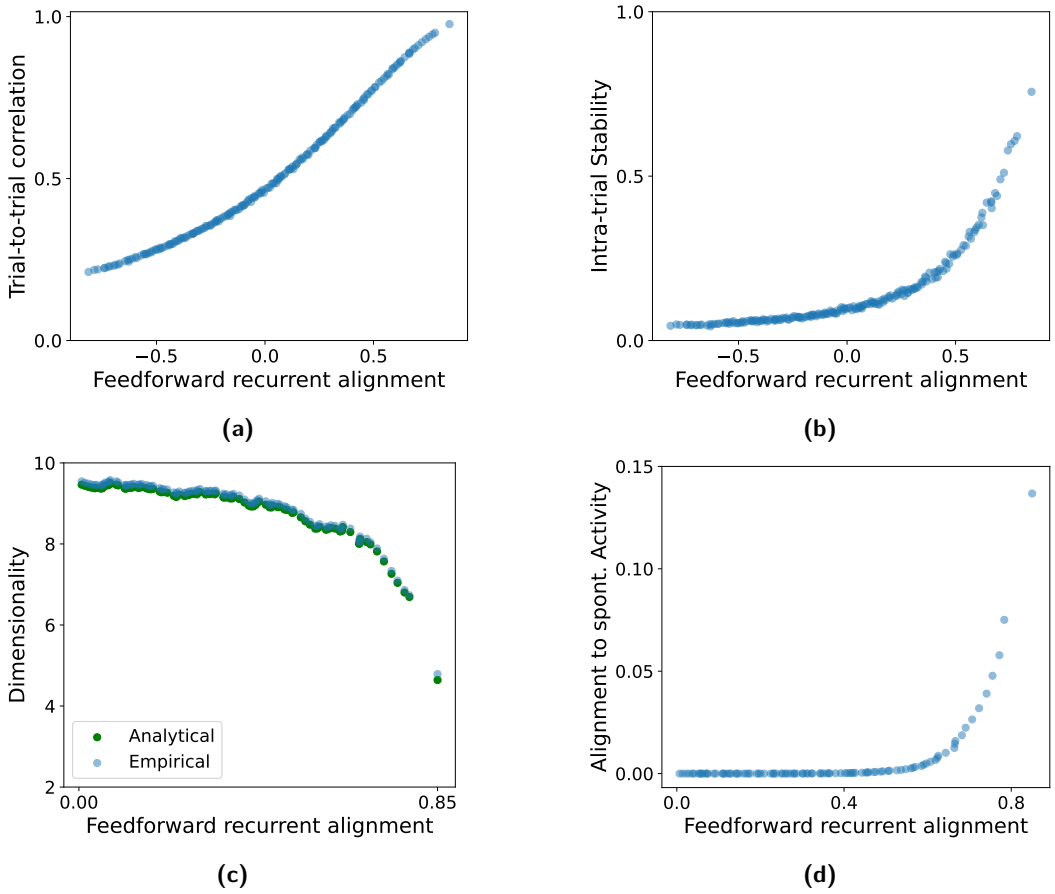


Figure 3.13 Correlation between response properties and feedforward recurrent alignment in symmetric low-rank RNNs with random noise. Besides the part of a low-rank matrix with rank G, low-rank RNNs with noise include an extra part of symmetrized random Gaussian matrix from eq.(3.13). $G = 1$ here.
(a) Trial-to-trial correlation (y-axis) against feedforward recurrent alignmnet score (x-axis).
(b) Intra-trial stability (y-axis) against feedforward recurrent alignment score (x-axis).
(c) Dimensionality (y-axis) in dependence of feedforward recurrent alignment score (x-axis). Green dots represent the analytical calculation of dimensionality eq.(2.33) and blue dots for empirical approximation of eq.(2.34)
(d) Alignment to spontaneous activity (y-axis) against feedforward recurrent alignment (x-axis).

Section conclusion Considering symmetric low-rank RNNs without random noise, the feedforward recurrent alignment keeps positive correlations to trial-to-trial correlation, intra-trial stability, and alignment to spontaneous activity. Besides, the expected negative correlation between dimensionality and feedforward recurrent alignment is also fulfilled. However, the network construction eq.(3.11) leads to a simple distribution of eigenvalues (Figure 3.11) and in turn also the distribution of feedforward recurrent alignment score. As a result, only a simple discontinuous correlation of feedforward recurrent alignment to response properties can be observed (Figure 3.12).

If adding random noise to disturb the simple dynamics of symmetric low-rank RNNs with eq.(3.11), the final construction eq.(3.13) is a full-rank symmetric RNN. As a result, the correlations of feedforward recurrent alignment to response properties are similar to the results observed in general symmetric RNNs from section 3.1. Although the expected phenomena are seen, the effect of low-rank RNNs cannot be significantly detected.

In total, after considering the symmetric low-rank RNNs with and without random noise, the feedforward recurrent alignment hypothesis can be modeled in both cases. The relationships between feedforward recurrent alignment and response properties also meet the expectations.

3.3.2 Different Constructions influence the impact of rank in Asymmetric Low-rank RNNs based on Response Properties

Asymmetric low-rank RNNs can have more complex dynamics than symmetric low-rank RNNs. Section 2.3.2 also introduces the construction of asymmetric low-rank RNNs with or without random noise. Generally, if taking the set of left connectivity vectors different from the set of right connectivity vectors, asymmetric matrices can be formulated by eq.(2.55), (2.56).

Since asymmetric low-rank RNNs are only a certain case of asymmetric RNNs with specific constructions, they also have the problem of complex eigenvectors and eigenvalues when modeling the feedforward recurrent alignment hypothesis. Therefore, from the previous results with general asymmetric RNNs, we choose modification 3 from section 2.2.3 with aligning inputs to symmetrized RNN because of its overall best performance among all considered modifications as mentioned in section 3.2.

Low-rank RNNs without random noise Asymmetric low-rank RNNs without random noise consist of non-equal left and right connectivity vectors, which are mutually orthogonal illustrated in Figure 2.4 and eq.(2.55).

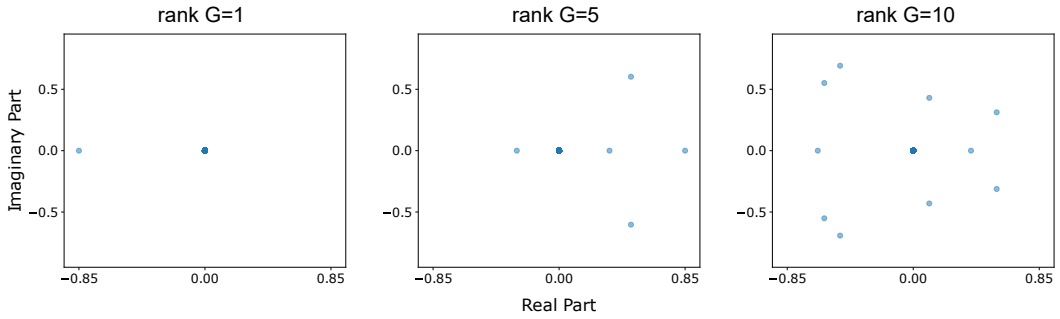


Figure 3.14 Eigenvalue distribution in dependence of rank with asymmetric low-rank RNNs without random noise. The asymmetric low-rank RNNs without random noise are constructed by mutually orthogonal non-equal left and right connectivity vectors defined by eq.(2.55). The number of vector pairs G determines the rank of the interaction matrix, which should be significantly smaller than the number of neurons n ($n = 200$). We can observe that the rank G can influence the distribution of eigenvalues for asymmetric low-rank RNNs in complex plane. Plotting eigenvalues in the complex plane with x-axis for the real part and y-axis for the imaginary part. Low-rank RNNs with rank G equals 1, 5, and 10.

For symmetric low-rank RNNs without random noise, as long as the rank G is significantly smaller than the number of neurons n as defined, the distribution of eigenvalues is separated into two groups independent of the rank (Figure 3.11). While for asymmetric low-rank RNNs without random noise, the influence of rank G is more significant, as shown in Figure 3.14.

The construction of asymmetric low-rank RNNs without random noise defined by eq.(2.55) can be understood as G pairs of presented connectivity vectors, while the rest of $n - G$ pairs are suppressed, rewritten with

$$J = \frac{1}{n}\sum_{g=1}^{G} l^{(g)} r^{(g)T} = \sum_{g=1}^{G} \frac{1}{n} l^{(g)} r^{(g)T} + \sum_{g=G+1}^{n} 0 l^{(g)} r^{(g)T} . \tag{3.14}$$

Therefore, exactly G number of eigenvalues do not concentrate at 0 but the rest $n-G$ eigenvalues are. Since the mathematical construction eq.(2.55) for asymmetric low-rank RNNs is not an eigendecomposition, the eigenvalues after re-scaling by normalization parameter R (method section 2.3.2) do not exactly have magnitude R as at symmetric low-rank RNNs from eq.(3.12).

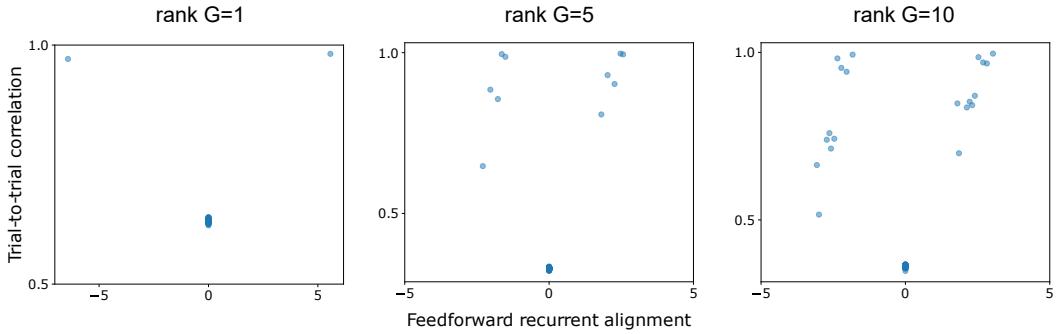


Figure 3.15 Relationship of trial-to-trial correlation and feedforward recurrent alignment in dependence of rank by asymmetric low-rank RNNs without random noise. The rank G in asymmetric low-rank RNNs from eq.(3.14) influence the eigenvalue distribution and thus also the distribution of trial-to-trial correlation. Rank G is significantly smaller than the number of neurons n by definition ($n = 200$). Illustrate the correlation between feedforward recurrent alignment and trial-to-trial correlation under varied ranks G equals 1, 5, and 10.

We expect that the distribution of the correlation between feedforward recurrent alignment and trial-to-trial correlation will be related to the eigenvalue distribution under different ranks. Besides, a positive correlation between feedforward recurrent alignment and trial-to-trial correlation should be maintained independent of the change of rank.

As expected, the rank influence the eigenvalue distribution and thus also the correlation distribution between feedforward recurrent alignment and trial-to-trial correlation. However, the positive correlation is not kept over the whole range of feedforward recurrent alignment. In fact, it can be observed that the distribution is almost identical in the negative and positive range of feedfroward recurrent alignment. In Figure 3.14, the number of dots in each half-range also exactly equal the number of range, which is also the number of non-zero eigenvalues. There are now doubled number of non-zero feedforward recurrent alignment scores as expected.

We hypothesize that the duplication could be due to the symmetrization of asymmetric low-rank RNNs with a simple construction of only applying connectivity vectors. With the symmetrization, a part of the reflected network also contributes to eigenvalue expression and therefore increases the number of presented feedforward recurrent alignment scores. Symmetrization of eq.(3.14) leads to

$$J_{\text{sym}} = \frac{J + J^T}{2} = \frac{1}{2n}\left(\sum_{g=1}^{G} l^{(g)} r^{(g)T} + \sum_{g=1}^{G} r^{(g)} l^{(g)T}\right) = \frac{1}{2n}\sum_{g=1}^{G} l^{(g)} r^{(g)T} + \frac{1}{2n}\sum_{g=1}^{G} r^{(g)} l^{(g)T}. \tag{3.15}$$

The second part of eq.(3.15) is presumed to be the reason for the distribution in the negative range of feedforward recurrent alignment in Figure 3.15. The construction from eq.(3.15) doubles the network in two sub-networks separately. Each sub-network is spread over the negative or the positive range of feedforward recurrent alignment. A positive correlation between feedforward recurrent alignment and trial-to-trial correlation can be observed in each half-range. But there is in the end no positive correlation overall in the feedforward recurrent alignment score range.

Thus, the modeling of the feedforward recurrent alignment hypothesis in the case of asymmetric low-rank RNNs without noise cannot fulfill the phenomenon over the whole range of alignment between inputs and symmetrized interaction matrix.

Low-rank RNNs with random noise Considering low-rank RNNs without random noise from above, the simple construction by eq.(3.14) does not perform well with the modification of alignment to symmetrized RNNs. Now, we add a Gaussian distributed asymmetric random part to the simple construction. As a result, the asymmetric recurrent network is composed of a low-rank part from connectivity vectors and a part of full-rank random noise defined by eq. (2.56) and explained in method section 2.3.2. The final interaction matrix J is therefore full-rank and asymmetric.

If the feedforward recurrent alignment hypothesis can be modeled in the asymmetric low-rank RNNs with random noise, positive correlations between feedforward recurrent alignment score and trial-to-trial correlation, intra-trial stability, and alignment to spontaneous activity are expected. Moreover, dimensionality should be negatively correlated with feedforward recurrent alignment. Based on observations from symmetric low-rank RNNs with random noise in Figure 3.13, the results can be similar to the case with general asymmetric full-rank RNNs in section 3.2.2. In other words, the dynamics of low-rank can be suppressed by the full-rank noise dynamics.

The results in Figure 3.16 confirmed our expectations of correlations between response properties and feedforward recurrent alignment. Besides, the correlations have high similarity to the results we got from general asymmetric RNNs. Under the strong effect of full-rank random noise, the difference caused by the low-rank part is not significant (comparing blue and orange dots in Figure 3.16). The results in section 3.2 indicate that the dispersion could be increased with an increased proportion of asymmetry in RNNs. Since the final construction for asymmetric low-rank RNNs with noise in eq.(2.56) consists only of asymmetric networks, the dispersion is expected to be large, especially at dimensionality.

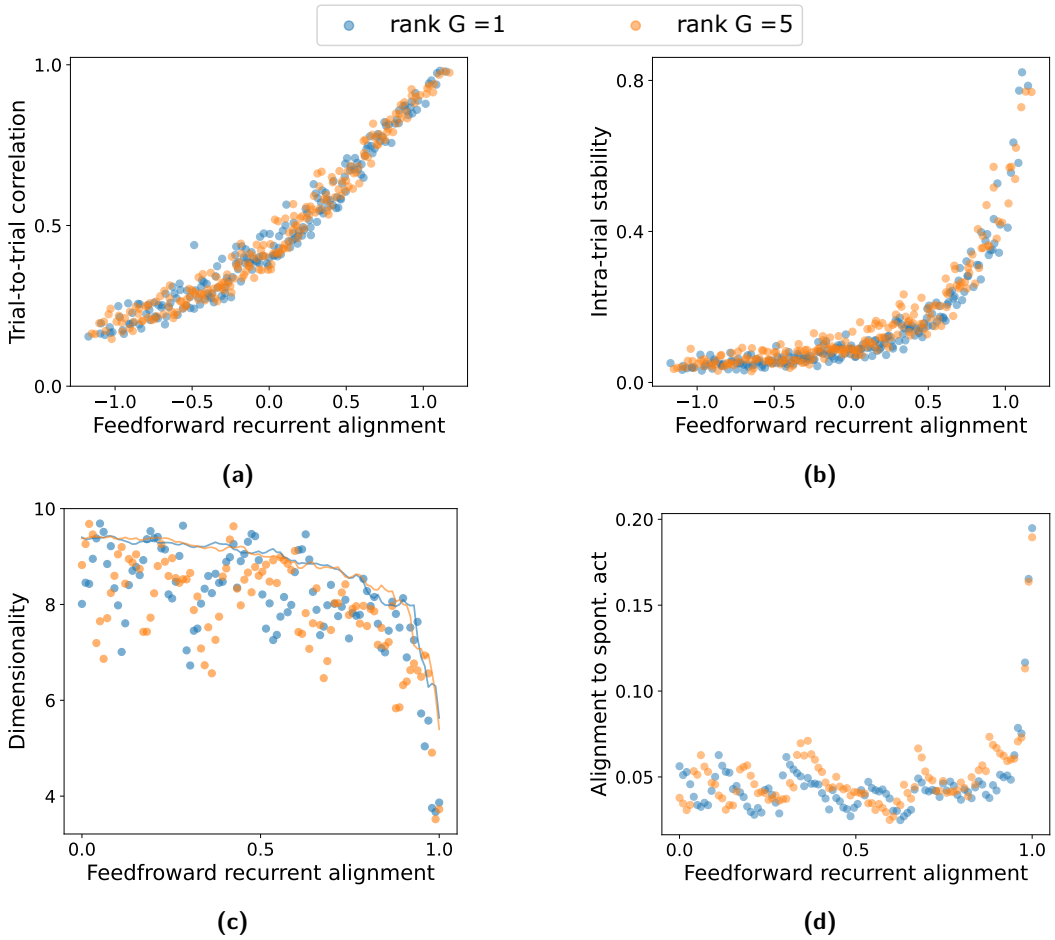


Figure 3.16 Correlations between feedforward recurrent alignment and selected response properties for asymmetric low-rank RNNs with random noise. In addition to the low-rank part constructed only by left and right connectivity vectors, a Gaussian distributed full-rank asymmetric random noise is included in the network dynamic eq.(2.56). Due to the full-rank random noise, the results have a large similarity to results from general asymmetric RNNs in section 3.2.2. Two different ranks are taken to assess the influence of ranks on correlations. The rank G should be significantly smaller than the number of neurons n. With $n = 200$, comparing ranks $G = 1$ in blue dots with $G = 5$ in orange:
(a) Correlation between feedforward recurrent alignment (x-axis) and trial-to-trial correlation (y-axis).
(b) Correlation between feedforward recurrent alignment (x-axis) and intra-trial stability (y-axis).
(c) Dimensionality (y-axis) calculated analytically (lines, blue line for $G = 1$ and orange line for $G = 5$) and empirically in correlation with feedforward recurrent alignment (x-axis).
(d) Alignment of evoked patterns to spontaneous activity (y-axis) in relationship with feedforward recurrent alignment (x-axis).

Section conclusion We explore in this section the modeling of feedforward recurrent alignment in asymmetric low-rank RNNs. Two types of constructions are taken into account: 1) single part of the non-equal left and right connectivity vectors and 2) with additional full-rank random noise defined in method section 2.3.2.

The simple construction without random noise has an eigenvalue distribution depending on the rank. Keeping the rank significantly smaller than the number of neurons, the number of non-zero eigenvalues equals the rank (Figure 3.14). Thus, the number of non-zero feedforward recurrent alignment scores also depends on the rank. Due to the modification of aligning inputs to a symmetrized interaction matrix for calculation of feedforward recurrent alignment score, the correlation between alignment score and trial-to-trial correlation is separated into two sub-groups. As a result, although in each sub-region the positive correlation between alignment score and trial-to-trial correlation is kept, there is no global positive correlation over the total range (Figure 3.15).

Adding the full-rank random noise can avoid the effect of doubling caused by symmetrization. The expected positive correlations between feedforward recurrent alignment and trail-to-trial correlation, intra-trial stability, and alignment to spontaneous activity are kept (Figure 3.16). Large dispersion is caused by the total asymmetry of the interaction matrix. Despite large dispersion, the negative correlation between dimensionality and feedforward recurrent alignment can be observed. However, the expression of low-rank connections is suppressed by the full-rank random noise, such that there is no significant difference between low-rank RNNs with various ranks.

3.4 White Noise Evoked Activity Can Help to Approximate Dominant Activity Direction in Response Space for Unknown Asymmetric Recurrent Networks

Normally during experimental procedures, the complete recurrent network structure of the laboratory animal is difficult to access. Therefore, even if the feedforward recurrent alignment hypothesis can be theoretically underpinned well in both symmetric and asymmetric recurrent networks (section 3.1 and section 3.2.2), the hypothesis cannot be well tested in experimental environments.

Motivated by some predictions from a series of theoretical frameworks [HM23, MYDF09] where was shown that the matching between inputs and spontaneous activity can lead to more reliable evoked responses and more efficient transmission across cortical networks, we consider the idea of generating spontaneous-like activity pattern for alignment as an approximation of the original recurrent network.

Our goal here is to achieve a theoretical framework for the feedforward recurrent alignment without knowing the exact recurrent network structure to mimic the experimental conditions. Besides, we study if this framework could also support the previous founding that better alignment between input and spontaneous activity leads to reliable response activity.

Mulholland et al. [HM] suggested an experimental method with white noise to generate spontaneous-like activity patterns. We model the spontaneous-like activity with white noise inputs and its evoked activity for the construction of a modified feedforward recurrent alignment score. Instead of eigenvectors of the interaction matrix, we apply the principal components of white-noise-evoked activity pattern samples for the calculation of feedforward recurrent alignment. Evaluation of our framework covers mainly four perspectives of response properties, namely trial-to-trial correlation, intra-trial stability, dimensionality, and alignment of evoked activity to spontaneous activity. They are introduced in the method section 2.4.1.

3.4.1 Input Alignment with White-noise-evoked Activity Pattern Support Previous Theoretical Frameworks

Positive Monotonic Correlation between Feedforward Recurrent Alignment Score and Eigenvalues of White-noise-evoked Activity Pattern

The feedforward recurrent alignment score should reflect how well the input is aligned with the dominant direction of activity patterns generated by RNNs. Since the original recurrent network is unknown, we apply the white-noise-evoked activity pattern for the alignment defined by eq.(2.60). Here we align the input to the eigenvectors of the covariance matrix of the white-noise-evoked activity pattern samples

instead of the eigenvectors of original recurrent network interactions.

A positive monotony correlation between the feedforward recurrent alignment score and variance ratio of the aligned white-noise-evoked activity pattern is essential for guaranteeing the functionality of the feedforward recurrent alignment hypothesis. Due to the selective response amplification eq.(3.1), the dominant variance ratios of the white-noise-evoked pattern should determine the strength and reliability of responses. Thus, the feedforward recurrent alignment should have a high value when aligning to the corresponding principal components of dominant variance ratios, indicating that the evoked response would be reliable. The monotony guarantees that the alignment to dominant principal components is the only source for the increase of the feedforward recurrent alignment score.

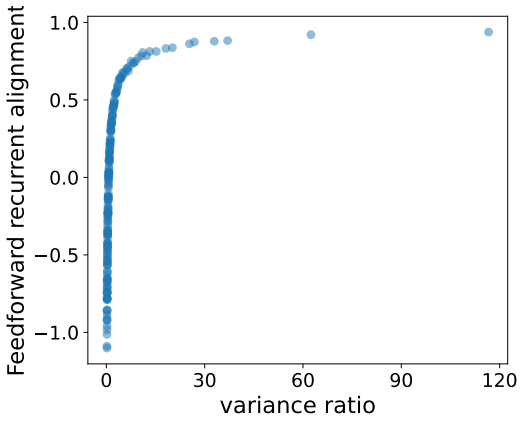


Figure 3.17 Correlation between feedforward recurrent alignment and variance ratio of white-noise-evoked activity patterns. With an unknown recurrent network structure, a white-noise-evoked spontaneous-like activity pattern is applied for aligning inputs. The feedforward recurrent alignment score is formulated with principal components of the covariance matrix from white-noise-evoked activity eq.(2.60) (500 samples). The alignment score should be positive and monotonously correlated with the corresponding variance ratio.

The result in Figure 3.17 illustrates the correlation between the feedforward recurrent alignment score calculated with principal components and its corresponding variance ratios of white-noise-evoked activity. The dominant variance ratios are also associated with higher feedforward recurrent alignment. Besides, a significant positive monotony can be observed.

Thus, the feedforward recurrent alignment measured with white-noise-evoked activity through eq.(2.60) can be a good candidate measurement for supporting the predictions that inputs aligning well to spontaneous activity can generate more reliable responses.

Evaluation of Approximation with White Noise through Response Properties

To evaluate if the candidate measurement through aligning inputs to white-noise-evoked activity eq.(2.60) can quantify reliable responses without knowing the recurrent structure, four response properties are taken into account: 1) trial-to-trial correlation, 2) intra-trial stability, 3) dimensionality, and 4) alignment of evoked activity to spontaneous activity. Those four properties are observed in experienced ferrets' primary visual cortex [TWFK23] for characterizing the reliability of neural responses. Both in symmetric and asymmetric RNNs, the theoretical modeling of the feedforward recurrent alignment hypothesis also supports the experimental observations in ferrets.

If the newly constructed feedforward recurrent alignment with white-noise-evoked activity can well quantify the feedforward recurrent alignment hypothesis for aligning to white-noise-evoked activity pattern eq.(2.59), the correlations between feedforward recurrent alignment score and four response properties that mentioned above should coincide with prior results from feedforward recurrent alignment hypothesis for full-rank RNNs in sections 3.1 and 3.2.

Therefore, we expect at least that high feedforward recurrent alignment corresponds with high trial-to-trial correlation, intra-trial stability, and alignment to spontaneous activity, while with low dimensionality. The results in Figure 3.18 fulfill our expectations even under different degrees of symmetry in the original networks.

However, relatively high trial-to-trial correlation, low dimensionality, and high alignment to spontaneous activity occur when the feedforward recurrent alignment score is small. In other words, a large discrepancy exists in a range of low feedforward recurrent alignment.

Multiple perspectives could lead to the discrepancy. One assumption for the discrepancy is that the influence of increased asymmetry in the network. Another reason could be that for principal components with small variance ratios, a single principal component cannot approximate the original eigenvectors as well as dominant principal components.

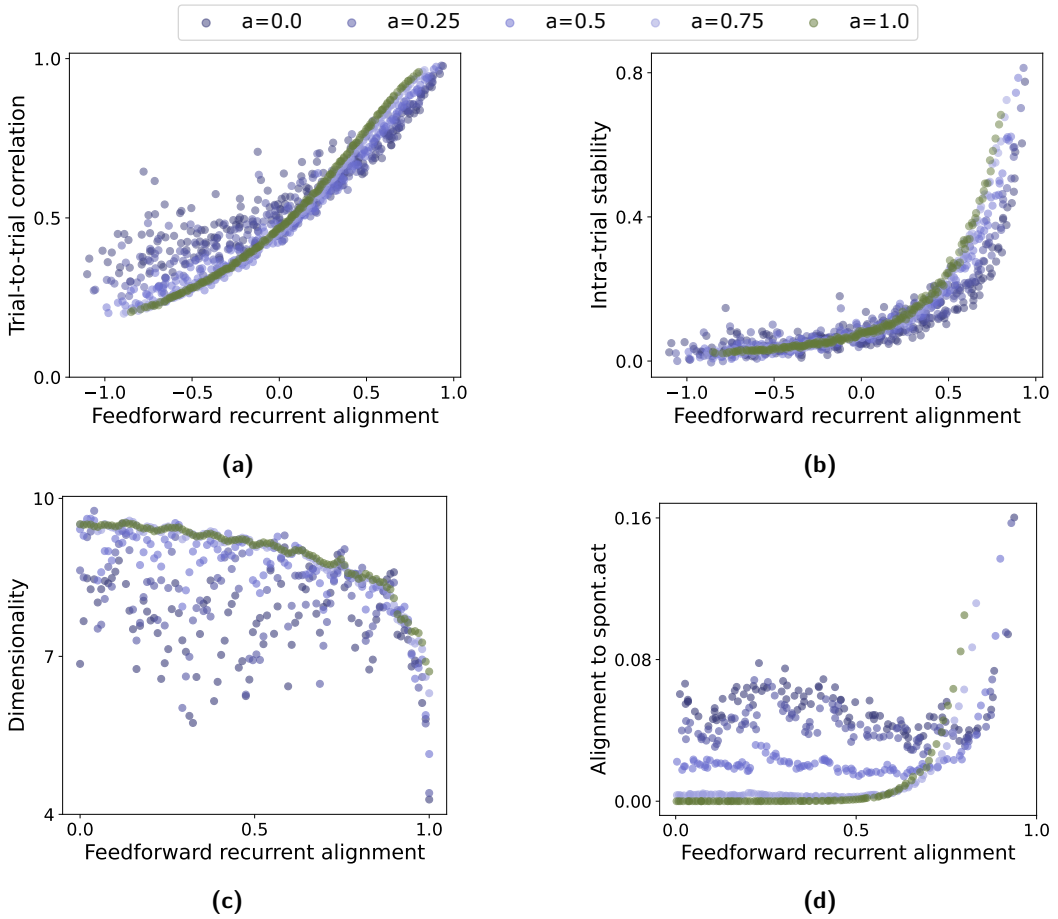


Figure 3.18 Correlation between feedforward recurrent alignment and selected response properties under aligning inputs to white-noise-evoked activity pattern. With inputs aligned to white-noise-evoked spontaneous-like activity, the modified feedforward recurrent alignment eq.(2.60) aligns inputs to principal components of white-noise-evoked activity pattern. The correlations between response properties and feedforward recurrent alignment are modeled with varied degrees of symmetry a for RNNs eq.(2.40) from $a = 0$ complete asymmetric to $a = 1.0$ complete symmetric.
(a) Correlation between feedforward recurrent alignment (x-axis) and trial-to-trial correlation (y-axis).
(b) Correlation between feedforward recurrent alignment (x-axis) and intra-trial stability (y-axis).
(c) Empirical approximation of effective dimensionality (y-axis) in correlation with feedforward recurrent alignment (x-axis).
(d) Relationship between feedforward recurrent alignment (x-axis) and alignment to spontaneous activity (y-axis).

3.4.2 Iterative Feedforward Recurrent Alignment from Low-dimensional Inputs Indicates Alignment Improvement

With experimental producible low-dimensional inputs, we aim to discover how the RNNs amplify those inputs and if the amplification can provide new insights about alignment development.

As described by the method section 2.4.2, low-dimensional inputs are modeled with eq.(2.70). Under the feedforward recurrent alignment hypothesis, the feedforward recurrent alignment score defined by eq.(2.72) should reflect how well the input is aligned to dominant activity patterns for generating reliable neural responses. Repeatedly applying the prior evoked activity pattern as inputs could reflect possible structural change after recurrent amplification thus indicating a possible plasticity adaptation of response activity patterns.

The updated dynamic of alignment in dependence of times for repeatedly applying prior response as inputs is an oscillation shown in Figure 3.19. At the n-th times using the prior response r_{n-1} as inputs, the feedforward recurrent alignment score ν_n defined in eq.(2.76) only depends on r_{n-1}. The oscillation can only be a consequence of the oscillation of evoked responses. So, repeatedly applying the prior response as feedforward input can lead to a stable oscillation of response activities r_n and therefore also the feedforward recurrent alignment score.

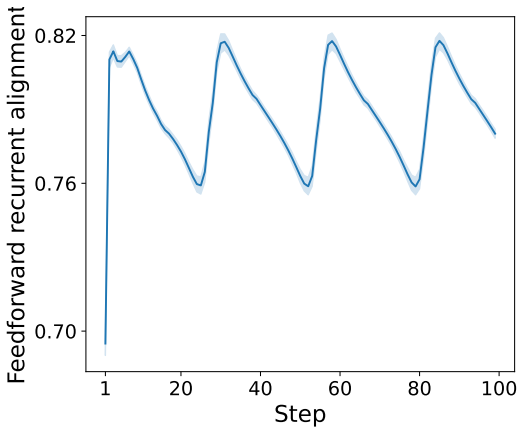


Figure 3.19 Dynamic of iterative feedforward recurrent alignment through applying prior response as input. Starting with low-dimensional input eq.(2.70) with $\beta_{\text{Low}} = 5$, the evoked activity pattern is applied as input for updating feedforward recurrent alignment. Iteratively repeating this procedure results in successive updates of feedforward recurrent alignment score eq.(2.76). For statistics 500 response samples are considered. The shadow indicates the 95% confidence interval.

3.5 Hebbian Learning of Feedforward Network Leads to Better Alignment between Feedforward Input and Recurrent Network

Now considering the different connection structures inside neural layers in the brain, we extend the focus from only on the recurrent network to also take the feedforward network structure into account, which delivers the feedforward inputs to the recurrent network illustrated in Figure 2.6.

The main difference compared to prior theoretical experiments is, that the feedforward recurrent network allows plasticity in the feedforward network. Furthermore, we want to verify the feedforward recurrent alignment hypothesis during the network learning process. That is the change of alignment between dynamic feedforward input and recurrent network.

For firstly a basic understanding of the influence of plasticity, we consider the feedforward network with single neuron input and fixed symmetric recurrent network illustrated in Figure 2.7). Only the feedforward network is updated with the Hebbian rule, which is the classic rule for activity-dependent synaptic plasticity [DA05]. The Hebbian rule describes the dynamics of feedforward weights through an ordinary differential equation eq.(2.80). Moreover, the rule reflects the hypothesis for a principle: neurons that fire together wire together. With the help of the Euler scheme, the time development of feedforward weight can be approximated eq.(2.85).

To track the alignment between feedforward input and recurrent network, two alternatives were regarded:

- Based on the time-dependent feedforward weight dynamics, project weights to the space spanned by eigenvectors of recurrent networks. Observe the projection coefficients distribution for dominant eigenvectors of recurrent networks.
- Update the feedforward recurrent alignment score simultaneously with a time-related update of feedforward input. Observe the development of feedforward recurrent alignment score in dependence on time.

3.5.1 Feedforward Weights are Determined by Dominant Eigenvectors after Learning

Projecting the feedforward weights to the space spanned by eigenvectors of recurrent networks results in a linear combination as the presentation of feedforward weights through eigenvectors with suitable coefficients eq.(2.87). The coefficients reflect the influence of corresponding eigenvectors on feedforward weights. A larger absolute value of one projection coefficient indicates a stronger influence of the corresponding eigenvector on feedforward weights.

According to the feedforward recurrent alignment hypothesis, a large alignment between feedforward input and recurrent network can be obtained by aligning

feedforward input proportional to dominant eigenvectors. Therefore, feedforward weights with large coefficients for dominant eigenvectors would quantify a good alignment between feedforward input and recurrent network. Since the projection of feedforward weights could consist of multiple dominant eigenvectors, the projection ratio defined by eq.2.89 takes the projection coefficients of the first twenty most dominant eigenvectors into account and could thus capture the dominance from more patterns.

The modeling of the projection ratio follows the update of feedforward weights according to the eq.(2.88). If during the learning, the feedforward network can generate inputs fitting better to the dominant eigenvectors of the recurrent network, the feedforward weight should concentrate on dominant eigenvectors. Thus, the coefficients for dominant eigenvectors would be expected to be larger during learning. As a result, the projection ratio should increase with the development of time. According to the feedforward recurrent hypothesis, this could lead to more reliable recurrent responses.

The results illustrated by Figure 3.20 support the assumption that during learning, the feedforward input aligns better with the recurrent network by strengthening the influence of dominant eigenvectors of the recurrent network. At least the first twenty most dominant eigenvectors gain more weights in the linear combination for the projection during the Hebbian learning process of the feedforward network. Until the stable state, almost all projection weights concentrate at the first twenty dominant eigenvectors since the projection ratio reaches almost 1.

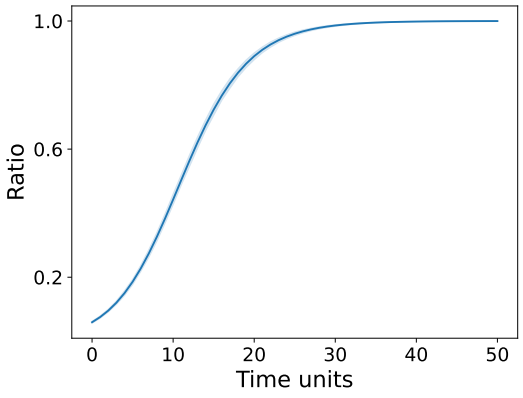


Figure 3.20 Time related development of projection ratio. Projection ratio defined by eq.(2.89) quantifies the strength of linear dependency between feedforward weights and the first twenty most dominant eigenvectors of the recurrent network. With the time-dependent update of feedforward weights eq.(2.85), the projection coefficients are also updated synchronously. Step width Δt for the Euler scheme is 0.1 over the total duration of $T = 50$ time units. For statistics, 50 repeats with different initial feedforward weights were implemented. The shadow indicates the 95% confidence interval.

3.5.2 Feedforward Recurrent Alignment Score Increases through Learning

Another alternative considers the change of feedforward recurrent alignment score directly over time simultaneously with the update of feedforward weights according to the eq.(2.91). Besides, the derivative of the feedforward recurrent alignment score dependent on time can also be explicitly formulated with the help of the Hebbian learning dynamic eq.(2.84), resulting the final derivative described by eq.(2.98).

According to the results from the first alternative, the feedforward weights improve their alignment to the recurrent network over time. The feedforward recurrent alignment score should increase if the feedforward inputs align better with the recurrent network. Since feedforward inputs are proportional to the feedforward weights due to the single input rate eq.(2.90), the development of the feedforward recurrent alignment score should be synchronized to the dynamic of projection ratio shown in Figure 3.20.

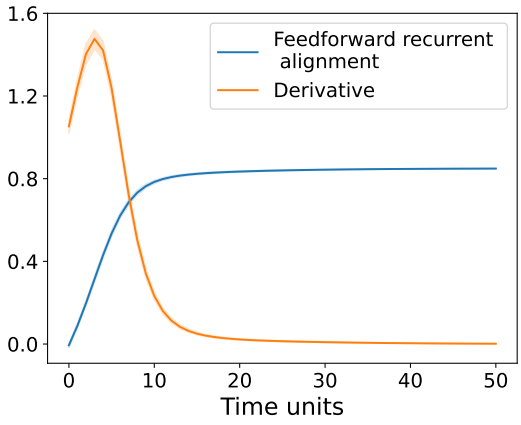


Figure 3.21 Dynamics of feedforward recurrent alignment score and its derivative. Simultaneously with the update of feedforward weight, the feedforward recurrent alignment score (shown with blue line) can also be updated simultaneously with eq.(2.91). The derivative of alignment score (shown with orange line) is determined by eq.(2.98). Step width Δt for the Euler scheme is 0.1 over a total duration of $T = 50$ time units. For statistics, 50 repeats with different initial feedforward weights were implemented.

Implementation of feedforward recurrent alignment score eq.(2.91) and its derivative eq. (2.98) along time coincide with the results from the first alternative in Figure 3.20. As shown in Figure 3.21, the feedforward recurrent alignment score also increases over time until stable state. This fulfills our expectations that the two alternatives reflect both the phenomenon of increased alignment between feedforward input and recurrent network during the Hebbian learning considering a single input rate.

4 Discussion

Initial model for symmetric RNNs

The previous work from M. Kaschube lab and D. Fitzpatrick lab [TWFK23] introduced the feedforward recurrent alignment hypothesis first. Their experimental results compared neural activities in visually naive and visually experienced ferrets' visual cortexes, demonstrating that experience plays a critical role in increasing neural response consistency both across and within trials. Moreover, they found out that the initial evoked responses at eye-opening consist of novel patterns, distinct from spontaneous activity of the endogenous network and that visual experience drives the development of low-dimensional, reliable representations aligned with spontaneous activity.

A computational model was developed with symmetric RNNs to give the proposal that high alignment of feedforward and recurrent networks can effectively amplify the novel activity patterns induced by the onset of visual experience [TWFK23]. The modeling results (section 3.1) support the experimental observations, indicating the potential of the feedforward recurrent alignment hypothesis. Since symmetric RNNs are simple conceptual models for cortical networks, the investigation of broader network structures could improve the modeling and even give new insights into the understanding of underlying mechanisms during experience-driven development.

Extension modifications for asymmetric RNNs

The closest extension of symmetric RNNs can be the asymmetric RNNs. Without symmetry, the recurrent interactions can generally enable the alignment of feedforward and recurrent networks in a complex plane. Complex patterns are however difficult to interpret in the context of neuroscience. Therefore, modifications to the original model are necessary to overcome this problem. We consider three possible modifications and evaluate them based on the experimental and modeling results from the prior work from M. Kaschube lab and D. Fitzpatrick lab [TWFK23]. After the verification, the modification considering the alignment of feedforward and symmetrized network has the best performance.

This indicates that in the case of general asymmetric RNNs, the alignment in the complex plane can be well approximated through symmetrization of the original RNNs. Besides, for full-rank asymmetric RNNs, symmetrization keeps the most of complex-plane information contained when transformed to a real number among all three modifications.

Test with low-rank RNNs

Beyond the full-rank RNNs, some recent work [MO18, BDV$^+$21, DVB$^+$22] introduced low-rank interaction structures, which was predicted to capture the actual cortical connectivity more realistically than full-rank RNNs. Thus, we test the modeling on both symmetric and asymmetric RNNs with two possible constructions. The difference between the two constructions is the dispersion from full-rank random noise. Symmetric low-rank RNNs coincide with the experimental observations with both constructions. If applying the best modification from full-rank asymmetric RNNs on low-rank asymmetric RNNs without random noise, the symmetrization leads to a duplication of alignment.

If considering the disturbance of random noise, the dynamics from low-rank RNNs will be overwritten and the results did not differ from full-rank RNNs. For asymmetric low-rank RNNs without random noise, other modifications would be necessary to guarantee the functionality of the model on low-rank RNNs in general.

Modification with white-noise evoked activity for black box RNNs

Considering the difficult accessibility of whole cortical interactions during the experimental environment, we try to modify the modeling of the feedforward recurrent alignment hypothesis with white-noise-evoked activity as feedforward input to align with the recurrent network. It turns out, that the experimental results can be roughly reflected. Besides, we test out the influence of repeatedly applying prior response activity as input on feedforward recurrent alignment. The dynamic of alignment indicates that the reuse of response as feedforward input to the recurrent network can the response to a stable oscillation.

Therefore, white-noise-evoked activity can be a good candidate for the design of experimental methods to verify the feedforward recurrent alignment hypothesis. The underlying mechanism for stable oscillation when reusing prior response activity still needs to be analytically and experimentally explored. It can give new insight into the feedforward recurrent alignment hypothesis and more understanding of feedforward recurrent network structure.

Modeling under dynamic feedforward network with Hebbian learning

Since the development of the ferrets' visual cortex was an experience-driven development, plasticity, and learning mechanisms cannot be ignored. For simplicity, start modeling the feedforward recurrent alignment with a dynamic feedforward network. The classic Hebbian learning rule was considered. The feedforward recurrent alignment increases over time until the steady state.

This indicates that during the learning, the feedforward network adapts itself to unsupervised learning such that its outputs align better with the recurrent network. According to the feedforward recurrent alignment hypothesis, this could imply that the responses become robust during learning. Therefore, modeling with plasticity and learning rules can be a promising direction to explore the underlying mechanism for experimental observations of ferrets' visual cortex.

5 Outlook

Possible constructions of feedforward recurrent alignment that does not be influenced by complex eigenvectors

In the case of general asymmetric RNNs, the modifications of the original model emerge from the problem that the alignment is in the complex plane. The modifications considered in this work focus on turning the aligned feedforward input from the complex plane into the real-number plane, while possibly keeping the original feedforward recurrent alignment definition from [TWFK23].

Another possible way for adapting the modeling of the feedforward recurrent alignment hypothesis could be a new construction of alignment score. The new construction could even allow the feedforward recurrent input to align in the complex plane resulting in the final score being a real number. For example, in a form similar to

$$\nu := (Jh)^*(Jh) = h^*J^*Jh = h^*JJh, \tag{5.1}$$

where h is the feedforward input and J the recurrent interaction matrix. The operator $*$ denotes the conjugate transformed of vectors or matrices. The last equation is followed by the fact that interactions between neurons are represented by real numbers.

With the new construction eq.(5.1), for both symmetric RNNs and asymmetric RNNs, the feedforward recurrent alignment is a real number when feedforward inputs are aligned to eigenvectors of the interaction matrix. However, the monotony between eigenvalues and alignment score cannot be kept. Although this construction is not a candidate, it could provide some ideas for new construction of feedforward recurrent alignment to overcome the problems and bring new insights to the feedforward recurrent alignment hypothesis in further studies.

Learning in recurrent network and learning rules with normalization

The simple tryout with only basic Hebbian learning in feedforward recurrent delivers the results of increased feedforward recurrent alignment during unsupervised learning. A next possible step to verify the feedforward recurrent hypothesis combined with plasticity could be also embedding learning in a recurrent network. This could help to understand the influence of recurrent dynamics on alignment and bring a closer understanding of how feedforward and recurrent networks cooperate to generate reliable responses.

Besides, the basic Hebbian rule is unstable and allows unbounded growth of feedforward network interaction strengths [DA05]. Therefore, further studies can consider normalization rules to limit the growth of network interactions, especially when considering more multiple input neurons. Moreover, other rules developed

based on Hebbian rules can also be implemented to explore the performance of different rules in the model.

Interaction structure considering excitatory and inhibitory neuron populations

A further important perspective that influences the activity dynamics and feedforward recurrent alignment is neuron population. Neurons are typically classified as either excitatory or inhibitory, meaning that they have either excitatory or inhibitory effects on all of their postsynaptic targets [DA05]. Excitatory-inhibitory networks with local heterogeneity were recently found to raise long-range neuron correlations and global organizations [DLD$^+$22]. Moreover, a recent work [EYG22] proposed that synapse-type-specific plasticity enables the joint development of stimulus selectivity and excitatory-inhibitory balance. Their model allows even implementation combining excitatory-inhibitory networks and Hebbian learning. Expanding the feedforward recurrent alignment hypothesis with excitatory and inhibitory neuron populations can be a potential focus for further studies.

Symbols and Modeling Values

List of Important Notations

$a \in \mathbb{R}$	degree of symmetry in construction of asymmetric RNNs in eq.(2.40)
$\beta_s \in \mathbb{R}$	trial-to-trial correlation defined in rq.(2.19)
$\beta \in \mathbb{R}$	parameter that defines the dimensionality in construction of covariance matrix for input. In varies contexts, there are β_{dim}, β_{spont}, and β_{Low} with $\beta_{\text{Low}} < \beta_{\text{dim}} < \beta_{\text{spont}}$.
$\bar{c}(\Delta\tilde{t}) \in \mathbb{R}$	intra-trial stability defined in eq.(2.25)
$d_{\text{eff}} \in \mathbb{R}$	the linear effective dimensionality defined in eq.(2.30). $d_{\text{eff, ana}}$ is the analytical formulation and $d_{\text{eff, emp}}$ the empirical for effective dimensionality
$\gamma \in \mathbb{R}$	the alignment of evoked activity to spontaneous activity
$h \in \mathbb{R}^{n\times 1}$	mean firing rates feedforward inputs to recurrent network. If not otherwise defined in contexts, characterizing generally the feedforward inputs.
$\tilde{h} \in \mathbb{R}^{n\times 1}$	feedforward inputs with modifications mentioned in section 2.2.2
$I_n \in \mathbb{R}^{n\times n}$	identity matrix
$J \in \mathbb{R}^{n\times n}$	interaction matrix for recurrent network
$n \in \mathbb{R}$	number of neurons in recurrent network
$N \in \mathbb{R}$	number of trials
$r \in \mathbb{R}^{n\times 1}$	response from recurrent network
$r^* \in \mathbb{R}^{n\times 1}$	steady state response from recurrent network determined by (2.6)
$R \in \mathbb{R}$	radius for eigenvalue distribution
$\sigma_{\text{trial}} \in \mathbb{R}$	variance constant for trial-to-trial correlation
$\sigma_{\text{time}} \in \mathbb{R}$	variance constant for intra-trial stability
$\Sigma \in \mathbb{R}^{n\times n}$	covariance matrix for input patterns. There are Σ^{Dim}, Σ_{spont}, and Σ_{Low} for different contexts.
$\nu \in \mathbb{R}$	feedforward recurrent alignment score
$\nu^* \in \mathbb{R}^{n\times 1}$	steady state for feedforward recurrent interaction during Hebbian learning of feedforward interaction
$W \in \mathbb{R}^{n\times 1}$	feedforward interaction matrix (vector) for feedforward recurrent network in section 2.5
$0_v \in \mathbb{R}^{n\times 1}$	zero vector in length of number of neurons n
$\rho \in \mathbb{R}$	projection ratio

Modeling Values

Notations	Values
n	200
R	0.85
τ_r	1
σ_{trial}	0.05
σ_{time}	0.3
M_{dim}	50
M_{spont}	100
Δt	0.1
$\Delta \tilde{t}$	20
T	120
N	500
β_{Low}	5
κ	5
β_{dim}	10
β_{spont}	20
M_{Low}	25
T_{Hebb}	50

List of Figures

Acknowledgements

I would like to express my sincere gratitude to all those who have contributed to the completion of this master's thesis.

First and foremost, I am deeply indebted to Professor Matthias Kaschube for delivering an inspiring lecture that ignited my interest in the field of theoretical neuroscience. Your guidance and encouragement have been invaluable throughout this journey.

I extend my thanks to Sigrid Trägenap for her invaluable assistance in brainstorming ideas and collaborating on project updates. Your insights and dedication have significantly enriched this research.

To the entire group, I am immensely grateful for creating a supportive environment that fostered growth and learning during the thesis period. Your constructive feedback and meticulous proofreading have significantly enhanced the quality of my work.

I am indebted to the Frankfurt Institute for Advanced Studies for providing me with a conducive working environment, enabling me to focus and excel in my research endeavors.

Moreover, I extend my sincere appreciation to the Deutschlandstipendium for the financial support that has made pursuing my academic goals possible.

To my cherished family and friends, your steadfast support and encouragement have been my guiding light on this journey. Your unwavering belief in me has provided continuous strength and motivation.

A heartfelt acknowledgment also goes to Tolga Tel for his meticulous proofreading, insightful feedback, and unwavering emotional support. Your presence beside me has been a consistent source of strength and inspiration.

Lastly, I am deeply grateful to every individual who has played a part in this journey, contributing to its success in their unique way. Thank you all for being by my side and making this achievement possible.

Supplementary material

Python code to reproduce the key results of this master's thesis is publicly available on GitHub under `https://github.com/rzhou-space/ffrec_alignment_hypothesis_masterthesis`

References

[Abb94] LF Abbott. Decoding neuronal firing and modelling neural networks. *Quarterly reviews of biophysics*, 27(3):291–331, 1994.

[AC14] James B Ackman and Michael C Crair. Role of emergent neural activity in visual map development. *Current opinion in neurobiology*, 24:166–175, 2014.

[AG18] Lilach Avitan and Geoffrey J Goodhill. Code under construction: neural coding over development. *Trends in Neurosciences*, 41(9):599–609, 2018.

[Bar75] H Bo Barlow. Visual experience and cortical development. *Nature*, 258(5532):199–204, 1975.

[BBKK21] David Badre, Apoorva Bhandari, Haley Keglovits, and Atsushi Kikumoto. The dimensionality of neural representations for control. *Current Opinion in Behavioral Sciences*, 38:20–28, 2021.

[BDV+21] Manuel Beiran, Alexis Dubreuil, Adrian Valente, Francesca Mastrogiuseppe, and Srdjan Ostojic. Shaping dynamics with multiple populations in low-rank recurrent networks. *Neural Computation*, 33(6):1572–1615, 2021.

[BMS+23] Manuel Beiran, Nicolas Meirhaeghe, Hansem Sohn, Mehrdad Jazayeri, and Srdjan Ostojic. Parametric control of flexible timing through low-dimensional neural manifolds. *Neuron*, 111(5):739–753, 2023.

[BSMA20] Ramon Bartolo, Richard C Saunders, Andrew R Mitz, and Bruno B Averbeck. Dimensionality, information and learning in prefrontal cortex. *PLoS computational biology*, 16(4):e1007514, 2020.

[BYBOS95] Rani Ben-Yishai, R Lev Bar-Or, and Haim Sompolinsky. Theory of orientation tuning in visual cortex. *Proceedings of the National Academy of Sciences*, 92(9):3844–3848, 1995.

[CGS98] Michael C Crair, Deda C Gillespie, and Michael P Stryker. The role of visual experience in the development of columns in cat visual cortex. *Science*, 279(5350):566–570, 1998.

[CMVH17] Ian K Christie, Paul Miller, and Stephen D Van Hooser. Cortical amplification models of experience-dependent development of selective columns and response sparsification. *Journal of neurophysiology*, 118(2):874–893, 2017.

[CSB96] Barbara Chapman, Michael P Stryker, and Tobias Bonhoeffer. Development of orientation preference maps in ferret primary visual cortex. *Journal of Neuroscience*, 16(20):6443–6453, 1996.

[CSFW17] Warasinee Chaisangmongkon, Sruthi K Swaminathan, David J Freedman, and Xiao-Jing Wang. Computing by robust transience: how the fronto-parietal network performs sequential, category-based decisions. *Neuron*, 93(6):1504–1517, 2017.

[CW01] Chiayu Chiu and Michael Weliky. Spontaneous activity in developing ferret visual cortex in vivo. *Journal of Neuroscience*, 21(22):8906–8914, 2001.

[CWF20] Jeremy T Chang, David Whitney, and David Fitzpatrick. Experience-dependent reorganization drives development of a binocularly unified cortical representation of orientation. *Neuron*, 107(2):338–350, 2020.

[DA05] Peter Dayan and Laurence F Abbott. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT press, 2005.

[DKM+95] Rodney J Douglas, Christof Koch, Misha Mahowald, Kevan AC Martin, and Humbert H Suarez. Recurrent excitation in neocortical circuits. *Science*, 269(5226):981–985, 1995.

[DLD+22] David Dahmen, Moritz Layer, Lukas Deutz, Paulina Anna Dąbrowska, Nicole Voges, Michael von Papen, Thomas Brochier, Alexa Riehle, Markus Diesmann, Sonja Grün, et al. Global organization of neuronal activity only requires unstructured local connectivity. *Elife*, 11:e68422, 2022.

[DVB+22] Alexis Dubreuil, Adrian Valente, Manuel Beiran, Francesca Mastrogiuseppe, and Srdjan Ostojic. The role of population structure in computations through neural dynamics. *Nature neuroscience*, 25(6):783–794, 2022.

[ES12] J Sebastian Espinosa and Michael P Stryker. Development and plasticity of the primary visual cortex. *Neuron*, 75(2):230–249, 2012.

[EYG22] Samuel Eckmann, Edward James Young, and Julijana Gjorgjieva. Synapse-type-specific competitive hebbian learning forms functional recurrent networks. *bioRxiv*, pages 2022–03, 2022.

[FI84] Yves Frégnac and Michel Imbert. Development of neuronal selectivity in primary visual cortex of cat. *Physiological reviews*, 64(1):325–434, 1984.

[FO10] David A Feldheim and Dennis DM O'Leary. Visual map development: bidirectional signaling, bifunctional guidance molecules, and competition. *Cold Spring Harbor perspectives in biology*, 2(11):a001768, 2010.

[FWS+96] Marla B Feller, David P Wellis, David Stellwagen, Frank S Werblin, and Carla J Shatz. Requirement for cholinergic synaptic transmission in the propagation of spontaneous retinal waves. *Science*, 272(5265):1182–1187, 1996.

[GG15] Peiran Gao and Surya Ganguli. On simplicity and complexity in the brave new world of large-scale neuroscience. *Current opinion in neurobiology*, 32:148–155, 2015.

[GKBS97] Lmke Gödecke, Dae-Shik Kim, Tobias Bonhoeffer, and Wolf Singer. Development of orientation preference maps in area 18 of kitten visual cortex. *European Journal of Neuroscience*, 9(8):1754–1762, 1997.

[Goo16] Geoffrey J Goodhill. Can molecular gradients wire the brain? *Trends in neurosciences*, 39(4):202–211, 2016.

[GPN+18] Juan A Gallego, Matthew G Perich, Stephanie N Naufel, Christian Ethier, Sara A Solla, and Lee E Miller. Cortical population activity within a preserved neural manifold underlies multiple motor behaviors. *Nature communications*, 9(1):4233, 2018.

[HFC08] Andrew D Huberman, Marla B Feller, and Barbara Chapman. Mechanisms underlying development of visual maps and receptive fields. *Annu. Rev. Neurosci.*, 31:479–509, 2008.

[HM] G. Smith H. Mulholland, M. Kaschube. Mechanisms underlying the self-organization of patterned activity in the developing visual cortex.

[HM23] M. Kaschube G.B. Smith H.N. Mulholland, S. Trägenap. Selective amplification of recurrent subnetworks in the developing visual cortex. *Neuroscience 2023, https://www.abstractsonline.com/pp8/!/10892/presentation/29431*, (PSTR149.11.), 2023.

[HMF13] Kenneth D Harris and Thomas D Mrsic-Flogel. Cortical connectivity and sensory coding. *Nature*, 503(7474):51–58, 2013.

[ILMR18] K Imaizumi, CC Lee, JN MacLean, and ES Ruthazer. Spontaneous activity in the sensory system. lausanne: Frontiers media. doi: 10.3389. 2018.

[Kar] Mehran Kardar. The mammalian visual system. *https://www.mit.edu/ kardar/research/seminars/CorticalMaps/VisualSystem.html*.

[Mil16] Kenneth D Miller. Canonical computations of cerebral cortex. *Current opinion in neurobiology*, 37:75–84, 2016.

[MO18] Francesca Mastrogiuseppe and Srdjan Ostojic. Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. *Neuron*, 99(3):609–623, 2018.

[Mon97] G Montgomery. Seeing, hearing and smelling the world. 1997.

[MRB10] Christian K Machens, Ranulfo Romo, and Carlos D Brody. Functional, but not anatomical, separation of "what" and "when" in prefrontal cortex. *Journal of Neuroscience*, 30(1):350–360, 2010.

[MSSN13] Valerio Mante, David Sussillo, Krishna V Shenoy, and William T Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 503(7474):78–84, 2013.

[MWBS91] Markus Meister, Rachel OL Wong, Denis A Baylor, and Carla J Shatz. Synchronous bursts of action potentials in ganglion cells of the developing mammalian retina. *Science*, 252(5008):939–943, 1991.

[MYDF09] Olivier Marre, Pierre Yger, Andrew P Davison, and Yves Frégnac. Reliable recall of spontaneous activity patterns in cortical networks. *Journal of neuroscience*, 29(46):14596–14606, 2009.

[Oh04] Se Jong Oh. *Learning to segment texture in 2D vs. 3D: A comparative study*. PhD thesis, Texas A&M University, 2004.

[PPV+20] Simon Peron, Ravi Pancholi, Bettina Voelcker, Jason D Wittenbach, H Freyja Ólafsdóttir, Jeremy Freeman, and Karel Svoboda. Recurrent interactions in local cortical circuits. *Nature*, 579(7798):256–259, 2020.

[PRFS98] Anna A Penn, Patricio A Riquelme, Marla B Feller, and Carla J Shatz. Competition in retinogeniculate patterning driven by spontaneous activity. *Science*, 279(5359):2108–2112, 1998.

[RA06] Kanaka Rajan and Larry F Abbott. Eigenvalue spectra of random matrices for neural networks. *Physical review letters*, 97(18):188104, 2006.

[RBW+13] Mattia Rigotti, Omri Barak, Melissa R Warden, Xiao-Jing Wang, Nathaniel D Daw, Earl K Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590, 2013.

[RM21] Stephen J Read and Lynn Carol Miller. Neural network models of personality structure and dynamics. In *Measuring and Modeling Persons and Situations*, pages 499–538. Elsevier, 2021.

[SGS99] Kerstin E Schmidt, Ralf AW Galuske, and Wolf Singer. Matching the modules: Cortical maps and long-range intrinsic connections in visual cortex during development. *Journal of neurobiology*, 41(1):10–17, 1999.

[SHW+18] Gordon B Smith, Bettina Hein, David E Whitney, David Fitzpatrick, and Matthias Kaschube. Distributed network interactions and their emergence in developing neocortex. *Nature neuroscience*, 21(11):1600–1608, 2018.

[SNMJ19] Hansem Sohn, Devika Narain, Nicolas Meirhaeghe, and Mehrdad Jazayeri. Bayesian computation through cortical latent dynamics. *Neuron*, 103(5):934–947, 2019.

[Soc19] et al. Soch, Joram. Statproofbook/statproofbook.github.io: Statproofbook 2022 (version 2022). *Proof: Linear transformation theorem for the multivariate normal distribution /Zenodo. https://doi.org/10.5281/ZENODO.4305949*, 2019.

[TWFK23] Sigrid Trägenap, David E Whitney, David Fitzpatrick, and Matthias Kaschube. The nature-nurture transform underlying the emergence of reliable cortical representations. *bioRxiv*, (2022.11. 14.516507), 2023.

[WF07] Leonard E White and David Fitzpatrick. Vision and cortical map development. *Neuron*, 56(2):327–338, 2007.

[WNHJ18] Jing Wang, Devika Narain, Eghbal A Hosseini, and Mehrdad Jazayeri. Flexible timing by temporal scaling of cortical responses. *Nature neuroscience*, 21(1):102–110, 2018.