Linking Neural Manifolds to Principles of Circuit Structure

Louis Pezon[1*], Valentin Schmutz[1,2†] and Wulfram Gerstner[1†]

[1*]School of Computer and Communication Sciences and School of Life Sciences, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland.

[2]UCL Queen Square Institute of Neurology, University College London WC1E 6BT London, United Kingdom.

*Corresponding author(s). E-mail(s): louis.pezon@epfl.ch; Contributing authors: v.schmutz@ucl.ac.uk; wulfram.gerstner@epfl.ch; †These authors contributed equally to this work.

Abstract

While large-scale recording techniques indicate that the activity of heterogeneous neuronal populations lies on low-dimensional "neural manifolds", it has remained challenging to reconcile this picture with the classical view of precisely tuned neurons interacting with each other in some ordered circuit structure. Using a modelling approach, we provide a conceptual, yet mathematically precise, link between these two contrasting views. We first show that there is no unique relationship between the circuit structure and the emergent low-dimensional dynamics that characterise the population activity. We then propose a method for retrieving the circuit structure from recordings of the population activity and test it on artificial data. Our approach provides not only a unifying framework for circuit and field models on one side, and low-rank networks on the other side, but also opens the perspective to identify principles of circuit structure from large-scale recordings.

Keywords: Neural activity vector, neural manifold, neural field model, computational neuroscience, neuronal dynamics, neural circuit structure

Introduction

Our understanding of the link between the brain activity and behaviour is continuously reshaped by the advance of experimental methods. While for a long time neural recordings were limited to one or a few neurons at a time [1], the availability of recent tools such as high-density probes and two-photon calcium imaging has enabled the simultaneous recording of thousands of neurons [2]. As a consequence, modern neuroscience offers two different coexisting, but not yet fully integrated, perspectives [3, 4].

The first perspective focuses on the representation and processing of information by single neurons embedded and talking to each other in a precise circuit structure. Responses of single neurons depend on "features" such as their physical location on the cortical sheet [5], or the tuning of their receptive field to aspects of sensory stimuli [1, 6, 7], spatial position and orientation in an environment [8–12], or task-related variables [13–15]. In this perspective, neural computation arises from the dynamic interaction of many precisely tuned neurons within recurrent brain circuits that are wired to combine and extract information from different neurons, as exemplified by numerous models [16–23]. For example, in a model of spatial orientation, the sense of orientation is the read-out from a population of model head-direction cells where each cell is characterised by its preferred head direction [18], leading to a one-dimensional circuit structure: see Fig. 1A. Importantly, the physical location of the neurons in the brain is irrelevant to this notion of dimensionality; neurons are assigned locations in an abstract space where neighbourhood is defined by tuning similarities, called the *tuning space* (Fig.1C). For example, in a model of early visual cortex where each neuron is described by the (x, y)-location and orientation of its receptive field [24, 25], the tuning space would be three-dimensional. In particular, patchy connectivity indicates that neurons that are not physical neighbours but have similar orientation preferences are "topological" neighbours with higher connection probability than other neurons at a similar distance. The idea that tuning similarities organize connectivity patterns as well as the population activity is captured by *field models* [26], i.e., in the limit of large networks the neuronal activity can be described as a function of the neurons' location in the continuous tuning space [27, 28].

The other perspective, pushed forward by the analysis of large-scale recordings, starts from a descriptive, population-level approach [29]. Several experimental studies revealed that the activity of large networks is sometimes restricted – or attracted – to low-dimensional manifolds

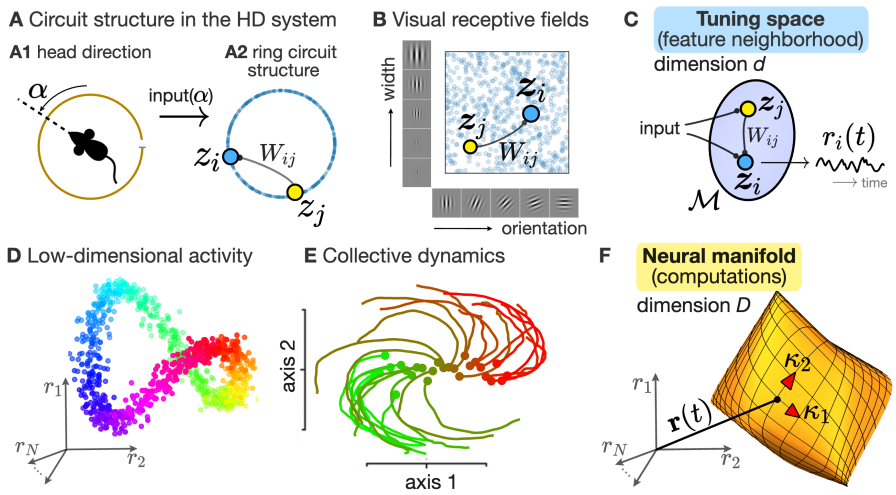

Fig. 1 *Two perspectives on neuronal activity.* **A.** The "circuit" perspective is illustrated in a ring model of the head direction (HD) system. **A1.** The animal's HD is represented by an angle α. **A2.** Each neuron in the population responds preferentially to a specific HD angle represented as an abstract "position" (blue dots) on a one-dimensional ring. The strength of recurrent connections W_{ij} between pairs of neurons i and j, as well as their response to the momentary HD angle α, are determined by their respective positions z_i and z_j. **B.** Similarly, in a model of early visual cortex, each neuron i can be characterised by the features of its receptive field (RF), here orientation and width, which determine the neuron's position $\boldsymbol{z}_i$ in a 2-dimensional abstract space. **C.** A specific circuit structure is characterised by the "positions" of the neurons in an abstract *tuning space* of dimension d, where neighbourhood represents tuning similarity. In circuit models, recurrent connections between pairs of neurons, as well as the external input they receive, are determined by the position $\boldsymbol{z}_i$ of each neuron i in the tuning space, which in turn determines its activity $r_i(t)$. **D.** The "neural manifold" perspective is illustrated by large-scale recordings of the neuronal activity in the rodent HD system. The time-dependent activity of a population of N cells representing the head direction angle, as illustrated in panel (A), traces out a one-dimensional ring-like manifold within the N-dimensional space, matching the topology of the represented variable [32]. Each point corresponds to a population activity vector, coloured according to the HD angle α it represents. **E.** In primate motor cortex, the linear projection of the population activity onto relevant axes reveals rotational population-level trajectories during reaching [33]. For each trial, the low-dimensional trajectory (solid lines) is mostly determined by its initial state (circles, coloured according to their projection on axis 1). **F.** The joint activity of N neurons, $\boldsymbol{r}(t) \in \mathbb{R}^N$, lies on a low-dimensional *neural manifold* with intrinsic dimension $D \ll N$ (in this schematic, a curved, $(D = 2)$-dimensional surface). The activity is parametrised by population-level latent variables $\kappa_1(t), \ldots, \kappa_D(t)$ whose dynamics reflect neural computations.
Panel (D) is inspired from [32]. Panel (E) is inspired from [33].

embedded within the state-space of the joint activity of all neurons, often referred to as the *neural manifold* [30–33]. The neuronal activity, as a point on this manifold, can thus be parameterised by a set of time-dependent *latent variables*, which encode computationally relevant quantities at the population level (Fig.1D,E,F). In this view, neural computation is described by the collective dynamics of thousands of neurons, driven by the flow of these latent variables on the neural manifold [34–39].

To summarise, the "circuit structure" perspective tries to grasp the mechanisms underlying neuronal activity at the level of individual neurons, whereas the "neural manifold" perspective focuses on the description of representations and computations at the population level [3]. Reconciling the two perspectives could help to combine the insights of mechanistic models with the statistical and descriptive power of population-level approaches [4]. Yet, a transparent, but mathematically precise, theoretical framework that connects the two perspectives is missing. This paper aims to provide the missing link. More precisely, we formulate the following questions. (i) Can the same collective dynamics result from different circuit structures? Even though the topology of the *tuning space* and the *neural manifold* has been found to match in specific cases [4, 11, 32] (for example, a ring in the head direction system, see Fig.1A,D), there is no evidence that it should be the case in general. (ii) Can we extract knowledge about the circuit structure from the neuronal activity? While many studies have focused on extracting the neural manifold from experimental data [31, 32], an established method for extracting the tuning space is missing.

In this paper, we propose a theoretical framework that allows us to address these questions in models of recurrent spiking neural networks. Before we start, let us introduce some terminology. The dimension of the tuning space introduced above will be denoted by (lower-case) d. In reference to neural field theories, d is also called the *field dimension* and we use this term occasionally. For example, ring models of head-direction [18] or orientation tuning of visual receptive fields [17] have $d = 1$; whereas a model of rodent place cells where each neuron is characterised by its place cell center in the two-dimensional environment [40] has field dimension $d = 2$. The dimension of the neural manifold is denoted by (capital) D and refers to the minimal number of variables necessary to describe the population activity [41], determined by linear [33, 35, 42] or nonlinear [31, 32, 43, 44] dimensionality reduction methods. Moreover, the collective dynamics in the neural manifold might generate a set of attractor states with dimension $D_A \leq D$. For example, a classic attractor state in the Hopfield model of associative memory [45] is a point with $D_A = 0$ whereas a line attractor in the ring model has $D_A = 1$. With this terminology, we can rephrase questions (i) and (ii) as the search for the missing link between these different notions of dimensionality.

On the modelling side, the perspective of "neural manifolds" has been illustrated by models of recurrent networks where the matrix of recurrent

connectivity in a network with N neurons has rank $D \ll N$, which generates a D-dimensional neural manifold [46, 47]. Such low-rank models enable to implement a wide range of computations – from pattern storage as fixed points [45] to approximating arbitrary dynamical systems [39, 47–49]. However, despite their success as versatile and interpretable computational models, the link of low-rank models to classical circuit models where the connectivity is structured by the tuning properties of individual neurons remains elusive. This problem has been identified by Langdon *et al.* [4]: "it is unclear whether low-rank RNNs utilize mechanisms similar to classical circuit models or implement truly novel solutions [...]". In order to answer the two questions posed above, we will in passing address this problem.

Results

Network model.

We study a model of a recurrent spiking neural network (SNN) composed of N Poisson neurons [50] (also called nonlinear Hawkes processes [28, 51]). In the model, each spike of a presynaptic neuron j, connected to a postsynaptic neuron i with a synapse strength W_{ij}, causes a stepwise increase or decrease of the postsynaptic membrane potential h_i. In the absence of synaptic input, the potential decays with the membrane time constant τ to the resting potential (set to zero), following the dynamics:

$$\frac{d}{dt}h_i(t) = -\frac{1}{\tau}h_i(t) + \frac{1}{N}\sum_{\substack{j=1\\ j\neq i}}^{N} W_{ij}S_j(t) + \sum_{k=1}^{K} w_{ik}^{\text{ext}} I_k(t) \tag{1}$$

where w_{ik}^{ext} denotes the input weights to neuron i corresponding to K external inputs $I_1, \ldots, I_K$; and $S_j(t)$ denotes the spike train of neuron j. A neuron with depolarization $h_i(t)$ will emit spikes with an instantaneous rate $\phi(h_i(t))$. For the sake of simplicity, we neglect refractoriness here; but to some extent, our framework can be adapted to more general neuron models, as examined in the Discussion.

As mentioned in the introduction, experimentalists have traditionally characterized neurons by their tuning properties to features of sensory input [1, 6, 7] or abstract task variables [8–11, 13–15, 40]. The d features used to characterise each neuron i can be represented as its *position*, denoted $\boldsymbol{z}_i$, in an abstract d-dimensional *tuning space* $\mathcal{M}$, in which

neighbourhood reflects tuning similarity. For example, the width Δ_i and orientation θ_i of a simple-cell RF in the visual input space are represented by the $(d = 2)$-dimensional "position" $\boldsymbol{z}_i = (\Delta_i, \theta_i)$, which lies in the range $\Delta_i > 0$ for the width, and $0 \leq \theta_i < \pi$ with periodic boundary condition for the orientation (Fig.1B); i.e., a tuning space with the topology of a cylinder. The position $\boldsymbol{z}_i = (z_{i,1}, z_{i,2}, ..., z_{i,d})$ is the vector containing the d tuning characteristics of the neuron.

With this in mind, we will write the recurrent connection W_{ij} from neuron j to neuron i in Eq.(1) as:

$$W_{ij} = \sum_{\mu=1}^{D} F_{i,\mu} G_{j,\mu} = \sum_{\mu=1}^{D} f_\mu(\boldsymbol{z}_i)\, g_\mu(\boldsymbol{z}_j) \tag{2}$$

where $\boldsymbol{f} = (f_1, \ldots, f_D)$ and $\boldsymbol{g} = (g_1, \ldots, g_D)$ are continuous functions that describe how a presynaptic neuron at the abstract position $\boldsymbol{z}_j$ is connected to a postsynaptic neuron at position $\boldsymbol{z}_i$. Similarly, external input weights from the external source I_k to neuron i will be written as $w_{ik}^{\text{ext}} = f_k^{\text{ext}}(\boldsymbol{z}_i)$.

We emphasize four aspects. First, the weight matrix in Eq. (2) has the well-known outer product structure [45, 46, 52, 53] that generates a matrix of rank $D < N$, analogous to the Hopfield model [45, 52] and other models of low-rank networks [46, 47, 53, 54], known to generate D-dimensional dynamics parameterised by latent variables $\boldsymbol{\kappa}(t) = (\kappa_1(t), \ldots, \kappa_D(t))$ (analogous to the overlap variables in the Hopfield model). The activity follows the flow defined by the collective dynamics of the latent variables in a neural manifold of dimension D. The closed-form expression of these collective dynamics, given in the Methods (Eq.(16)), is the central ingredient of our theory, for it links the circuit structure to the flow of the population activity in the neural manifold.

Second, the weight matrix in Eq. (2) can also be interpreted as a D-dimensional inner product (scalar product) $\sum_{\mu=1}^{D} F_{i,\mu} G_{j,\mu} = \vec{F}_i \cdot \vec{G}_j$ (see Fig.2). Many known models of low-rank networks are defined by a specific distribution of the two vectors $(\vec{F}_i, \vec{G}_i)$ associated with each neuron i across the population [45–47, 52–56]. In this respect, Eq.(2) is a general formulation of low-rank connectivity, where specific models correspond to a specific distribution of the positions $\boldsymbol{z}_i$ and choice of the functions $\boldsymbol{f}, \boldsymbol{g}$. For example, a Hopfield model with P binary patterns [50, 52] corresponds to positions $\boldsymbol{z}_i$ uniformly distributed on the edges $\{\pm 1\}^P$ of the P-dimensional hypercube.

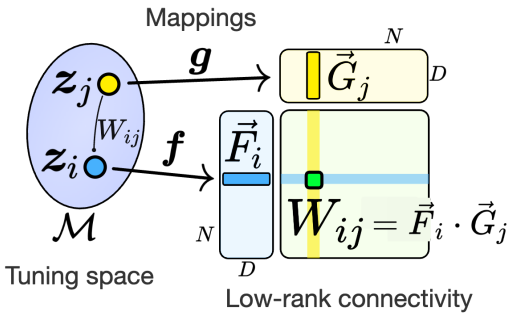

Fig. 2 *Linking the circuit structure to the low-rank connectivity.* Neurons with positions $\boldsymbol{z}_i, \boldsymbol{z}_j$ in the tuning space $\mathcal{M}$ (left) are mapped by two functions $\boldsymbol{f}, \boldsymbol{g} : \mathcal{M} \to \mathbb{R}^D$ to D-dimensional vectors $\vec{F}_i = \boldsymbol{f}(\boldsymbol{z}_i)$ and $\vec{G}_j = \boldsymbol{g}(\boldsymbol{z}_j)$. Each entry W_{ij} of the rank-D connectivity matrix (right) is given by the scalar product $\vec{F}_i \cdot \vec{G}_j$.

Third, the weight matrix in Eq. (2) could be the result of Hebbian learning [52, 53]. Indeed, we can interpret each function $f_\mu(\boldsymbol{z}_i)$ as the probability that neuron i with tuning properties $\boldsymbol{z}_i$ forms, in the experimental situation μ, a dendritic spine; and $g_\mu(\boldsymbol{z}_j)$ as the probability that neuron j forms a presynaptic bouton. The product $f_\mu(\boldsymbol{z}_i)\, g_\mu(\boldsymbol{z}_j)$ gives then the total probability of forming a connection (or increasing an existing one) in the experimental situation μ. Across several experimental situations, the weight matrix in Eq.(2) will build up. In a memory-related brain area, the "experimental situation" could correspond to the presentation of a novel memory item that was never encountered before; whereas in early visual cortex the experimental situation could correspond to elongated spatial edges observed during the critical period.

Finally, the connectivity of Eq.(2) corresponds to a connectivity kernel over the tuning space, given by $w(\boldsymbol{z}_i, \boldsymbol{z}_j) = \boldsymbol{f}(\boldsymbol{z}_i) \cdot \boldsymbol{g}(\boldsymbol{z}_j)$. In the limit of a large network ($N \to \infty$), the above model corresponds to a neural field over the tuning space [54], whose dimension d is not necessarily related to the rank D of the connectivity, i.e. to the dimension of the neural manifold. This is true in general for circuit models where the synaptic weights scale as $1/N$ when the number N of neurons increases [28]. For that reason, we sometimes use the terms "circuit model" and "field model" interchangeably in the rest of the paper.

Tuning space and neural manifold can mismatch: examples

In the following, we illustrate the main insights provided by the above model in a few examples. For the sake of simplicity, we assume in the following examples that the network does not receive any external input;

and that the instantaneous firing rate is a step function of the membrane potential h, with $\phi(h) = R$ for $h > 0$ and zero otherwise. Qualitatively results do not change for other choices (Appendix B).

A one-dimensional circuit structure generates a two-dimensional neural manifold

As a first example, let the tuning space be the real line, over which the neurons' positions $z_i \in \mathbb{R}$ are normally distributed (Fig. 3A1). If the interactions are given by $W_{ij} = J\, z_i z_j$, this corresponds to a Hopfield network with a single Gaussian pattern. In this case, the latent dynamics have a single stable non-trivial fixed point, up to a symmetry degeneracy, and analogously for the case of P patterns [55]. We study a variant of this model where synaptic weights are given by:

$$W_{ij} = \begin{cases} J\, z_i z_j & (z_i z_j \geq 0) \\ \cos(\alpha) J\, z_i z_j & (z_i z_j < 0) \end{cases} \tag{3}$$

with $0 < \alpha < \pi/2$. The weights can be written as a 2-dimensional scalar product $W_{ij} = \boldsymbol{f}(z_i) \cdot \boldsymbol{g}(z_j)$, where $\boldsymbol{g} = J\boldsymbol{f}$, and the images $\boldsymbol{f}(z_i)$ of the neuron's positions are arranged along a line in $\mathbb{R}^2$, bent at the origin by an angle α (Fig.3A2; more details in the Methods). In agreement with the theory, a SNN with N Poisson neurons (Eq.(1)) and weights according to Eq. 3 generates ($D = 2$)-dimensional collective dynamics that decay towards three stable fixed points spread out in the plane and one saddle point at the origin. Yet, the tuning space is the real line, meaning that the model with neural manifold of dimension $D = 2$ has a tuning space and circuit structure of dimension $d = 1$.

Two different circuit structures generate identical collective dynamics

We now show that identical population dynamics in a neural manifold of dimension $D = 2$ can be implemented by two different circuit models: for the first model the tuning space is the 1-dimensional ring, and for the second one the 2-dimensional plane. Moreover, for both models, the set of attractor states of the collective dynamics, inside the two-dimensional neural manifold, is a one-dimensional limit cycle.

The first circuit model is a *ring model*, in which neurons are uniformly assigned angular positions $z_i \in [0, 2\pi[$ (Fig.3C1) [17, 18]. We express the connectivity as:

$$W_{ij} = J \cos(z_i - z_j - \delta) \tag{4}$$

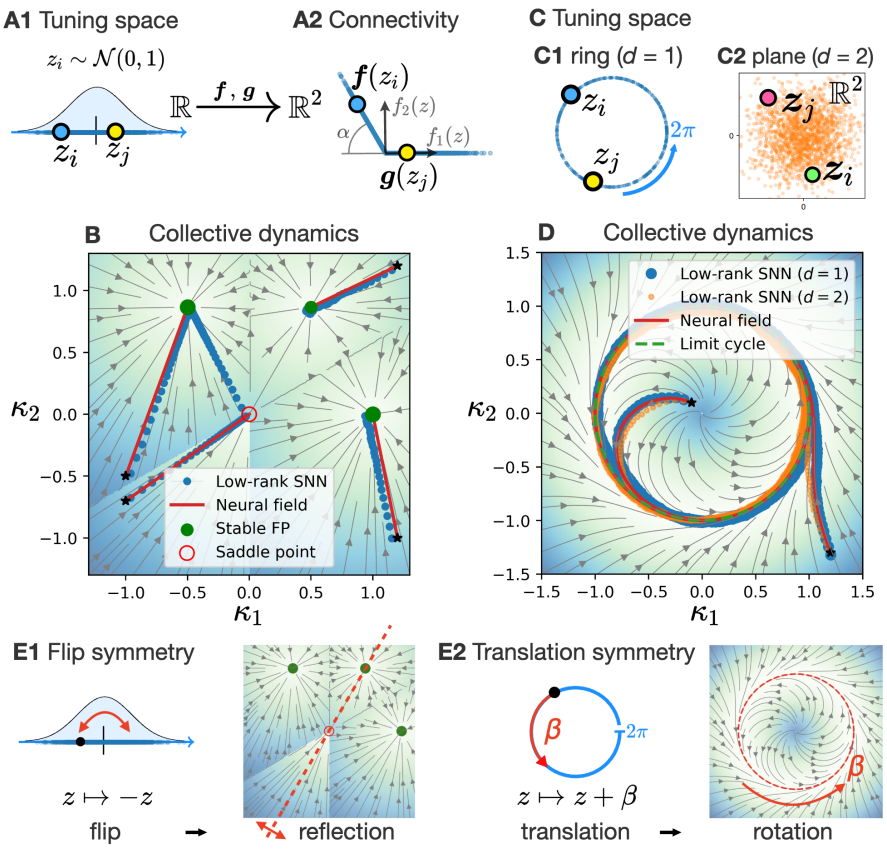

Fig. 3 *The dimension d of the tuning space and the dimension D of the neural manifold do not always match.* **A1.** Neurons are normally distributed over the real line: the dimension of the tuning space is $d = 1$. **A2.** The connectivity is determined by 2-dimensional mappings $\boldsymbol{f}$ and $\boldsymbol{g}$: the dimension of the neural manifold is $D = 2$. **B.** Flow of the collective dynamics of the circuit model in panel (A) (darker blue-shading indicates faster dynamics). The population dynamics of an SNN with $N = 2000$ neurons (blue dots, following Eqs.(1-2)) is well approximated by the collective dynamics (red lines, Eq. (16)), starting from the same initial condition (black stars). The axes are graduated in units of $k = \frac{\tau J R}{\sqrt{2\pi}}$. **C-D.** Two different circuit models can produce identical collective dynamics. **C.** Schematic of the distribution of neuron positions: **C1** uniform over the ring ($d = 1$); **C2** standard Gaussian in $\mathbb{R}^2$ ($d = 2$). **D.** Same as panel (B), for the circuit models in panel (C). The collective dynamics generate a globally attractive limit cycle. The axes are graduated in units of $\rho_0 = \frac{\tau J R}{\pi} \cos\delta$. **E** The symmetries of a circuit model result in symmetries of the collective dynamics. **E1.** The symmetry w.r.t. a simultaneous flip of all neuron positions (flip symmetry) in panel (a) causes a reflection symmetry of the collective dynamics in panel (b). **E2.** Similarly, the symmetry w.r.t. simultaneous shifts of all neurons along the ring (translation symmetry) in panel (C1) causes a rotation symmetry of the collective dynamics in panel (D).
The relevant numerical parameters are: $\tau J R = 10$, and $\alpha = \pi/3$ in (a-b) (Eq. (3)), $\delta = \pi/10$ in (c-d) (Eq. (4)). The latent trajectory of the SNN is obtained from Eq. (15): $\boldsymbol{\kappa}^{\mathrm{SNN}}(t) := \frac{1}{N}\sum_i M\boldsymbol{f}(z_i)h_i(t)$. Notice that in panel (B), the limit field stays in the saddle point at the origin, whereas the SNN leaves towards one of the attractors due to finite-size fluctuations and spike firing noise.

where $J > 0$, and $\delta \in]-\pi, \pi]$ is an arbitrary dephasing. For $\delta = 0$, we get a standard ring attractor model with stationary bumps of activity [17]; whereas for $\delta \neq 0$, the attractor states are travelling bumps [18]. Using the sum expansion of the cosine function [57], the mapping from neuronal positions z_i to the low-rank connectivity in Eq.(2) are given by $\boldsymbol{f}(z_i) = (\cos z_i, \sin z_i)$ and $\boldsymbol{g}(z_j) = J\boldsymbol{f}(z_j + \delta)$.

In the second model, the neuronal positions are normally distributed over the *real plane* $\mathbb{R}^2$ (Fig.3C2). Similarly to the ring model above, let the connections be maximal for an angle δ between two neurons:

$$W_{ij} = \tilde{J}\, \boldsymbol{z}_i R^{(\delta)} \boldsymbol{z}_j \tag{5}$$

where $R^{(\delta)}$ is the matrix of rotation by an angle δ in $\mathbb{R}^2$; and $\tilde{J} > 0$ is the interaction strength (different from that of the ring model above). The mappings from the neuronal position to the low-rank connectivity in Eq.(2) is given by $\boldsymbol{f}(\boldsymbol{z}_i) = \boldsymbol{z}_i$ and $\boldsymbol{g}(\boldsymbol{z}_j) = \tilde{J} R^{(\delta)} \boldsymbol{z}_j$.

Importantly, for $J = \sqrt{\frac{\pi}{2}}\tilde{J}$ (i.e. up to a rescaling), the two models obey *identical* two-dimensional collective dynamics (given in the Methods, Eq.(21)). The theory predicts that for any initial condition, the two latent variables that parameterise the population activity converge to a globally attractive limit cycle, which is confirmed by simulations of SNNs with Poisson neurons (Fig.3D).

We would like to highlight two aspects. (i) For both the "ring" and the "planar" model, there exists a simultaneous transformation of all neuronal positions in the tuning space that leaves the connectivity unchanged, i.e., a translation along the ring in Eq. (4), or rotation in the plane in Eq. (5). The flow of the collective dynamics – and therefore the set of its attractor states – in turn exhibits a rotation symmetry, illustrated in Fig.3D. The general link between symmetries of tuning space and neuronal manifold is established in the next section. (ii) Neither the dimension D of the neural manifold nor the flow of the collective dynamics enable us to distinguish the ring model from the planar model. How the same collective dynamics can be implemented using different circuit structures is discussed more generally in the section "Dimensionality Mismatch". Nevertheless, we present later (section "The footprint of the circuit structure") a technique that allows us to extract the tuning space from the the dynamics of the population activity.

Symmetries of the connectivity generate symmetries of the collective dynamics

A *symmetry* of a circuit model is a simultaneous transformation of all neuronal positions preserving the network dynamics. For instance, in the ring model defined by Eq.(4), the translation along the ring by an arbitrary (angular) distance β is a symmetry: intuitively, it means that the recurrent input received by each neuron is not affected upon shifting every neuron i from angular position z_i to position $z_i + \beta$. This reflects the fact that the choice of the "zero" position (or of HD angle $\alpha = 0$) on a ring is arbitrary; thus, only differences between angles, but not absolute angles, are meaningful.

More precisely, a *symmetry* of a circuit model is a transformation $S : \mathcal{M} \to \mathcal{M}$ of neuronal positions that transforms solutions to the dynamics into other solutions; i.e. if membrane potentials given by the field $h_i(t) = v(t, \boldsymbol{z}_i)$ over the tuning space solve the dynamics of Eq.(1), then $h_i(t) = v(t, S\boldsymbol{z}_i)$ is also a valid solution. In particular, any mapping S that leaves the connectivity and the distribution of the positions in the tuning space unchanged is a symmetry. We show that each such symmetry S of a circuit model induces a symmetry of the associated collective dynamics in the D-dimensional neural manifold, expressed as a linear mapping $T^{(S)}$ over $\mathbb{R}^D$ (Eq.(22) in the Methods).

For example, in the ring model above, with connectivity given by Eq.(4), each translation symmetry $S : z \mapsto z + \beta$ results in a symmetry of the collective dynamics which is given by the matrix:

$$T^{(S)} = \begin{pmatrix} \cos\beta & -\sin\beta \\ \sin\beta & \cos\beta \end{pmatrix} \tag{6}$$

which is the matrix of rotation of an angle β in $\mathbb{R}^2$ (Fig.3E2). Similarly, the earlier example of a one-dimensional circuit model with connectivity given by Eq.(3) has a "flip" symmetry $S : z \mapsto -z$. The corresponding symmetry of the collective dynamics (Fig.3E1) is the reflection w.r.t. the axis spanned by the vector $(1-\cos\alpha, \sin\alpha)$. More generally, the symmetry group of a circuit model in the tuning space is also a symmetry group of the collective dynamics in the neural manifold. More details are given in the Methods and Appendix C.

Dimensionality mismatch

While the symmetries of the connectivity establish a link between the circuit structure and the collective dynamics of a RNN, the previous example of two models generating the same limit-cycle dynamic suggests that the tuning space cannot be retrieved from the flow of the collective dynamics, and that their respective dimensions d and D might differ. Here, we discuss the potential mechanisms of this mismatch.

Heterogeneity as a mechanism of dimensionality mismatch.

The "planar" circuit model with a two-dimension tuning space (Eq.(5)) can be interpreted as an instance of the ring model that includes single-neuron heterogeneities. Since neuronal positions are normally distributed in $\mathbb{R}^2$, the polar coordinates (ξ_i, φ_i) of the position $\boldsymbol{z}_i$ of any neuron i are two independent random variables: an angle φ_i, distributed uniformly over the ring $[0, 2\pi[$, and a radius ξ_i, distributed according to the χ_2-distribution. The radius ξ_i can be interpreted as a heterogeneity affecting both the neuron's transfer function and its connectivity compared to the initial ring model (for which $\xi_i = 1$). The distribution of this heterogeneity does not affect the collective dynamics of the ring model (up to a rescaling; see Appendix B.2). Along the same line of argument, additional neuronal features that parameterise heterogeneity in the population can be incorporated into any circuit model, thus extending the dimension of the tuning space, without affecting the collective dynamics. Simple examples include the multiplication of the outgoing weights of each neuron by a random factor, modelling subsets of neurons with "weak" or depleted synapses; and heterogeneous maximum firing rates among neurons (see Appendix D).

Several universal approximators.

The possibility of a dimensionality mismatch also arises from the fact that any target collective dynamics can be equivalently implemented by several circuit models, defined over different tuning spaces. Two different "universal approximators" have been identified. On one hand, a Gaussian mixture low-rank network of rank D can approximate arbitrary collective dynamics in a D-dimensional neural manifold [47], and corresponds to a field model with tuning space dimension $d = 2D$ (Appendix E.1). On the other hand, field models with a one-dimensional tuning space can in principle also approximate arbitrary collective dynamics if the mappings

$\boldsymbol{f}, \boldsymbol{g}$ in Eq. (2) can be chosen arbitrarily with no requirement of "smoothness", analogous to the universal approximation property of feed-forward neural networks (Appendix E.2).

The footprint of the circuit structure

We conclude from the previous sections that there is no unique relationship between the collective dynamics that drive the population activity in the neural manifold, and the tuning space that reflects the principles of the circuit structure. If the collective dynamics do not provide information about the tuning space, can the latter still be extracted by analysing the neuronal activity? In this section, we propose an approach to this problem, that we illustrate in a model of contextual decision-making.

Individual statistics provide embeddings of the neurons' positions.

We analyze recordings of the activity of N neurons at T time points, represented as a $N \times T$ matrix $\mathbf{r}$, whose entries $\mathbf{r}_{i,k} = r_i(t_k)$ correspond to the firing rate of neuron i at time t_k. Each *column* $\mathbf{r}_{:,k} \in \mathbb{R}^N$ of this matrix represents population activtiy at time t_k. If sufficiently many time points are recorded, these vectors will accumulate in $\mathbb{R}^N$, allowing to extract the neural manifold and the collective dynamics from the activity [31, 32, 35, 38, 43, 44], as illustrated in Fig. 1F.

In contrast, in a field model, the firing rate $r_i(t) = r(t, \boldsymbol{z}_i)$ of any neuron i depends on its positions $\boldsymbol{z}_i$. Therefore, each *row* $\mathbf{r}_{i,:}$ of the activity matrix $\mathbf{r}$ represents the trajectory of a single neuron i as a T-dimensional vector that depends on its position $\boldsymbol{z}_i$:

$$\mathbf{r}_{i,:} = (r(t_1, \boldsymbol{z}_i), r(t_2, \boldsymbol{z}_i), \ldots, r(t_T, \boldsymbol{z}_i)) \in \mathbb{R}^T \tag{7}$$

If the dependence on $\boldsymbol{z}_i$ is smooth and sufficiently many neurons are recorded, these points will accumulate in $\mathbb{R}^T$ to form a T-dimensional embedding of the d-dimensional tuning space.

We start with the two previous example models (Eqs.(4) and (5)) that generate a one-dimensional limit-cycle in a two-dimensional neural manifold, but assume a sigmoidal rate function $\phi(x) = R\,(\tanh(x) + 1)/2$. Although in this case the collective dynamics of the two models are not strictly identical, they remain similar (Appendix B.2), so that the activity of corresponding SNNs follow similar 2-dimensional trajectories (Fig.4A,B). Importantly, the values of each neuron's firing rate at

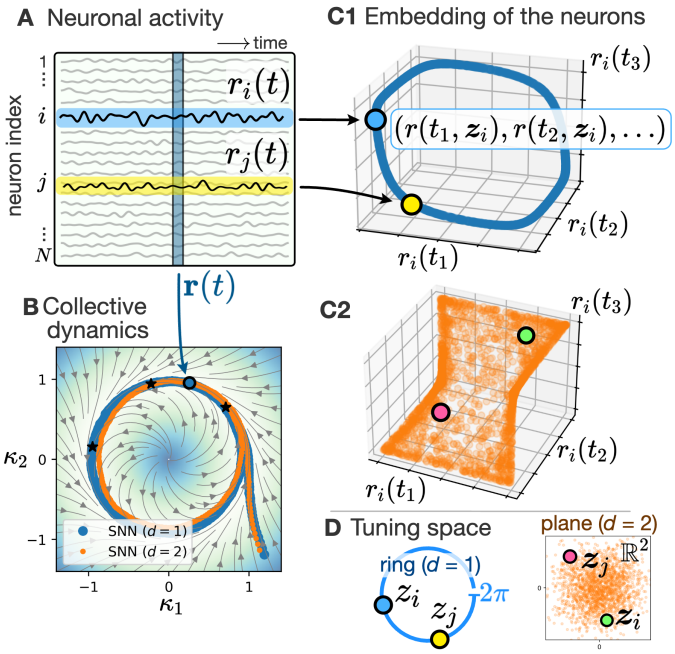

Fig. 4 *The dimension d of the tuning space can be retrieved from neuronal activity.* **A.** In a circuit model, each *column* of the population activity, i.e., the *momentary* firing rate vector $\boldsymbol{r}(t) \in \mathbb{R}^N$, lies on the neural manifold; whereas each *row* corresponds to a given neuron's time-dependent firing rate $r_i(t) = r(t, \boldsymbol{z}_i)$ which depends on its position $\boldsymbol{z}_i$ in the tuning space. **B.** The N-dimensional activity of two qualitatively different SNNs of $N = 2000$ neurons each follow similar two-dimensional limit-cycle dynamics on the neural manifold, even though the tuning space of the two SNNs have different dimension d (blue: $d = 1$; orange: $d = 2$). The plotting conventions are the same as in Fig. 3B,D. **C1.** Recording the firing rate $(r_i(t_1), r_i(t_2), r_i(t_3))$ of each neuron i at three different time points (black stars in panel (b)) yields a three-dimensional embedding of the neuronal position $\boldsymbol{z}_i$ in the tuning space. For the first SNN, the point-cloud of the embeddings of all the neurons in the population forms a ring, consistent with the topology of the underlying tuning space. **C2.** Similarly, for the second SNN, the set of all three-dimensional embeddings lies on a curved two-dimensional surface. **C.** The actual tuning space of the blue ($d = 1$) and orange ($d = 2$) model used in (B) agrees with the manifolds extracted in (C).

three different time points yield a 3-dimensional point-cloud whose topology reveals the differences between the tuning space of the two models (Fig.4C).

This reasoning can be extended to sets of T neuron-specific statistical indicators such as z-scored spike counts of a given neuron i across T different time bins, or, the projections of the activity of neuron i ("loadings") onto the T first principal components of the activity. Such statistics, recorded across many neurons, will accumulate to form the (potentially low-dimensinoal) tuning space embedded in the T-dimensional analysis space.

Extracting the tuning space in a model of contextual decision-making

We illustrate the extraction of the tuning space with a simple model of contextual decision-making (CDM) in the presence of external inputs [35]. During each trial, the network receives a sensory input consisting of two independent signals (I_1, "color" and I_2, "motion") that each fluctuates in time around some fixed positive or negative mean value; as well as a binary context input $I_c(t)$ which indicates the relevant signal, taking value $+1$ (respectively -1) to indicate the "color" (resp. "motion") signal (Fig.5A) The task consists in indicating whether the relevant signal is positive or negative at trial onset. Such a task can be solved by two-dimensional collective dynamics in rank-two networks [46]. We reformulate the task in our framework.

Building on the previous examples, we study two circuit models defined over different tuning spaces. Both models, implemented by SNNs, generate similar low-dimensional trajectories and task-related responses. For both models, the input weights, i.e., the contribution of the sensory and contextual signals to the external input received by each neuron i, are parameterised by functions of the neuron's position $\boldsymbol{z}_i$ (Fig.5A). Similarly, the network outputs a one-dimensional readout of the spiking activity with output weights also determined by the neuron's position.

Since the connectivity has rank two, the spiking network generates collective dynamics in a neural manifold of dimension $D = 2$ (see Methods). Between two trials (with context input $I_c = 0$), the latent variables decay towards a baseline state; but in the "color" context ($I_c = +1$), symmetric attractive fixed points appear along a "color"-relevant axis, and similarly in the "motion" context ($I_c = -1$), as illustrated in Fig.5B. The collective dynamics are implemented using two different SNNs corresponding to two circuit models called the "blue" and the "orange" model for now. Our aim is to identify the tuning space of the two models from the neuronal activity.

Retrieving the circuit structure from the population activity.

We used targeted dimensionality reduction to obtain low-dimensional trajectories corresponding to the population activity of each SNN (Fig.5C,D). Two-dimensional trajectories were obtained by projecting the N-dimensional spiking activity onto "color" and "motion"-relevant axes. Upon receiving the same sensory and context inputs, the low-dimensional trajectories and the choice readouts of the two models are similar (Fig.5D,E). On the other hand, statistical indicators of individual

neurons based on measurements of the activity within three different time windows yield a 3-dimensional embedding of each neuron. For both models, the resulting point-cloud lies in a two-dimensional subspace (Fig.5F). Using persistent homology to characterise the topology of these point-clouds [31, 32, 58, 59], the ring topology of the tuning space is confirmed by a persistent single ring-feature for the "blue" model, but not the "orange" one (Fig. 5G). This identification of the tuning space from statistical features is indeed correct and corresponds to the tuning space of the two models, even though knowledge about the tuning space was not used for model identification. The "blue" circuit model is a ring model with the neuronal positions uniformly distributed over the ring whereas the tuning space of the "orange" circuit model is the 2-dimensional domain $[-1, 1]^2$ in which neurons are uniformly distributed (Fig.5H). In both models individual neurons exhibit mixed selectivity. Yet in the ring model, the selectivity of any neuron i to each of the two sensory input features are correlated, so that the vector $\boldsymbol{f}(z_i)$ lies on the circle in $\mathbb{R}^2$; while in the second model, the selectivity of each neuron i to the two features, encoded by the two components of its position $\boldsymbol{z}_i$, are uncorrelated.

Discussion

In order to establish links between the tuning space (characterizing the principles of circuit structure) and the neural manifold (characterizing computation via the collective dynamics of neuronal activity) we asked two specific questions in the introduction. The first question needs to be answered negatively: there is no unique relationship between the tuning space and the neural manifold. In particular, different circuit structures can yield equivalent collective dynamics. Nevertheless, there are links since we showed that the symmetries of the circuit structure (in the tuning space) are inherited by the collective dynamics (in the neural manifold). Moreover, we highlighted heterogeneity in the population as a potential cause for the mismatch between the tuning space and the neural manifold.

As an answer to the second question, we proposed a method for extracting the topology and dimension of the tuning space from the population activity, independently of the collective dynamics. It is worth noticing that our method relies on the fact that the circuit structure of a model does not determine the dimension of the neural manifold; yet, it shapes its *geometry*, i.e. the shape of its embedding in the high-dimensional space of the activity [60]. In this regard, the proposed method relates to an analysis of the geometry of the population activity.

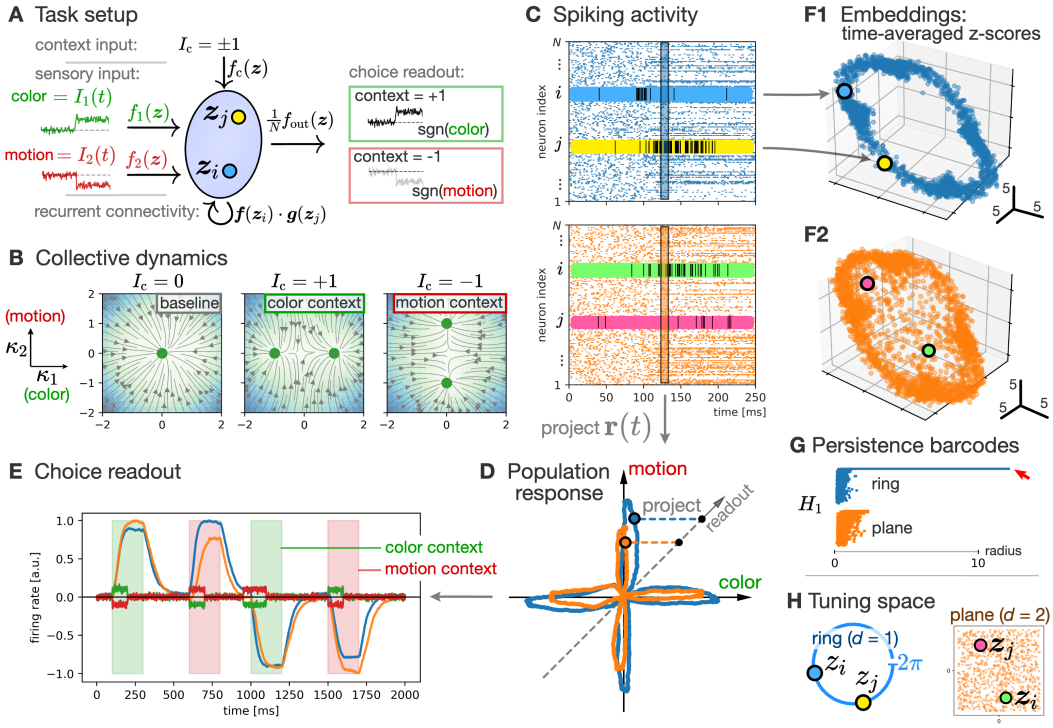

Fig. 5 *Contextual decision-making in two field models of different tuning-space dimension d.* **A.** The network receives a sensory input consisting of two features ($I_1(t)$, "color" and $I_2(t)$, "motion"), each fluctuating around a fixed mean value that can be positive or negative; and a binary context input $I_c = \pm 1$. The task of the network is to indicate the sign of the sensory feature that is specified by the context (I_1 for $I_c = +1$, I_2 for $I_c = -1$). Input, recurrent, and output synaptic weights are determined by functions of the neurons' positions $\boldsymbol{z}$ in the tuning space. **B.** The task is solved by two-dimensional collective dynamics, where the context input $I_c = \pm 1$ generates attractors (green dots) along the axis that corresponds to the relevant sensory input feature. **C.** Spike trains of 150 neurons from two SNNs (blue SNN, top; orange SNN, bottom; $N = 3000$ neurons each). Spike trains of example neurons i and j are highlighted. **D.** Low-dimensional trajectory of the two SNNs during 4 trials, obtained by projecting the N-dimensional spiking activity $\boldsymbol{r}(t)$ onto "color" and "motion"-relevant axes. The choice readout is the projection of the population activity on a diagonal readout axis which is a linear combination of the "color" and "motion" axes. **E.** Time-course of the external inputs (green and red traces, for the "color" and "motion" features, respectively) and normalised choice readout of the two networks (blue and orange traces) during the 4 trials. The context input takes value $I_c = +1$ (respectively -1) in the "color" (resp. "motion") context indicated by green (resp. red) shadings, and $I_c = 0$ in between trials. Both networks correctly solve the task. **F.** 3-dimensional embeddings of the neurons, generated by time-averages of each neuron's z-scored activity within three different time intervals. Both point-clouds in panels (F1) and (F2) lie within a 2-dimensional linear subspace (99% variance explained by 2 first principal components). **G.** Persistence barcodes of the ring features H_1 of the simplicial complexes associated with the point-clouds of panels (F1)-(F2). The ring topology of the point-cloud in (F1) is characterised by the persistent ring feature (long bar indicated with red arrow). **H.** The actual tuning spaces of the blue ($d = 1$) and orange models ($d = 2$) agree with the barcode predictions in panel (G).

In passing we also addressed the open problem of the relationship between classical circuit models structured by the tuning properties of the neurons, and models of low-rank RNNs. Our modeling framework shows that, in the limit of large networks, both classical circuit models and low-rank models are encompassed by neural field models. Our formalism allowed us to express the collective dynamics that govern the population activity, and to link the population activity with the concepts of tuning space (associated with classical circuit models) and neural manifold.

While our model opens the door to a unified perspective on the link between the principles of circuit structure, neuronal activity, and behaviour, it does not incorporate important features of more realistic models of biological neural networks. First, in our model, the recurrent connectivity and neuronal properties are dominated by a homogeneous structural principle whereas cortical connectivity appears to be, at least partly, random [61]; moreover, neuronal responses exhibit a degree of disorder [15, 35, 62]. These aspects have been incorporated into several earlier models and can lead to chaotic activity and trial-to-trial variability [46, 63]. The generalisation of our theory to the case of mixed structured and random connectivity, following [46], is left for future work.

Second, the neurons in our model are Poisson neurons, which do not account for refractoriness. Our theory can be partly extended to models of renewal neurons (such as the Spike Response Model or Integrate-and-Fire models [50, 64]). In this case, the collective dynamics can no longer be described as an ODE. Yet, in the regime of asynchronous firing, each neuron's firing rate can be expressed as a non-linear function of its synaptic input [65, 66], so that our formalism can be applied to the description of the fixed points of the asynchronous population activity.

Third, cortical networks exhibit a balance between excitation and inhibition [67, 68], a typical feature of balanced networks [69–71], including spatially structured networks [72]. Whether our formalism could be applied to obtain a low-dimensional description of the population activity in spatially structured balanced networks remains an open problem.

Finally, we did not apply our method for the extraction of the circuit structure to experimental data. The question of whether such structure could be observed in experimental recordings is, at this stage, still open. A positive answer would suggest that, even in a population where neurons exhibit mixed selectivity and seemingly heterogeneous responses, the degree of heterogeneity in the population is reducible to a low-dimensional description. This would imply that the general framework

of field models is suitable for the description of the neuronal activity. Our work contributes to link this general mechanistic framework to a descriptive, population-level approach that leverages the statistical power of large-scale neural recordings.

February 28, 2024

References

[1] Hubel, D.H., Wiesel, T.N.: Receptive fields of single neurones in the cat's striate cortex. The Journal of Physiology **148**(3), 574–591 (1959) https://doi.org/10.1113/jphysiol.1959.sp006308 https://physoc.onlinelibrary.wiley.com/doi/pdf/10.1113/jphysiol.1959.sp006308

[2] Urai, A.E., Doiron, B., Leifer, A.M., Churchland, A.K.: Large-scale neural recordings call for new insights to link brain and behavior. Nature Neuroscience **25**(1), 11–19 (2022)

[3] Barack, D.L., Krakauer, J.W.: Two views on the cognitive brain. Nature Reviews Neuroscience **22**(6), 359–371 (2021)

[4] Langdon, C., Genkin, M., Engel, T.A.: A unifying perspective on neural manifolds and circuits for cognition. Nature Reviews Neuroscience **24**(6), 363–377 (2023)

[5] Muller, L., Chavane, F., Reynolds, J., Sejnowski, T.J.: Cortical travelling waves: mechanisms and computational principles. Nature Reviews Neuroscience **19**(5), 255–268 (2018)

[6] J.P. Rauschecker, M.H. B. Tian: Processing of complex sounds in the macaque nonprimary auditory cortex. Science **268**, 111–114 (1995)

[7] Georgopoulos, A.P., Schwartz, A., Kettner, R.E.: Neuronal population coding of movement direction. Science **233**, 1416–1419 (1986)

[8] McNaughton, B.L., Barnes, C.A., O'Keefe, J.: The contributions of position, direction, and velocity to single unit activity in the hippocampus of freely-moving rats. Experimental Brain Research **52**, 41–49 (1983)

[9] Moser, E.I., Moser, M.-B., McNaughton, B.L.: Spatial representation in the hippocampal formation: a history. Nature Neuroscience **20**(11), 1448–1464 (2017)

[10] Taube, J., Muller, R., JB Ranck, J.: Head-direction cells recorded from the postsubiculum in freely moving rats. 1. description and quantitative analysis. Journal of Neuroscience **10**(2), 420–435 (1990)

[11] Kim, S.S., Rouault, H., Druckmann, S., Jayaraman, V.: Ring attractor dynamics in the drosophila central brain. Science

356(6340), 849–853 (2017) https://doi.org/10.1126/science.aal4835 https://www.science.org/doi/pdf/10.1126/science.aal4835

[12] Doeller, C.F., Barry, C., Burgess, N.: Evidence for grid cells in a human memory network. Nature **463**(7281), 657–661 (2010)

[13] Hirokawa, J., Vaughan, A., Masset, P., Ott, T., Kepecs, A.: Frontal cortex neuron types categorically encode single decision variables. Nature **576**(7787), 446–451 (2019)

[14] Raposo, D., Kaufman, M.T., Churchland, A.K.: A category-free neural population supports evolving demands during decision-making. Nature Neuroscience **17**(12), 1784–1792 (2014)

[15] Rigotti, M., Barak, O., Warden, M.R., Wang, X.-J., Daw, N.D., Miller, E.K., Fusi, S.: The importance of mixed selectivity in complex cognitive tasks. Nature **497**(7451), 585–590 (2013)

[16] Amari, S.-i.: Dynamics of pattern formation in lateral-inhibition type neural fields. Biological Cybernetics **27**(2), 77–87 (1977)

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

[18] Zhang, K.: Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. Journal of Neuroscience **16**(6), 2112–2126 (1996) https://doi.org/10.1523/JNEUROSCI.16-06-02112.1996 https://www.jneurosci.org/content/16/6/2112.full.pdf

[19] Wilson, H.R., Cowan, J.D.: A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. Kybernetik **13**(2), 55–80 (1973)

[20] Burak, Y., Fiete, I.R.: Accurate Path Integration in Continuous Attractor Network Models of Grid Cells. PLOS Computational Biology **5**(2), 1000291 (2009)

[21] McNaughton, B.L., Battaglia, F.P., Jensen, O., Moser, E.I., Moser, M.-B.: Path integration and the neural basis of the 'cognitive map'. Nature Reviews Neuroscience **7**(8), 663–678 (2006)

[22] Coombes, S., Beim Graben, P., Potthast, R., Wright, J.: Neural Fields: Theory and Applications, (2014). https://doi.org/10.1007/978-3-642-54593-1

[23] Licurgo de Almeida, Marco Idiart, John E. Lisman: The Input–Output Transformation of the Hippocampal Granule Cells: From Grid Cells to Place Fields. The Journal of Neuroscience **29**(23), 7504 (2009)

[24] Bednar, J.A., Wilson, S.P.: Cortical maps. The Neuroscientist **22**, 604–617 (2016)

[25] Bonhoeffer, T., Grinvald, A.: lso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns. Nature **353**, 429–431 (1991)

[26] Bressloff, P.C.: Spatiotemporal dynamics of continuum neural fields. Journal of Physics A: Mathematical and Theoretical **45**(3), 033001 (2011)

[27] Gerstner, W.: Time structure of the activity in neural network models. Phys. Rev. E **51**, 738–758 (1995) https://doi.org/10.1103/PhysRevE.51.738

[28] Chevallier, J., Duarte, A., Löcherbach, E., Ost, G.: Mean field limits for nonlinear spatially extended hawkes processes with exponential memory kernels. Stochastic Processes and their Applications **129**(1), 1–27 (2019)

[29] Saxena, S., Cunningham, J.P.: Towards the neural population doctrine. Current Opinion in Neurobiology **55**, 103–111 (2019)

[30] Khona, M., Fiete, I.R.: Attractor and integrator networks in the brain. Nature Reviews Neuroscience **23**(12), 744–766 (2022)

[31] Gardner, R.J., Hermansen, E., Pachitariu, M., Burak, Y., Baas, N.A., Dunn, B.A., Moser, M.-B., Moser, E.I.: Toroidal topology of population activity in grid cells. Nature **602**(7895), 123–128 (2022)

[32] Chaudhuri, R., Gerçek, B., Pandey, B., Peyrache, A., Fiete, I.: The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep. Nature Neuroscience **22**(9), 1512–1520 (2019)

[33] Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Foster, J.D., Nuyujukian, P., Ryu, S.I., Shenoy, K.V.: Neural population dynamics during reaching. Nature **487**(7405), 51–56 (2012)

[34] Vyas, S., Golub, M.D., Sussillo, D., Shenoy, K.V.: Computation through neural population dynamics. Annual Review of Neuroscience **43**(1), 249–275 (2020)

[35] Mante, V., Sussillo, D., Shenoy, K.V., Newsome, W.T.: Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature **503**(7474), 78–84 (2013)

[36] Shenoy, K.V., Sahani, M., Churchland, M.M.: Cortical control of arm movements: A dynamical systems perspective. Annual Review of Neuroscience **36**(1), 337–359 (2013) https://doi.org/10.1146/annurev-neuro-062111-150509 . PMID: 23725001

[37] Safaie, M., Chang, J.C., Park, J., Miller, L.E., Dudman, J.T., Perich, M.G., Gallego, J.A.: Preserved neural dynamics across animals performing similar behaviour. Nature **623**(7988), 765–771 (2023)

[38] Gallego, J.A., Perich, M.G., Miller, L.E., Solla, S.A.: Neural manifolds for the control of movement. Neuron **94**(5), 978–984 (2017)

[39] DePasquale, B., Sussillo, D., Abbott, L.F., Churchland, M.M.: The centrality of population-level factors to network computation is demonstrated by a versatile approach for training spiking networks. Neuron **111**(5), 631–64910 (2023) https://doi.org/10.1016/j.neuron.2022.12.007

[40] Burgess, N., Recce, M., O'Keefe, J.: A model of hippocampal function. Neural Networks **7**, 1065–1081 (1994)

[41] Jazayeri, M., Ostojic, S.: Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. Current Opinion in Neurobiology **70**, 113–120 (2021)

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

[43] Jensen, K.T., Kao, T.-C., Tripodi, M., Hennequin, G.: Manifold gplvms for discovering non-euclidean latent structure in neural data. NIPS'20. Curran Associates Inc., Red Hook, NY, USA (2020)

[44] Schneider, S., Lee, J.H., Mathis, M.W.: Learnable latent embeddings for joint behavioural and neural analysis. Nature **617**(7960), 360–368 (2023)

[45] Hopfield, J.J.: Neurons with graded response have collective computational properties like those of two-state neurons. Proceedings of the National Academy of Sciences **81**(10), 3088–3092 (1984) https://doi.org/10.1073/pnas.81.10.3088 https://www.pnas.org/doi/pdf/10.1073/pnas.81.10.3088

[46] Mastrogiuseppe, F., Ostojic, S.: Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. Neuron **99**(3), 609–62329 (2018)

[47] Beiran, M., Dubreuil, A., Valente, A., Mastrogiuseppe, F., Ostojic, S.: Shaping Dynamics With Multiple Populations in Low-Rank Recurrent Networks. Neural Computation **33**(6), 1572–1615 (2021) https://doi.org/10.1162/neco_a_01381

[48] Sussillo, D., Abbott, L.F.: Generating coherent patterns of activity from chaotic neural networks. Neuron **63**(4), 544–557 (2009) https://doi.org/10.1016/j.neuron.2009.07.018

[49] Dubreuil, A., Valente, A., Beiran, M., Mastrogiuseppe, F., Ostojic, S.: The role of population structure in computations through neural dynamics. Nature Neuroscience **25**(6), 783–794 (2022) https://doi.org/10.1038/s41593-022-01088-4

[50] Wulfram, G., Kistler, W.M., Richard, N., Liam, P.: Neuronal Dynamics. From Single Neurons to Networks and Models of Cognition. Cambridge University Press, ??? (2014)

[51] Delattre, S., Fournier, N., Hoffmann, M.: Hawkes processes on large networks. The Annals of Applied Probability **26**(1), 216–261 (2016)

[52] Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences **79**(8), 2554–2558 (1982) https://doi.org/10.1073/pnas.79.8.2554 https://www.pnas.org/doi/pdf/10.1073/pnas.79.8.2554

[53] Pereira, U., Brunel, N.: Attractor dynamics in networks with learning rules inferred from in vivo data. Neuron **99**(1), 227–2384 (2018)

[54] Schmutz, V., Brea, J., Gerstner, W.: Emergent rate-based dynamics in duplicate-free populations of spiking neurons (2024)

[55] Amit, D.J., Gutfreund, H., Sompolinsky, H.: Spin-glass models of neural networks. Phys. Rev. A **32**, 1007–1018 (1985) https://doi.org/10.1103/PhysRevA.32.1007

[56] Schuessler, F., Dubreuil, A., Mastrogiuseppe, F., Ostojic, S., Barak, O.: Dynamics of random recurrent networks with correlated low-rank structure. Phys. Rev. Res. **2**, 013111 (2020) https://doi.org/10.1103/PhysRevResearch.2.013111

[57] Veltz, R., Faugeras, O.: Local/global analysis of the stationary solutions of some neural field equations. SIAM Journal on Applied Dynamical Systems **9**(3), 954–998 (2010) https://doi.org/10.1137/090773611 https://doi.org/10.1137/090773611

[58] Rybakken, E., Baas, N., Dunn, B.: Decoding of Neural Data Using Cohomological Feature Extraction. Neural Computation **31**(1), 68–93 (2019)

[59] Ghrist, R.: Barcodes: the persistent topology of data. Bull. Amer. Math. Soc. (N.S.) **45**(1), 61–75 (2008) https://doi.org/10.1090/S0273-0979-07-01191-3

[60] Kriegeskorte, N., Wei, X.-X.: Neural tuning and representational geometry. Nature Reviews Neuroscience **22**(11), 703–718 (2021)

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

[62] Churchland, M.M., Shenoy, K.V.: Temporal complexity and heterogeneity of single-neuron activity in premotor and motor cortex. Journal of Neurophysiology **97**(6), 4235–4257 (2007) https://doi.org/10.1152/jn.00095.2007 https://doi.org/10.1152/jn.00095.2007. PMID: 17376854

[63] Sompolinsky, H., Crisanti, A., Sommers, H.J.: Chaos in random neural networks.

Phys. Rev. Lett. **61**, 259–262 (1988) https://doi.org/10.1103/PhysRevLett.61.259

[64] Gerstner, W.: Population dynamics of spiking neurons: Fast transients, asynchronous states, and locking **12**(1), 43–89 https://doi.org/10.1162/089976600300015899 . Accessed 2023-02-09

[65] Spiridon, M., Gerstner, W.: Effect of lateral connections on the accuracy of the population code for a network of spiking neurons. Network: Computation in Neural Systems **12**(4), 409 (2001) https://doi.org/10.1088/0954-898X/12/4/301

[66] Ostojic, S., Brunel, N.: From spiking neuron models to linear-nonlinear models. PLOS Computational Biology **7**(1), 1–16 (2011) https://doi.org/10.1371/journal.pcbi.1001056

[67] Shu, Y., Hasenstaub, A., McCormick, D.A.: Turning on and off recurrent balanced cortical activity. Nature **423**(6937), 288–293 (2003)

[68] Haider, B., Duque, A., Hasenstaub, A.R., McCormick, D.A.: Neocortical network activity in vivo is generated through a dynamic balance of excitation and inhibition. Journal of Neuroscience **26**(17), 4535–4545 (2006) https://doi.org/10.1523/JNEUROSCI.5297-05.2006 https://www.jneurosci.org/content/26/17/4535.full.pdf

[69] Brunel, N.: Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. Journal of Computational Neuroscience **8**(3), 183–208 (2000)

[70] Renart, A., Rocha, J., Bartho, P., Hollender, L., Parga, N., Reyes, A., Harris, K.D.: The asynchronous state in cortical circuits. Science **327**(5965), 587–590 (2010) https://doi.org/10.1126/science.1179850 https://www.science.org/doi/pdf/10.1126/science.1179850

[71] Litwin-Kumar, A., Doiron, B.: Slow dynamics and high variability in balanced cortical networks with clustered connections. Nature Neuroscience **15**(11), 1498–1505 (2012)

[72] Rosenbaum, R., Doiron, B.: Balanced networks of spiking neurons with spatially dependent recurrent connections. Phys. Rev. X **4**, 021039 (2014) https://doi.org/10.1103/PhysRevX.4.021039

[73] Fournier, N., Löcherbach, E.: On a toy model of interacting neurons. Annales de l'Institut Henri Poincaré, Probabilités et Statistiques **52**(4), 1844–1876 (2016)

[74] Faugeras, O., Grimbert, F., Slotine, J.-J.: Absolute stability and complete synchronization in a class of neural fields models. SIAM Journal on Applied Mathematics **69**(1), 205–250 (2008) https://doi.org/10.1137/070694077 https://doi.org/10.1137/070694077

[75] Renardy, M., Rogers, R.C.: Operator Theory, 2nd edn., pp. 228–282. Springer, ??? (2004)

[76] Conway, J.B.: A Course in Functional Analysis, pp. 42–43267. Springer, ??? (2007). https://doi.org/10.1007/978-1-4757-4383-8_9 . https://doi.org/10.1007/978-1-4757-4383-8$_9$

[77] Cybenko, G.: Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems **2**(4), 303–314 (1989)

Methods

A unifying framework: field model of low-rank networks

Network model.

We turn here to a more precise mathematical formulation. We consider a model of Recurrent Neural Network (RNN) composed of N Poisson neurons [50] (or nonlinear Hawkes processes [28, 51]). The spike train $S_i(t)$ of each neuron $i = 1, \ldots, N$ is generated by an inhomogeneous Poisson process with intensity $\phi(h_i(t))$, where $h_i(t)$ represents the neuron's membrane potential, and $\phi : \mathbb{R} \to \mathbb{R}^+$ is a non-linear bounded voltage-to-rate function. The potential $h_i(t)$ is a leaky integrator of the synaptic input; which, for the ease of exposition, we first consider as coming only from recurrent connections. The dynamics of the potential thus follow:

$$\frac{d}{dt}h_i(t) = -\frac{1}{\tau}h_i(t) + \frac{1}{N}\sum_{\substack{j=1\\ j\neq i}}^{N} W_{ij}S_j(t) \tag{8}$$

with given initial condition $h_i(0)$. Here, $\tau > 0$ is the membrane time constant; and the synaptic weights, scaled as $1/N$, are given by a matrix of rank $D \in \mathbb{N}$ fixed:

$$W_{ij} = \sum_{\mu=1}^{D} F_{i,\mu}G_{j,\mu} = \vec{F}_i \cdot \vec{G}_j \tag{9}$$

where the dot denotes the scalar product in $\mathbb{R}^D$. Each entry W_{ij} of the low-rank connectivity matrix corresponds to the scalar product of two D-dimensional "connectivity vectors", $\vec{F}_i = (F_{i,1}, \ldots, F_{i,D})$ and $\vec{G}_i = (G_{i,1}, \ldots, G_{i,D})$, associated with neurons i and j (see Fig.2). As discussed in the Results section, Eq. (9) is a general formulation of the

constraint of a low-rank connectivity; specific models of low-rank networks are essentially built by specifying the distribution of these vectors across the population [45–47, 52–56]. In what follows, we introduce a general model for the distribution of the connectivity vectors. The resulting class of RNNs encompasses most low-rank models encountered in the literature. Moreover, for fixed rank D, as $N \to \infty$, it corresponds to a field model whose field dimension d is not necessarily related to the rank D of the connectivity, i.e. to the dimension of the neural manifold.

A general model of low-rank connectivity.

As we are interested in the limit of large N, let's start from the simple observation that, if one wants to add a neuron labelled i^* to the network, one has to specify its connectivity, through the associated vectors $\vec{F}_{i^*}$ and $\vec{G}_{i^*}$, as well as its initial condition $h_{i^*}(0)$. It is commonly assumed that these features are independent across neurons: intuitively, it means one can add a neuron i^* to the network without knowing the $\{\vec{F}_i, \vec{G}_i, h_i(0)\}_{i=1}^N$ associated with all other neurons. However, for a given neuron, those might be correlated. For the new neuron i^*, we thus model these features as being determined by some individual properties $\boldsymbol{z}_{i^*}$ of the neuron, sampled from a given distribution.

Formally, we represent $\boldsymbol{z}_{i^*}$ by the neuron's *position* on a smooth manifold $\mathcal{M}$ of intrinsic dimension d, drawn at random according to a probability distribution $\rho(d\boldsymbol{z})$. Given the position $\boldsymbol{z}_i$ of any neuron i, the associated connectivity vectors $\vec{F}_i$, $\vec{G}_i$ are determined by mappings $\boldsymbol{f}, \boldsymbol{g} : \mathcal{M} \to \mathbb{R}^D$:

$$F_{i,\mu} = f_\mu(\boldsymbol{z}_i), \; G_{i,\mu} = g_\mu(\boldsymbol{z}_i) \quad (\mu = 1, \dots, D) \tag{10}$$

with f_μ, g_μ the μ-th components of $\boldsymbol{f}$, $\boldsymbol{g}$. This is illustrated on Fig.2. Analogously, the initial condition is determined by a map $h_0 : \mathcal{M} \to \mathbb{R}$, according to $h_i(0) = h_0(\boldsymbol{z}_i)$. Eqs.(9),(10) define a kernel $w(\boldsymbol{z}, \boldsymbol{z}') = \boldsymbol{f}(\boldsymbol{z}) \cdot \boldsymbol{g}(\boldsymbol{z}')$ such that the recurrent connectivity is given by $W_{ij} = w(\boldsymbol{z}_i, \boldsymbol{z}_j)$. The manifold $\mathcal{M}$ thus represents the tuning space, in which the neurons' positions determine the connectivity. The distribution of the neurons over the tuning space, described by $\rho(d\boldsymbol{z})$, defines a scalar product between functions over $\mathcal{M}$:

$$\langle \varphi, \psi \rangle_\rho := \int_{\mathcal{M}} \rho(d\boldsymbol{z}) \varphi(\boldsymbol{z}) \psi(\boldsymbol{z}) \tag{11}$$

In the following, we assume that the components f_μ, g_μ ($\mu = 1, \dots, D$) of the connectivity mappings, as well as the initial condition h_0, are continuous and belong to the vector space of ρ-square-integrable functions over $\mathcal{M}$. This last condition defines a Hilbert space $\mathcal{H}$ equipped with the ρ-scalar product $\langle \cdot, \cdot \rangle_\rho$.

Corresponding field model.

As $N \to \infty$, the RNN defined by eqs.(8,9,10) is described by a field model: the membrane potential at time t of a neuron at position $\boldsymbol{z}$ is given by the neural field $v_t(\boldsymbol{z})$, which obeys the *field equation*

$$\partial_t v_t(\boldsymbol{z}) = -\frac{1}{\tau} v_t(\boldsymbol{z}) + \int_{\mathcal{M}} \rho(d\boldsymbol{z}')\, \boldsymbol{f}(\boldsymbol{z}) \cdot \boldsymbol{g}(\boldsymbol{z}')\, \phi(v_t(\boldsymbol{z}')) \tag{12}$$

with initial condition $v_0(\boldsymbol{z}) = h_0(\boldsymbol{z})$. It is the continuum version of Eq. (8) [27, 50], where the Poisson noise from the spikes vanishes [28, 54]. Under suitable regularity and boundedness assumptions concerning the transfer function ϕ, the convergence as $N \to \infty$ of each membrane potential $h_i(t)$ to $v_t(\boldsymbol{z}_i)$ is guaranteed, with rate $1/\sqrt{N}$. It follows that neurons become pairwise independent [28, 73].

Low-dimensional dynamics.

Analogously to a rank-D RNN, the dynamics of the neural field can be written as a D-dimensional system [57], as presented here in a simple fashion. The field v_t, at any time t, is an element of the Hilbert space $\mathcal{H}$ (provided that $h_0, \{f_\mu\}_\mu \subset \mathcal{H}$). The field equation (12) can thus be written as an ODE in $\mathcal{H}$: [74]

$$\partial_t v_t = -\frac{1}{\tau} v_t + \sum_{\mu=1}^{D} f_\mu \langle g_\mu, \phi \circ v_t \rangle_\rho \tag{13}$$

where ($\circ$) denotes the composition of functions, and $\langle \cdot, \cdot \rangle_\rho$ the ρ-scalar product of Eq. (11). The field v_t undergoes an exponential decay, except in the linear subspace $\mathcal{F} := \mathrm{span}\{f_\mu\} \subset \mathcal{H}$. It can thus be written as a linear combination of the basis functions $\{f_\mu\}$, whose coefficients $\kappa_\mu(t)$ are the "latent variables":

$$v_t(\boldsymbol{z}) = \sum_{\mu=1}^{D} f_\mu(\boldsymbol{z}) \kappa_\mu(t) \; + e^{-t/\tau} h_0^{\perp}(\boldsymbol{z}) \tag{14}$$

where $h_0^\perp = h_0 - \boldsymbol{f} \cdot \boldsymbol{\kappa}(0)$ is the projection of the initial condition h_0 onto the orthogonal complement of $\mathcal{F}$ in $\mathcal{H}$. By projection of Eq. (14) onto each basis function f_μ, the latent variables can be written in vector form as:

$$\boldsymbol{\kappa}(t) = M \langle \boldsymbol{f}, v_t \rangle_\rho \tag{15}$$

where we introduced the inverse correlation matrix M, given by $(M^{-1})_{\mu\nu} = \langle f_\mu, f_\nu \rangle_\rho$ (assuming without loss of generality that the functions $\{f_\mu\}$ are linearly independent in $\mathcal{H}$). Eqs.(13-14-15) yield D-dimensional collective dynamics for $\boldsymbol{\kappa}$. It can be written compactly:

$$\frac{d}{dt}\boldsymbol{\kappa}(t) = -\frac{1}{\tau}\boldsymbol{\kappa}(t) + \langle \boldsymbol{g}, \phi \circ \big(\boldsymbol{f} \cdot \boldsymbol{\kappa}(t)\big) \rangle_\rho \tag{16}$$

under the assumption that $h_0^\perp = 0$. This equation for the collective dynamics (and its more general expression, Eq.(38) in the Appendix) is the central ingredient of our theory. The derivation is given in Appendix A.

Neural manifold and attractor states.

Eq. (14) implies that the firing rate of each neuron i, given by $r_i(t) = \phi(h_i(t)) = \phi(v_t(\boldsymbol{z}_i))$, is parametrised by the latent variables $\boldsymbol{\kappa}$; and so is the N-dimensional network activity $\boldsymbol{r}(t)$. Besides, $\boldsymbol{r}(t)$ exponentially decays towards the D-dimensional neural manifold

$$\{\phi(F \cdot \boldsymbol{k}) \mid \boldsymbol{k} \in \mathbb{R}^D\} \subset \mathbb{R}^N \tag{17}$$

generated by the set of all possible latent variables; where $F \in \mathbb{R}^{N \times D}$ denotes the matrix of connectivity vectors $\vec{F}_i$ given by Eq. (10), and ϕ is applied element-wise. In the absence of external input, the latent variables $\boldsymbol{\kappa}(t)$ will converge to the set of attractor states of the collective dynamics of Eq. (16), which specifically depends on the considered model. In the Results, we give concrete examples, where attractor states are discrete fixed points and limit cycles.

Generalisations.

This derivation can be readily extended to the case of a network receiving a low-dimensional external input. Assume that the latter is composed of K features $I_\nu(t)$ ($\nu = 1, \ldots, K$), and that the tuning of any neuron i to each input feature I_ν is determined by a corresponding function $f_\nu^{\text{ext}}(\boldsymbol{z}_i)$

of its position $\boldsymbol{z}_i$. The external input it receives can be written in the form:

$$I_i^{\text{ext}}(t) = \sum_{\nu=1}^{K} f_\nu^{\text{ext}}(\boldsymbol{z}_i) I_\nu(t) \tag{18}$$

This results in a drive to the collective dynamics of Eq. (16), while expanding their dimensionality (at most up to $D+K$). The corresponding general expression of the collective dynamics is given by Eq.(38) in Appendix A.2. The derivation of the collective dynamics can also be generalised to the case of field models with arbitrary square-integrable interaction kernel over $\mathcal{M}^2$. In this case, the collective dynamics are not guaranteed to be finite-dimensional; but the field can always be well approximated by a low-rank model by truncation of the dimension [57]. Both generalisations are given in Appendix A.

Latent dimension vs. field dimension.

There is no assumption in our framework regarding any link between the dimension D of the neural manifold (which is determined by the rank of the connectivity) on one side, and the dimension d of the tuning space on the other side. Moreover, it is not clear whether the tuning space restricts the range of collective dynamics that can be implemented in an RNN, expressed by Eq. (16). These are the points we address in the Results.

Tuning space and neural manifold can mismatch: examples

A one-dimensional circuit structure generates two-dimensional collective dynamics

In our first example, the tuning manifold is the real line, over which the neurons' positions are normally distributed; i.e. $\mathcal{M} = \mathbb{R}$ and $z_i \sim \rho = \mathcal{N}(0,1)$ i.i.d. for $i = 1, \ldots, N$ (Fig.3A1). The interactions are given by:

$$W_{ij} = \begin{cases} J\, z_i z_j & (z_i z_j \geq 0) \\ \cos(\alpha) J\, z_i z_j & (z_i z_j < 0) \end{cases} \tag{19}$$

with $0 < \alpha < \pi/2$. This can be written as the 2-dimensional scalar product $W_{ij} = \boldsymbol{f}(z_i) \cdot \boldsymbol{g}(z_j)$, with:

$$\boldsymbol{f}(z) = \begin{pmatrix} [z]_+ + \cos(\alpha)[z]_- \\ -\sin(\alpha)[z]_- \end{pmatrix} \in \mathbb{R}^2 \quad ; \quad \boldsymbol{g} = J\boldsymbol{f} \tag{20}$$

where $[z]_\pm = z \cdot \mathbb{1}_{\{\mathrm{sgn}(z)=\pm 1\}}$ denotes the positive or negative part of z. This is illustrated on Fig.3A2. The corresponding collective dynamics resulting from Eq.(16) and displayed in Fig.3B are derived in Appendix B.1.

Two different circuit structures generate identical limit-cycle dynamics
Let $J = \sqrt{\frac{\pi}{2}}\tilde{J}$. The two models with connectivity given by Eqs.(4)-(5) obey *identical* collective dynamics: for the polar coordinates $(\rho(t), \theta(t))$ of the latent variable $\boldsymbol{\kappa}(t) \in \mathbb{R}^2$, Eq. (16) yields

$$\dot{\rho} = -\frac{1}{\tau}\rho + \frac{J}{\pi}R\cos\delta \quad ; \quad \rho\dot{\theta} = \frac{J}{\pi}R\sin\delta \tag{21}$$

for *both* models (derivation in Appendix B.2). For any initial condition, the latent variables converge to a globally attractive limit cycle with radius $\rho_0 = \frac{J}{\pi}\tau R\cos\delta$ and constant angular velocity $\dot{\theta} = \frac{1}{\tau}\tan\delta$.

Symmetries

In mathematical terms, a *symmetry* of the field model is a mapping $S : \mathcal{M} \to \mathcal{M}$ over the tuning space for which the field dynamics of Eq. (12) are equivariant; so that given any solution v_t to the dynamics, $v_t \circ S$ is also a solution. Similarly, a symmetry T of the collective dynamics is an equivariance of Eq.(16); so that for any solution $\boldsymbol{\kappa}(t)$, $T\boldsymbol{\kappa}(t)$ is also a solution. We show the following (see Theorem 1 in Appendix C):

1. Let S be (i) a symmetry of the field model, that (ii) does not change the D-dimensional space containing the recurrent input. This is expressed by the condition: $\boldsymbol{f}_\nu \circ S \in \mathcal{F} = \mathrm{span}\{f_\mu\}_{\mu=1}^D$, for all $\nu = 1, \ldots, D$. Then, the linear mapping $T^{(S)}$, given by the $D \times D$ matrix

$$T^{(S)}_{\mu\nu} = \sum_\sigma M_{\mu\sigma}\langle f_\sigma, f_\nu \circ S\rangle_\rho \tag{22}$$

   with $(M^{-1})_{\mu\nu} = \langle f_\mu, f_\nu\rangle_\rho$, is a symmetry of the collective dynamics.
2. If S is a bijection for which both the connectivity kernel w and the distribution of the positions ρ are invariant, i.e. $w(Sz, Sz') = w(z, z')$ and $\rho \circ S(dz) = \rho(dz)$, then both requirements (i) and (ii) of point 1. above are satisfied.

Furthermore, the mapping $S \mapsto T^{(S)}$ defines the action on $\mathbb{R}^D$ of the symmetry group G of the field model, which will also be a symmetry group of the collective dynamics. Yet, the order of composition of group elements is reversed: for any elements S, S' of G acting on $\mathcal{M}$, the action on $\mathbb{R}^D$ of their composition $S \circ S'$ is given by: $T^{(S\circ S')} = T^{(S')}T^{(S)}$. Therefore, a "right" group action on $\mathcal{M}$ becomes a "left" group action on $\mathbb{R}^D$: see

Corollary 1 in Appendix C. This will have an effect if the symmetry group G is non-commutative; like, for example, the group of rotations in $\mathbb{R}^3$.

Model of context-dependent decision-making

For both models, the contribution of the sensory and contextual input features I_1, I_2 and I_c to the external input received by each neuron i are parameterised by functions $f_1^\mathrm{ext}(z_i), f_2^\mathrm{ext}(z_i)$ and $f_\mathrm{c}(z_i)$, respectively. We impose the recurrent and the sensory input to be colinear by setting $f_\mu^\mathrm{ext} = f_\mu$ ($\mu = 1, 2$). The total external input received by each neuron i reads:

$$I(t, z_i) = \boldsymbol{f}(z_i) \cdot \boldsymbol{I}(t) + J_\mathrm{c} f_\mathrm{c}(z_i) I_\mathrm{c}(t) \tag{23}$$

where $\boldsymbol{I} = (I_1, I_2)$ is the sensory input, and J_c is an adjustable input strength parameter.

Since the maps f_1, f_2, f_c determine the contribution of the sensory features I_1, I_2 and context I_c to the external input each neuron receives, they can be understood as encoding the neuron's selectivity to these features, while f_out encodes each neuron's contribution to the output choice.

Collective dynamics

We assume symmetric couplings, i.e. $\boldsymbol{g} = J\boldsymbol{f}$; and arbitrary voltage-to-rate function ϕ. For a stationary context input $I_\mathrm{c}(t) = \pm 1$, the collective dynamics write (from Eq. (38)):

$$\frac{d}{dt}\boldsymbol{\kappa}(t) = -\frac{1}{\tau}\boldsymbol{\kappa}(t) + J\langle \boldsymbol{f}, \phi \circ (\boldsymbol{f} \cdot \boldsymbol{\kappa}(t) + \tau J_\mathrm{c} f_\mathrm{c} I_\mathrm{c})\rangle_\rho + \boldsymbol{I}(t) \tag{24}$$

The feature maps $\boldsymbol{f}(z)$, the context selectivity $f_\mathrm{c}(z)$ and the strength parameters J, J_c can be conveniently chosen such that in the absence of external input ($I_1 = I_2 = u_\mathrm{c} = 0$), these dynamics admit no non-trivial fixed point; but in the "color" context ($I_\mathrm{c} = +1$), symmetric attractive fixed points appear along the "color" axis at $\boldsymbol{\kappa} = (\pm k, 0)$, with $k > 0$. Similarly, the "motion" context ($I_\mathrm{c} = -1$) yields attractive fixed points along the "motion" axis at $\boldsymbol{\kappa} = (0, \pm k)$, as illustrated in Fig.5B.

This is achieved by setting $f_\mathrm{c} = |f_1| - |f_2|$. In the "color" context $I_\mathrm{c} = +1$, neurons that are more selective to the first input feature ("color"), i.e. for which $f_1(\boldsymbol{z}_i) > f_2(\boldsymbol{z}_i)$, are excited; while those that are more selective to the second input feature ("motion") are inhibited; and reciprocally in the "motion" context $I_\mathrm{c} = -1$.

At the beginning of each trial, the context input is switched on to $I_\mathrm{c} = \pm 1$, indicating the context. The population activity is then driven

towards one of the fixed points by the relevant component of the sensory input (I_1, "color" if $I_c = +1$, or I_2, "motion" if $I_c = -1$), while ignoring the irrelevant component; and it relaxes towards the baseline state when the context input is switched off, being ready for a new trial.

Choice readout

The choice readout is a one-dimensional readout of the activity with output weights $f_{\text{out}}(\boldsymbol{z}_i)$:

$$\text{readout}(t) = \frac{1}{N} \sum_{i=1}^{N} f_{\text{out}}(\boldsymbol{z}_i) S_i(t) \tag{25}$$

with S_i the spike train of neuron i. The map of output weights is set to $f_{\text{out}} = f_1 + f_2$. Thus, the readout is positive for the two fixed points that correspond to a positive sensory input ($(+k, 0)$ in the "color" context if $I_1 > 0$, or $(0, +k)$ in the "motion" context if $I_2 > 0$), and similarly, negative for negative sensory inputs.

Low-dimensional trajectories of the SNNs in Fig.5D were obtained by projecting the low-passed spiking activity onto "color" and "motion"-related axes, defined by $\{f_\mu(\boldsymbol{z}_i)\}_{i=1}^N$ (with $\mu = 1$ for the "color" and $\mu = 2$ for the "motion" axis). Since $f_{\text{out}} = f_1 + f_2$, the "choice" readout of Eq. (25) is equivalent to the projection of this 2-dimensional population response on a diagonal readout axis (Fig.5D,E).

Appendix

Note: in the Appendix, we abandon the boldface notation $\boldsymbol{z}$ to denote positions in the tuning space. In place, the notation z still represents a d-dimensional variable.

A Collective dynamics

A.1 Derivation of the collective dynamics

We start from eq.(13), rewritten here for convenience:

$$\partial_t v_t = -\frac{1}{\tau} v_t + \sum_{\mu=1}^{D} f_\mu \langle g_\mu, \phi \circ v_t \rangle_\rho \tag{26}$$

The recurrent input belongs to the linear subspace $\mathcal{F} = \text{span}\{f_\mu\} \subset \mathcal{H}$. We can decompose the field v_t as its projection onto $\mathcal{F}$, plus a component $v_t^\perp$ belonging to its orthogonal complement $\mathcal{F}^\perp$:

$$v_t = \sum_\mu f_\mu \kappa_\mu(t) + v_t^\perp \tag{27}$$

which implies:

$$\partial_t v_t = \sum_\mu f_\mu \dot{\kappa}_\mu(t) + \partial_t v_t^\perp \tag{28}$$

where the dot $(\dot{\ })$ denotes the time derivative. Together, equations (26)-(27) yield:

$$\partial_t v_t = \sum_\mu f_\mu \left(-\frac{1}{\tau} \kappa_\mu(t) + \langle g_\mu, \phi \circ v_t \rangle_\rho \right) - \frac{1}{\tau} v_t^\perp \tag{29}$$

As both $v_t^\perp$ and $\partial_t v_t^\perp$ are orthogonal to all of the f_μ, we get by identification of eqs.(28) and (29):

$$\dot{\kappa}_\mu(t) = -\frac{1}{\tau} \kappa_\mu(t) + \langle g_\mu, \phi \circ v_t \rangle_\rho \tag{30}$$

$$\partial_t v_t^\perp = -\frac{1}{\tau} v_t^\perp \tag{31}$$

Thus, $v_t^\perp = e^{-t/\tau} h_0^\perp$, so that eq.(27) yields eq.(14). It allows replacing v_t in eq.(30) to obtain, in vector form:

$$\frac{d}{dt}\boldsymbol{\kappa}(t) = -\frac{1}{\tau}\boldsymbol{\kappa}(t) + \langle \boldsymbol{g}, \phi \circ \big(\boldsymbol{f} \cdot \boldsymbol{\kappa}(t) + e^{-t/\tau} h_0^\perp\big)\rangle_\rho \tag{32}$$

which yields eq.(16) if we assume $h_0 \in \mathcal{F}$.

Eventually, the expression of the latent variables $\boldsymbol{\kappa}$ can be obtained by projecting eq.(27) on $\mathcal{F}$: $\langle f_\nu, v_t\rangle_\rho = \sum_\mu \langle f_\nu, f_\mu\rangle_\rho \, \kappa_\mu$; or in vector form: $\langle \boldsymbol{f}, v_t\rangle_\rho = C\boldsymbol{\kappa}$, with the correlation matrix $C_{\nu\mu} := \langle f_\nu, f_\mu\rangle_\rho$. We can assume without loss of generality that the set of functions $\{f_\mu\}$ are linearly independent in $\mathcal{H}$. If that is not the case, a new basis of $\mathcal{F}$ with fewer elements can always be chosen (see also the next section A.2). Thus, the matrix C has full rank D, and is thus invertible, yielding eq.(15):

$$\boldsymbol{\kappa}(t) = C^{-1}\langle \boldsymbol{f}, v_t\rangle_\rho \tag{33}$$

A.2 Including external inputs

Following the arguments of section 2, assume that the network receives an external input composed of K features $I_\nu(t)$ ($\nu = 1, \ldots, K$), to which the tuning of any neuron i is determined by a corresponding function $f_\nu^I(z_i)$ of its position z_i. The external input received by any neuron can be written: $I^{\text{ext}}(t, z_i) = \sum_{\nu=1}^K f_\nu^I(z_i) I_\nu(t)$. The total (recurrent + external) input thus now belongs to the linear subspace $\tilde{\mathcal{F}} \subset \mathcal{H}$ spanned by $\{f_\mu\}_{\mu=1}^D \cup \{f_\nu^I\}_{\nu=1}^K$, of dimension $\tilde{D} \leq D + K$. Let $\{\tilde{f}_\alpha\}_{\alpha=1}^{\tilde{D}}$ be a basis of $\tilde{\mathcal{F}}$. Since any $f_\mu \in \mathcal{F}$ is a linear combination of the $\{\tilde{f}_\alpha\}$, then for each α, there exists a linear combination $\tilde{g}_\alpha$ of the feature maps $\{g_\mu\}$ such that

$$w(z, z') = \sum_{\mu=1}^{D} f_\mu(z) g_\mu(z') = \sum_{\alpha=1}^{\tilde{D}} \tilde{f}_\alpha(z) \tilde{g}_\alpha(z') \tag{34}$$

and similarly, there exists a linear combination $\tilde{I}_\alpha$ of the input features $\{I_\nu\}$ such that

$$I^{\text{ext}}(t, z) = \sum_{\nu=1}^{K} f_\nu^I(z) I_\nu(t) = \sum_{\alpha=1}^{\tilde{D}} \tilde{f}_\alpha(z) \tilde{I}_\alpha(t) \tag{35}$$

Without loss of generality, we can thus write the total synaptic input received at position z as:

$$I(t,z) = \sum_{\alpha=1}^{\tilde{D}} \tilde{f}_\alpha(z)\Big(\int_{\mathcal{M}} \tilde{g}_\alpha(z')\phi(v_t(z'))\rho(dz) \; + \tilde{I}_\alpha(t)\Big) \tag{36}$$

From now on, we omit the tildes ($\tilde{\text{o}}$) for readability. The field equation in the case of a low-dimensional external input can be written:

$$\partial_t v_t = -\frac{1}{\tau} v_t + \sum_\alpha f_\alpha\Big(\langle g_\alpha, \phi \circ v_t\rangle_\rho + I_\alpha(t)\Big) \tag{37}$$

Repeating the derivation of sec.A.1 gives the general expression of the collective dynamics:

$$\frac{d}{dt}\boldsymbol{\kappa}(t) = -\frac{1}{\tau}\boldsymbol{\kappa}(t) + \langle \boldsymbol{g}, \phi \circ \big(\boldsymbol{f}\cdot\boldsymbol{\kappa}(t) + e^{-t/\tau} h_0^{\perp}\big)\rangle_\rho \; + \boldsymbol{I}(t) \tag{38}$$

where $\boldsymbol{I}(t) = (I_1(t), \ldots, I_D(t))$ is the vector of external input features.

A.3 Generalisation to arbitrary L^2 kernels

We turn now to the case of field models where the connectivity is determined by a kernel $w(z,z')$ over $\mathcal{M}^2$ which is only assumed to be in $L^2(\rho\times\rho)$, i.e.

$$\|w\|^2_{2,\rho\times\rho} := \int_{\mathcal{M}^2} \rho(dz)\rho(dz')w(z,z')^2 < \infty \tag{39}$$

yielding the general field equation:

$$\begin{aligned} \partial_t v_t(z) &= -\frac{1}{\tau} v_t(z) + \int_{\mathcal{M}} \rho\left(dz'\right) w(z,z')\phi\left(v_t(z')\right) \\ &= -\frac{1}{\tau} v_t + K_w[\phi\circ v_t] \end{aligned} \tag{40}$$

where $K_w : \mathcal{H} \to \mathcal{H}$ is the Hilbert-Schmidt integral operator associated with the kernel w: [75]

$$K_w[\varphi](z) := \int_{\mathcal{M}} \rho(dz')w(z,z')\varphi(z') \tag{41}$$

Since K_w is a Hilbert-Schmidt operator, it can always be decomposed as:

$$K_w[\varphi] = \sum_{\alpha\in\mathbb{N}} \lambda_\alpha x_\alpha \langle y_\alpha, \varphi\rangle_\rho \tag{42}$$

where the sets of left and right eigenvectors $\{x_\alpha\}, \{y_\alpha\}$ are orthonormal families in $\mathcal{H}$. The l^2-norm of the series of singular values $\lambda_\alpha \geq 0$ is finite, and is equal to the Hilbert-Schmidt norm of K_w, i.e. $\sum_{\alpha\in\mathbb{N}} \lambda_\alpha^2 = \|w\|^2_{2,\rho\times\rho} < \infty$ [75, 76]. The dynamics of general fields, eq.(40), have thus a very similar structure to that of low-rank fields, eq.(12) – except (precisely) that the rank of K_w is not necessarily finite. In particular, one can still write the associated collective dynamics for the projections $\kappa_\alpha(t) = \langle x_\alpha, u_t\rangle_\rho$, following the line of sec.A.1:

$$\begin{aligned} \frac{d}{dt}\kappa_\alpha(t) &= -\frac{1}{\tau}\kappa_\alpha(t) + \lambda_\alpha \langle y_\alpha, \phi\circ v_t\rangle_\rho \qquad (\alpha\in I) \\ v_t &= \sum_{\alpha\in I} x_\alpha \kappa_\alpha(t) + e^{-t/\tau} v_0^\perp \end{aligned} \tag{43}$$

where $I = \{\alpha \mid \lambda_\alpha > 0\} \subseteq \mathbb{N}$, and $v_0^\perp = v_0 - \sum_\alpha x_\alpha \kappa_\alpha(0)$. From this point of view, a low-rank field is only characterised by a finite set of non-zero singular values $\lambda_\alpha > 0$. The dynamics of any arbitrary field can be approximated using a truncation of the sum in eq.(42) [57], which corresponds to a model of low-rank network according to our framework (see eq.(10)).

B Examples of collective dynamics

B.1 Fixed points in a generalised Hopfield model

We consider here the connectivity mappings given by eq.(20), which read:

$$\boldsymbol{f}(z) = \begin{pmatrix} [z]_+ + \cos\alpha [z]_- \\ -\sin\alpha [z]_- \end{pmatrix} = \begin{pmatrix} 1 & \cos\alpha \\ 0 & -\sin\alpha \end{pmatrix} \begin{pmatrix} [z]_+ \\ [z]_- \end{pmatrix} =: \mathbf{A}\,\tilde{\boldsymbol{f}}(z)\ ;\ \boldsymbol{g} = J\boldsymbol{f} \tag{44}$$

with $\tilde{\boldsymbol{f}}(z) = ([z]_+, [z]_-)$, $\mathbf{A}$ the 2×2 matrix above, and $[z]_\pm = z \cdot \mathbb{1}_{\{\mathrm{sgn}(z)=\pm 1\}}$. Let the neurons be normally distributed over the real line, i.e. $z_i \sim \mathcal{N}(0,1)$. The collective dynamics of eq.(16) write:

$$(\frac{d}{dt} + \frac{1}{\tau})\boldsymbol{\kappa} = \int_{\mathbb{R}} Dz\, J\mathbf{A}\tilde{\boldsymbol{f}}(z)\phi(\boldsymbol{\kappa}\mathbf{A}\tilde{\boldsymbol{f}}(z)) \tag{45}$$

with Dz the standard Gaussian measure; and we assume for now a general transfer function ϕ. We make the following change of variable: $\boldsymbol{\lambda} = \mathbf{A}^T \boldsymbol{\kappa}$. Noticing that $(\frac{d}{dt} + \frac{1}{\tau})\boldsymbol{\lambda} = \mathbf{A}^T(\frac{d}{dt} + \frac{1}{\tau})\boldsymbol{\kappa}$, we have:

$$\begin{aligned}(\frac{d}{dt} + \frac{1}{\tau})\boldsymbol{\lambda} &= \int_{\mathbb{R}} Dz\, J\mathbf{A}^T\mathbf{A}\tilde{\boldsymbol{f}}(z)\phi(\boldsymbol{\lambda}\cdot\tilde{\boldsymbol{f}}(z)) \\ &= J\mathbf{A}^T\mathbf{A}\int_{\mathbb{R}} Dz \begin{pmatrix}[z]_+ \\ [z]_-\end{pmatrix}\phi(\boldsymbol{\lambda}\cdot\tilde{\boldsymbol{f}}(z))\end{aligned} \tag{46}$$

Let r, θ be the polar coordinates of $\boldsymbol{\lambda}$. From the expression of $(\lambda_1, \lambda_2) = r(\cos\theta, \sin\theta)$, we get:

$$\boldsymbol{\lambda}\cdot\tilde{\boldsymbol{f}}(z) = \begin{cases} rz\cos\theta \ \ (z \geq 0) \\ rz\sin\theta \ \ (z < 0)\end{cases} \tag{47}$$

so that eq.(46) can be written:

$$(\frac{d}{dt}+\frac{1}{\tau})\boldsymbol{\lambda} = J\mathbf{A}^T\mathbf{A}\left[\begin{pmatrix}1\\0\end{pmatrix}\int_0^\infty Dz\, z\,\phi(rz\cos\theta) + \begin{pmatrix}0\\1\end{pmatrix}\int_{-\infty}^0 Dz\, z\,\phi(rz\sin\theta)\right] \tag{48}$$

We now derive further the dynamics for the step transfer function $\phi(x) = R\mathbb{1}_{\{x>0\}}$. Notice that (i) for $z > 0$, $rz\cos\theta > 0 \Leftrightarrow \lambda_1 > 0$; and (ii) for $z < 0$, $rz\sin\theta > 0 \Leftrightarrow \lambda_2 < 0$. The above gives:

$$(\tau\frac{d}{dt} + 1)\boldsymbol{\lambda} = \tau JR\langle[z]_+\rangle\mathbf{A}^T\mathbf{A}\left[\begin{pmatrix}1\\0\end{pmatrix}\mathbb{1}_{\{\lambda_1>0\}} - \begin{pmatrix}0\\1\end{pmatrix}\mathbb{1}_{\{\lambda_2<0\}}\right] \tag{49}$$

where $\langle[z]_+\rangle = \int_0^\infty Dz\, z = 1/\sqrt{2\pi}$. Let us define $k = \tau JR\langle[z]_+\rangle$, and:

$$\boldsymbol{\lambda}^{(+)} = k\mathbf{A}^T\mathbf{A}\begin{pmatrix}1\\0\end{pmatrix} = k\begin{pmatrix}1\\ \cos\alpha\end{pmatrix} \quad ; \quad \boldsymbol{\lambda}^{(-)} = -k\mathbf{A}^T\mathbf{A}\begin{pmatrix}0\\1\end{pmatrix} = k\begin{pmatrix}-\cos\alpha\\ -1\end{pmatrix} \tag{50}$$

We can write:

$$
\begin{aligned}
(\tau\frac{d}{dt}+1)\boldsymbol{\lambda} &= \boldsymbol{\lambda}^{(+)}\mathbb{1}_{\{\lambda_1>0\}}+\boldsymbol{\lambda}^{(-)}\mathbb{1}_{\{\lambda_2<0\}} \\
&= \begin{cases} \boldsymbol{\lambda}^{(+)} & (\boldsymbol{\lambda}\in Q_1) \\ 0 & (\boldsymbol{\lambda}\in Q_2) \\ \boldsymbol{\lambda}^{(-)} & (\boldsymbol{\lambda}\in Q_3) \\ \boldsymbol{\lambda}^{(+)}+\boldsymbol{\lambda}^{(-)} & (\boldsymbol{\lambda}\in Q_4) \end{cases}
\end{aligned} \tag{51}
$$

with $Q_{1\leq i\leq 4}$ the four quadrants of $\mathbb{R}^2$, numbered in trigonometric order. For $\alpha\in]0,\pi/2[$, $\boldsymbol{\lambda}^{(+)}\in Q_1$, $\boldsymbol{\lambda}^{(-)}\in Q_3$ and $(\boldsymbol{\lambda}^{(+)}+\boldsymbol{\lambda}^{(-)})\in Q_4$. They are thus stable fixed points of the dynamics, while $\boldsymbol{\lambda}=0$ is a saddle point. The dynamics of the original variables $\boldsymbol{\kappa}$, shown in fig.3(B), are obtained through the change of coordinates $\boldsymbol{\kappa}=(\mathbf{A}^T)^{-1}\boldsymbol{\lambda}$. In particular, the three corresponding fixed points are:

$$
\boldsymbol{\kappa}^{(+)}=k\begin{pmatrix}1\\0\end{pmatrix} \quad ; \quad \boldsymbol{\kappa}^{(-)}=k\begin{pmatrix}-\cos\alpha\\+\sin\alpha\end{pmatrix}\approx k\begin{pmatrix}-1/2\\0.87\end{pmatrix} \tag{52}
$$

and $(\boldsymbol{\kappa}^{(+)}+\boldsymbol{\kappa}^{(-)})$. The numerical values correspond to Fig.3(B), where $\alpha=\pi/3$.

B.2 Limit cycles in two different models

Ring model.

For the ring model, we have from eq.(4):

$$
\boldsymbol{f}(z)=\begin{pmatrix}\cos z\\ \sin z\end{pmatrix} \quad ; \quad \boldsymbol{g}=J\boldsymbol{f}(z+\delta) \tag{53}
$$

and the collective dynamics resulting from eq.(16) write:

$$
(\frac{d}{dt}+\frac{1}{\tau})\boldsymbol{\kappa}=\int_0^{2\pi}\frac{dz}{2\pi}\boldsymbol{g}(z)\phi(\boldsymbol{f}(z)\cdot\boldsymbol{\kappa}) \tag{54}
$$

Let r,θ be the polar coordinates of $\boldsymbol{\kappa}$. In the polar coordinate system $(\hat{e}_r,\hat{e}_\theta)$ associated with $\boldsymbol{\kappa}$, we have:

$$
(\frac{d}{dt}+\frac{1}{\tau})\boldsymbol{\kappa}=(\dot{r}+\frac{1}{\tau}r)\hat{e}_r+r\dot{\theta}\hat{e}_\theta \quad ; \quad \boldsymbol{f}(z)=\cos(z-\theta)\hat{e}_r+\sin(z-\theta)\hat{e}_\theta \tag{55}
$$

hence:

$$\begin{aligned}\begin{pmatrix}\dot{r}+\frac{1}{\tau}r\\ r\dot{\theta}\end{pmatrix} &= J\int_0^{2\pi}\frac{dz}{2\pi}\begin{pmatrix}\cos(z+\delta-\theta)\\ \sin(z+\delta-\theta)\end{pmatrix}\phi(r\cos(z-\theta))\\ &= J\int_0^{2\pi}\frac{dz}{2\pi}\begin{pmatrix}\cos(z+\delta)\\ \sin(z+\delta)\end{pmatrix}\phi(r\cos z)\end{aligned} \tag{56}$$

For the step transfer function $\phi(x) = R\mathbb{1}_{\{x>0\}}$, we have $\phi(r\cos z) = R\mathbb{1}_{\{-\pi/2<z<\pi/2\}}$, hence:

$$\begin{pmatrix}\dot{r}+\frac{1}{\tau}r\\ r\dot{\theta}\end{pmatrix} = \frac{JR}{2\pi}\int_{-\pi/2}^{\pi/2}dz\begin{pmatrix}\cos(z+\delta)\\ \sin(z+\delta)\end{pmatrix} = \frac{JR}{\pi}\begin{pmatrix}\cos\delta\\ \sin\delta\end{pmatrix} \tag{57}$$

which is identical to eq.(21).

Gaussian positions.

Let the neurons being normally distributed in $\mathbb{R}^2$, $z_i =\sim \mathcal{N}(0, \mathbb{1}_2)$. As discussed in the Results section 2, the distribution can be parameterised in polar coordinates: each neuron is assigned a radius $\xi_i \sim \chi_2$ and an angle $\varphi_i \sim \mathcal{U}([0, 2\pi[)$. The connectivity of eq.(5) thus writes:

$$W_{ij} = \tilde{J} z_i R^{(\delta)} z_j = \tilde{J}\begin{pmatrix}\xi_i\cos\varphi_i\\ \xi_i\sin\varphi_i\end{pmatrix}\begin{pmatrix}\xi_j\cos(\varphi_j+\delta)\\ \xi_j\sin(\varphi_j+\delta)\end{pmatrix} \tag{58}$$

which gives the feature maps, as functions of a neuron's radius and angle:

$$\tilde{\boldsymbol{f}}(\xi,\varphi) = \begin{pmatrix}\xi\cos\varphi\\ \xi\sin\varphi\end{pmatrix} = \xi\boldsymbol{f}(\varphi) \quad ; \quad \tilde{\boldsymbol{g}}(\xi,\varphi) = \tilde{J}\tilde{\boldsymbol{f}}(\xi,\varphi+\delta) = \xi\frac{\tilde{J}}{J}\boldsymbol{g}(\varphi) \tag{59}$$

where $\boldsymbol{f}, \boldsymbol{g}$ are those of eq.(53). The collective dynamics of eq.(16) thus write:

$$(\frac{d}{dt}+\frac{1}{\tau})\boldsymbol{\kappa} = \frac{\tilde{J}}{J}\int_0^{2\pi}\frac{d\varphi}{2\pi}\int_{\mathbb{R}^+}D\xi\ \xi\boldsymbol{g}(\varphi)\phi(\xi\boldsymbol{f}(\varphi)\cdot\boldsymbol{\kappa}) \tag{60}$$

where $D\xi$ denotes the χ_2-probability measure. Notice that for the step transfer function, $\phi(\xi x) = \phi(x)\ \forall\xi > 0$, hence:

$$(\frac{d}{dt}+\frac{1}{\tau})\boldsymbol{\kappa} = \frac{\tilde{J}}{J}\Big(\int_{\mathbb{R}^+}\xi D\xi\Big)\int_0^{2\pi}\frac{d\varphi}{2\pi}\ \boldsymbol{g}(\varphi)\phi(\boldsymbol{f}(\varphi)\cdot\boldsymbol{\kappa}) \tag{61}$$

which is identical to eq.(54), up to a factor $\frac{\tilde{J}}{J}\langle\xi\rangle$ on the RHS, where $\langle\xi\rangle := \int \xi D\xi$ ($= \sqrt{\pi/2}$ for the χ_2-distribution). The rescaling factor equals 1 if $J = \langle\xi\rangle\tilde{J}$; hence, the two models' collective dynamics are identical up to a rescaling of the connectivity.

Notice that the derivation is not restricted to a normal distribution over $\mathbb{R}^2$, but applies to any rotationally invariant distribution; in this case, the only difference will be the distribution of the radii ξ_i, which will affect the rescaling factor $\langle\xi\rangle$.

General transfer function

Consider the collective dynamics (Eq.(16)) for a general transfer function, corresponding to the connectivity functions of Eq.(59):

$$(\frac{d}{dt} + \frac{1}{\tau}\boldsymbol{\kappa}) = \tilde{J}\int_0^{2\pi}\frac{d\varphi}{2\pi}\int_{\mathbb{R}^+} D\xi\ \xi\boldsymbol{f}(\varphi+\delta)\phi(\xi\boldsymbol{f}(\varphi)\cdot\boldsymbol{\kappa}) \tag{62}$$

For the radial coordinate (r,θ) of $\boldsymbol{\kappa}$, it writes (using Eq.(55)):

$$\begin{aligned}\begin{pmatrix}\dot{r}+\frac{1}{\tau}r\\ r\dot{\theta}\end{pmatrix} &= \tilde{J}\int_0^{2\pi}\frac{d\varphi}{2\pi}\int_{\mathbb{R}^+} D\xi\ \xi\begin{pmatrix}\cos(\varphi+\delta-\theta)\\ \sin(\varphi+\delta-\theta)\end{pmatrix}\phi(\xi r\cos(\varphi-\theta))\\ &= \tilde{J}\int_0^{2\pi}\frac{d\varphi}{2\pi}\int_{\mathbb{R}^+} D\xi\ \xi\begin{pmatrix}\cos(\varphi+\delta)\\ \sin(\varphi+\delta)\end{pmatrix}\phi(\xi r\cos(\varphi))\\ &= \tilde{J}\begin{pmatrix}\cos\delta\\ \sin\delta\end{pmatrix}\cdot\Phi(r)\end{aligned} \tag{63}$$

where $\Phi(r)$ is given by:

$$\Phi(r) := \int_{-1}^{1}\frac{dx}{\pi\sqrt{1-x^2}}\int_{\mathbb{R}^+} D\xi\ \xi x\,\phi(r\xi x) \tag{64}$$

Thus, the model generates rotational dynamics that converge towards a limit cycle of radius r_0, implicitly defined by $r_0 = \tau\tilde{J}\cos\delta\Phi(r_0)$ (provided that a solution exists).

Radius parameter as a heterogeneity

The radius ξ_i of each neuron can be interpreted as a heterogeneity, which affects both the connectivity and the transfer functions of the neurons. The connectivity of Eq. (5) reads $W_{ij} = \xi_i\xi_j w^{\text{ring}}(\varphi_i,\varphi_j)$, where w^{ring} denotes the connectivity kernel of the ring model (Eq.(4)). On the other

hand, given the polar coordinates (ρ, θ) of the latent variable $\boldsymbol{\kappa}$, each neuron's firing rate is given by $r_i(t) = \phi(\xi_i \cdot v_t^{\text{ring}}(\varphi_i))$, where $v_t^{\text{ring}}(\varphi_i) = \rho \cos(\varphi_i - \theta)$ is the potential of a neuron at angle φ_i in the initial ring model.

C Symmetries of the dynamics

We present here the results of Sec.2, relating the symmetries of the field model to those of the collective dynamics. We take an operator theory perspective. The dynamics of the field, eq.(13), can be written:

$$\Box v_t := \left(\frac{d}{dt} + \frac{1}{\tau}\right) v_t = \mathcal{D}^{\mathcal{H}} v_t \tag{65}$$

where $\Box = \left(\frac{d}{dt} + \frac{1}{\tau}\right)$ is linear, and $\mathcal{D}^{\mathcal{H}} : \mathcal{H} \to \mathcal{F}$ (where $\mathcal{F} = \text{span}\{f_\mu\}$) is a nonlinear operator defined by:

$$\mathcal{D}^{\mathcal{H}} : \varphi \mapsto \sum_\mu f_\mu \langle g_\mu, \phi \circ \varphi \rangle_\rho \tag{66}$$

Symmetries.

Let $S : \mathcal{M} \to \mathcal{M}$ be a bijection such that $\varphi \circ S \in \mathcal{H}$, $\forall \varphi \in \mathcal{H}$. For such a mapping S, $L^{(S)} : \varphi \mapsto \varphi \circ S$ defines a linear operator on $\mathcal{H}$. We call S a *symmetry* of a field model if the operator $\mathcal{D}^{\mathcal{H}}$ describing the dynamics commutes with $L^{(S)}$; i.e. $\mathcal{D}^{\mathcal{H}} \circ L^{(S)} = L^{(S)} \circ \mathcal{D}^{\mathcal{H}}$. It implies the equivariance of the field dynamics for $L^{(S)}$: if v_t solves the dynamics of eq.(65), then $L^{(S)} v_t$ is also a solution to the dynamics, with initial condition $L^{(S)} v_0$.

Similarly, the collective dynamics of eq.(16) are described by a nonlinear operator $\mathcal{D}^D$ acting on $\mathbb{R}^D$:

$$\begin{aligned} \Box \boldsymbol{\kappa} &= \mathcal{D}^D \boldsymbol{\kappa} \\ \text{with} \quad \mathcal{D}^D &: \boldsymbol{\kappa} \mapsto \langle \boldsymbol{g}, \phi \circ (\boldsymbol{f} \cdot \boldsymbol{\kappa}) \rangle_\rho \end{aligned} \tag{67}$$

and a symmetry of the collective dynamics is a mapping $T : \mathbb{R}^D \to \mathbb{R}^D$ that commutes with $\mathcal{D}^D$.

Theorem 1. *1. If (i) S is a symmetry of the field model, and (ii) $\mathcal{F} = \text{span}\{f_\mu\}$ is an invariant subspace of $L^{(S)}$, then a symmetry of the collective dynamics is given by the linear operator $T^{(S)} : \mathbb{R}^D \to \mathbb{R}^D$,*

defined by the matrix

$$T^{(S)} = M(\langle f_\mu, f_\nu \circ S\rangle_\rho) \tag{68}$$

where $M^{-1} = (\langle f_\mu, f_\nu\rangle_\rho)$*, and* $(a_{\mu\nu})$ *denotes the* $D\times D$ *matrix with entries* $a_{\mu\nu}$*.*

2. If both the distribution ρ *and the connectivity kernel* $w(z,z') = \boldsymbol{f}(z)\cdot\boldsymbol{g}(z')$ *are invariant for* S*, then the requirements (i) and (ii) of point 1 are satisfied.*

The proof can be found at the end of this section.

Left and right symmetry groups.

Let $S : (g,z) \mapsto S_g(z)$ denote the action of a group G on $\mathcal{M}$. G is a *symmetry group* of the field model if for any $g \in G$, S_g is a symmetry. We call it a *left* symmetry group if S is a left group action (for which $S_{gh} = S_g \circ S_h$, i.e. $S_{gh}(z) = S_g(S_h(z))$, $\forall g,h \in G$); and a *right* symmetry group for a right group action (for which $S_{gh} = S_h \circ S_g$). Similarly, a symmetry group of the collective dynamics is a group acting on $\mathbb{R}^D$ whose action of any element is a symmetry of the collective dynamics.

Corollary 1. *Let* G *be a* left *symmetry group of the field model, whose action* S_g *of each element* g *on* $\mathcal{M}$ *satisfies the requirements of Theorem 1.1. Then,* G *is a* right *symmetry group of the collective dynamics, with a linear group action on* $\mathbb{R}^D$ *given by* $(g,\boldsymbol{\kappa}) \mapsto T^{(S_g)}\boldsymbol{\kappa}$*.*

Proofs.

In the following, $S : \mathcal{M} \to \mathcal{M}$ denotes a bijection satisfying $\varphi \circ S \in \mathcal{H}$, $\forall \varphi \in \mathcal{H}$; so that $L^{(S)} : \mathcal{H} \to \mathcal{H} : \varphi \mapsto \varphi \circ S$ is well defined.

We start by defining the surjection $K : \mathcal{H} \to \mathbb{R}^D$ that maps a field to the components of its projection on $\mathcal{F}$, and the reciprocal injection $V : \mathbb{R}^D \to \mathcal{F} \subset \mathcal{H}$:

$$K : \varphi \mapsto M\langle \boldsymbol{f}, \varphi\rangle_\rho \quad ; \quad V : \boldsymbol{k} \mapsto \boldsymbol{k}\cdot\boldsymbol{f} \tag{69}$$

with $M^{-1} = (\langle f_\mu, f_\nu\rangle_\rho)$ the inverse correlation matrix. They satisfy the following properties:

$$KV = \mathbb{1}_{\mathbb{R}^D} \quad ; \quad VK = \Pi_{\mathcal{F}} \tag{70}$$

where $\Pi_{\mathcal{F}} : \varphi \mapsto \sum_{\mu,\nu} M_{\mu\nu} f_\mu \langle f_\nu, \varphi \rangle_\rho$ is the orthogonal projection on $\mathcal{F}$ (by definition of the matrix M). Thus, $K : v \mapsto \boldsymbol{\kappa}$ and $V : \boldsymbol{\kappa} \mapsto v$ establish a bijection between $\mathcal{F}$ and $\mathbb{R}^D$.

We start with a few useful lemmas.

Lemma 1. $\mathcal{D}^D = K\mathcal{D}^{\mathcal{H}}V$.

Proof. By setting $\boldsymbol{\kappa}(t) = Kv_t$ and $v_t = V\boldsymbol{\kappa}(t)$, and noticing that $\Box = (\partial_t - 1/\tau)$ commutes with K, one obtains from eq.(65) the following expression for the dynamics of $\boldsymbol{\kappa}$:

$$\Box\boldsymbol{\kappa} = \Box K v_t = K\Box v_t = K\mathcal{D}^{\mathcal{H}} v_t = K\mathcal{D}^{\mathcal{H}} V\boldsymbol{\kappa}(t) \tag{71}$$

which can be identified with eq.(67)). □

Lemma 2. *If $\mathcal{F}$ is an invariant subspace of an operator L on $\mathcal{H}$, then $LV = \Pi_{\mathcal{F}} LV$.*

Proof. For any $\varphi \in \mathcal{H}, V\varphi \in \mathcal{F}$. $\mathcal{F}$ is an invariant subspace of L iff $Lv \in \mathcal{F}\ \forall v \in \mathcal{F}$. Then $\forall \varphi \in \mathcal{H}$, $LV\varphi \in \mathcal{F}$, implying $\Pi_{\mathcal{F}} LV\varphi = LV\varphi$. □

Lemma 3. *If ρ is invariant for the bijection S, then $(L^{(S)})^\dagger = L^{(S^{-1})}$.*

Proof. $L^{(S)^\dagger}$ is defined by: $\langle \varphi, L^{(S)}\psi \rangle_\rho = \langle L^{(S)^\dagger}\varphi, \psi \rangle_\rho$ for all $\varphi, \psi \in \mathcal{H}$. If ρ is invariant for S, i.e. $\rho \circ S^{-1} = \rho$, then for all $\varphi, \psi \in \mathcal{H}$,

$$\begin{aligned}
\langle \varphi, L^{(S)}\psi \rangle_\rho = \langle \varphi, \psi \circ S \rangle_\rho &= \int_{\mathcal{M}} \varphi(z)\psi(Sz)\rho(dz) \\
&= \int_{\mathcal{M}} \varphi(S^{-1}z)\psi(z)\,(\rho \circ S^{-1})(dz) = \int_{\mathcal{M}} \varphi(S^{-1}z)\psi(z)\,\rho(dz) \\
&= \langle \varphi \circ S^{-1}, \psi \rangle_\rho = \langle L^{(S^{-1})}\varphi, \psi \rangle_\rho
\end{aligned} \tag{72}$$

where we used the change of variable formula in the third equality. □

Proof of Theorem 1.

Point 1.

Let L be a linear operator on $\mathcal{H}$ that satisfies: (i) $\mathcal{D}^{\mathcal{H}} L = L\mathcal{D}^{\mathcal{H}}$, and (ii) $\mathcal{F}$ is an invariant subspace of L. We have from (i):

$$K(L\mathcal{D}^{\mathcal{H}})V = K(\mathcal{D}^{\mathcal{H}} L)V \tag{73}$$

The left-hand side can be written:

$$KL\mathcal{D}^{\mathcal{H}}V = KL(VK\mathcal{D}^{\mathcal{H}})V = (KLV)(K\mathcal{D}^{\mathcal{H}}V) = T\mathcal{D}^D \tag{74}$$

with $T := KLV$. The first equality is obtained by noticing that the image of $\mathcal{D}^{\mathcal{H}}$ is in $\mathcal{F}$, so that $\Pi_{\mathcal{F}}\mathcal{D}^{\mathcal{H}} = \mathcal{D}^{\mathcal{H}}$; combined with $VK = \Pi_{\mathcal{F}}$, eq.(70); and we used Lemma 1 in the last equality.

In turn, the right-hand side writes:

$$K\mathcal{D}^{\mathcal{H}}(LV) = K\mathcal{D}^{\mathcal{H}}(VKLV) = (K\mathcal{D}^{\mathcal{H}}V)(KLV) = \mathcal{D}^{D}T \tag{75}$$

where we used Lemma 2, with $VK = \Pi_{\mathcal{F}}$.

The operator $T = KLV$ thus commutes with the operator $\mathcal{D}^D$: it is a symmetry of the collective dynamics. Given a transformation S such that $L = L^{(S)}$, we can write, for all $\boldsymbol{\kappa} \in \mathbb{R}^D$:

$$\begin{aligned}(T\boldsymbol{\kappa})_\mu &= \sum_\nu M_{\mu\nu}\langle f_\nu, L^{(S)}(\sum_\sigma f_\sigma\kappa_\sigma)\rangle_\rho = \sum_\nu M_{\mu\nu}\langle f_\nu, \sum_\sigma \kappa_\sigma(L^{(S)}f_\sigma)\rangle_\rho \\ &= \sum_\sigma\sum_\nu M_{\mu\nu}\langle f_\nu, f_\sigma\circ S\rangle_\rho\kappa_\sigma =: \sum_\sigma T^{(S)}_{\mu\sigma}\,\kappa_\sigma\end{aligned} \tag{76}$$

which yields $T^{(S)}_{\mu\sigma} = \sum_\nu M_{\mu\nu}\langle f_\nu, f_\sigma\circ S\rangle_\rho$. We used the linearity of $L^{(S)}$ in the second equality.

Point 2.

Let both ρ and the connectivity w be invariant for S (i.e. $w(z,z') = w(Sz,Sz')\ \forall z,z' \in \mathcal{M}$). We first show that it implies: $\mathcal{D}^{\mathcal{H}}L^{(S)} = L^{(S)}\mathcal{D}^{\mathcal{H}}$. For all $v \in \mathcal{H}$ and $z \in \mathcal{M}$, we have:

$$\begin{aligned}(\mathcal{D}^{\mathcal{H}}L^{(S)}v)(z) &= \sum_\mu f_\mu(z)\langle g_\mu, \phi\circ(L^{(S)}v)\rangle_\rho \\ &= \sum_\mu f_\mu(z)\langle g_\mu, \phi\circ v\circ S\rangle_\rho = \sum_\mu f_\mu(z)\langle g_\mu, L^{(S)}(\phi\circ v)\rangle_\rho \\ &= \sum_\mu f_\mu(z)\langle {L^{(S)}}^\dagger g_\mu, \phi\circ v\rangle_\rho \overset{(*)}{=} \sum_\mu f_\mu(z)\langle g_\mu\circ S^{-1}, \phi\circ v\rangle_\rho \\ &= \int_{\mathcal{M}} w(z,S^{-1}z')\phi\circ v(z)\rho(dz) \overset{(**)}{=} \int_{\mathcal{M}} w(Sz,z')\phi\circ v(z)\rho(dz) \\ &= \sum_\mu f_\mu(Sz)\langle g_\mu, \phi\circ v\rangle_\rho = \sum_\mu (L^{(S)}f_\mu)(z)\langle g_\mu, \phi\circ v\rangle_\rho \\ &= (L^{(S)}\mathcal{D}^{\mathcal{H}}v)(z)\end{aligned} \tag{77}$$

where we used Lemma 3 in $(*)$, and the invariance of w in $(**)$.

We now show that it implies that $\mathcal{F}$ is invariant under $L^{(S)}$. For this, it is sufficient to show that the images of the basis functions $L^{(S)} f_\mu \in \mathcal{F},\ \forall \mu = 1, \dots, D$. We have:

$$\begin{aligned} w \text{ invariant for } S \ &\Leftrightarrow \ \sum_\mu f_\mu(Sz) g_\mu(Sz') = \sum_\mu f_\mu(z) g_\mu(z') \quad \forall z, z' \in \mathcal{M} \\ &\Leftrightarrow \ \sum_\mu (L^{(S)} f_\mu)(z)(L^{(S)} g_\mu)(z') = \sum_\mu f_\mu(z) g_\mu(z') \quad \forall z, z' \\ &\overset{(*)}{\Rightarrow} \ \sum_\mu (L^{(S)} f_\mu)(z) \langle L^{(S)} g_\nu, L^{(S)} g_\mu \rangle_\rho = \sum_\mu f_\mu(z) \langle L^{(S)} g_\nu, g_\mu \rangle_\rho \quad \forall z, \forall \nu \\ &\Rightarrow \ \sum_\mu G_{\nu\mu} L^{(S)} f_\mu = \sum_\mu f_\mu \langle L^{(S)} g_\nu, g_\mu \rangle_\rho \in \mathcal{F} \quad \forall \nu = 1, \dots, D \end{aligned} \tag{78}$$

where $(*)$ comes by projection of the functions of z' on $L^{(S)} g_\nu$; and $G_{\nu\mu} := \langle L^{(S)} g_\nu, L^{(S)} g_\mu \rangle_\rho = \langle g_\nu, g_\mu \rangle_\rho$ from the invariance of ρ and Lemma 3.

Without loss of generality, one can assume that the functions $\{g_\mu\}$ are linearly independent. If not verified, one can always choose a new basis of $\mathcal{F}$ such that this is the case. Thus, the matrix G has full rank, and is therefore invertible. The last line of eq.(78) yields:

$$L^{(S)} f_\mu = \sum_\nu (G^{-1})_{\mu\nu} \sum_\sigma \langle L^{(S)} g_\nu, g_\sigma \rangle_\rho \, f_\sigma \in \mathcal{F} \quad \forall \mu = 1, \dots, D \tag{79}$$

i.e. $L^{(S)} f_\mu \in \mathcal{F},\ \forall \mu = 1, \dots, D$, which concludes the proof.

Proof of Corollary 1.

In the following, we use the shorthand notation $L^{(S_g)} = L^{(g)}$ and $T^{(S_g)} = T^{(g)}$. To prove the corollary, it is sufficient to show that $T^{(gh)} = T^{(h)} T^{(g)}$, $\forall g, h \in G$; and that $T^{(e)} = \mathbb{1}_{\mathbb{R}^D}$ for the identity $e \in G$.

Given that S is a left group action, we have $S_{gh} = S_g \circ S_h$; and thus, $L^{(gh)} = L^{(S_{gh})} = L^{(S_g \circ S_h)}$. We have by associativity of the composition: $L^{(S_g \circ S_h)} \varphi = \varphi \circ (S_g \circ S_h) = (\varphi \circ S_g) \circ S_h = L^{(h)}(L^{(g)} \varphi)$, for all $\varphi \in \mathcal{H}$; hence: $L^{(gh)} = L^{(h)} L^{(g)}$. Thus, $(g, \varphi) \mapsto L^{(g)} \varphi$ defines a *right* group action of G on $\mathcal{H}$.

By recalling $T^{(S)} = K L^{(S)} V$, we get:

$$T^{(gh)} = K L^{(gh)} V = K L^{(h)} L^{(g)} V = K L^{(h)} (VK) L^{(g)} V = T^{(h)} T^{(g)} \tag{80}$$

where we used Lemma 2 in the third equality, with $VK = \Pi_{\mathcal{F}}$ (eq(70)).

Notice eventually that the identity e yields identity group action: we have $S_e = \mathbb{1}_{\mathcal{M}} \;\Rightarrow\; L^{(e)} = \mathbb{1}_{\mathcal{H}} \;\Rightarrow\; T^{(e)} = K\mathbb{1}_{\mathcal{H}}V = \mathbb{1}_{\mathbb{R}^D}$ according to eq.(70).

D Adding heterogeneity in a field model

Rescaling of outgoing weights

Consider a field model, obeying the low-dimensional dynamics of Eq.(16), in which the outgoing weights of each neuron i are rescaled by an heterogeneity parameter $\xi_i > 0$, drawn i.i.d. across neurons according to a distribution ρ_ξ. In the resulting "new" field model, each neuron's position in the "new" tuning space now contains this heterogeneity parameter; we denote it

$$\tilde{z}_i = (z_i, \xi_i) \in \mathcal{M} \times \mathbb{R}_+ =: \tilde{\mathcal{M}} \tag{81}$$

and the "new" connectivity reads:

$$\tilde{W}_{ij} = \tilde{\boldsymbol{f}}(\tilde{z}_i) \cdot \tilde{\boldsymbol{g}}(\tilde{z}_j) = \xi_j W_{ij} = \boldsymbol{f}(z_i) \cdot \xi_j \boldsymbol{g}(z_j) \tag{82}$$

which allows us to identify $\tilde{\boldsymbol{f}}(\tilde{z}_i) = \boldsymbol{f}(z_i)$, and $\tilde{\boldsymbol{g}}(\tilde{z}_j) = \xi_j \boldsymbol{g}(z_j)$. The collective dynamics, Eq.(16), of the latent variables $\tilde{\boldsymbol{\kappa}}$ associated with this "new" model with rescaled connectivity write:

$$(\frac{d}{dt} + \frac{1}{\tau})\tilde{\boldsymbol{\kappa}} = \Big(\int \xi \rho_\xi(d\xi) \Big) \cdot \langle \boldsymbol{g}, \phi \circ (\boldsymbol{f} \cdot \tilde{\boldsymbol{\kappa}}) \rangle_\rho \tag{83}$$

which is identical to the dynamics of the original model, up to a rescaling of the latent variables by a factor $\int \xi \rho_\xi(d\xi)$, which represents the mean value of the heterogeneity parameter ξ.

Heterogeneous maximum firing rates

In a similar fashion, starting from an original field model, each neuron's firing rate can be rescaled by a parameter $a_i > 0$ drawn according to a distribution ρ_a. The firing rate of any neuron i, given its membrane potential h_i, is now given by $r_i = a_i \phi(h_i)$. The transfer function $\tilde{\phi}$ of the "new" model thus depends on each neuron's "new" position $\tilde{z}_i = (z_i, a_i)$: we write $\tilde{\phi}(h, \tilde{z}_i) = a_i \phi(h)$. Similar to the example above ("Rescaling of output weights"), the collective dynamics of the "new" and of the original field models are identical, up to a rescaling by the mean value of the heterogeneity parameter $\int a \rho_a(da)$.

E Field models as universal approximators

E.1 Gaussian mixture low-rank networks

Gaussian mixture low-rank networks are a class of low-rank RNNs recently introduced to study how the collective dynamics are shaped by the statistics of the connectivity vectors [47, 49]. In this model, the rank-D interaction between neurons i, j is given by $\boldsymbol{m}_i \cdot \boldsymbol{n}_j$, where the connectivity vectors $\boldsymbol{m}_i, \boldsymbol{n}_i \in \mathbb{R}^D$ associated with each neuron i are drawn according to the Gaussian mixture of P populations:

$$z_i = (\boldsymbol{m}_i, \boldsymbol{n}_i) \sim \rho = \sum_{p=1}^{P} \alpha_p \rho^{(p)} \tag{84}$$

where for each $p = 1 \dots, P$, $\rho^{(p)}$ is a multivariate Gaussian over $\mathbb{R}^{2D}$; and $\sum_{p=1}^{P} \alpha_p = 1$. Here, each neuron's position $z_i \in \mathbb{R}^{2D}$ corresponds to the concatenation of the connectivity vectors, and one can directly identify $\boldsymbol{f}(\boldsymbol{m}, \boldsymbol{n}) = \boldsymbol{m}$, $\boldsymbol{g}(\boldsymbol{m}, \boldsymbol{n}) = \boldsymbol{n}$.

These RNNs have been shown to be universal approximators of arbitrary D-dimensional collective dynamics, if the number P of populations can be chosen arbitrarily [47].

E.2 Field models over one-dimensional space

For the sake of simplicity, we focus on the case of bounded one-dimensional tuning space $\mathcal{M} = [0, 1]$, with uniform distribution $\rho(dz) = dz$. We furthermore restrict to the subspace of the connectivity functions that vanish at the boundary, that we denote $\mathcal{H}_0$.

First of all, biases can be introduced by adding a constant latent variable $\kappa_0(t) = 1$, associated with a bias function $f_0(z)$. Let the set of approximating functions, called AF_0, be the set of functions $\hat{D} : \mathbb{R}^D \to \mathbb{R}$ of the form

$$\hat{D}(\boldsymbol{\kappa}) = \int_0^1 g(z) \phi\Big(\sum_{\nu=1}^{D} \kappa_\nu f_\nu(z) + f_0(z) \Big) dz \qquad \text{with } g, \{f_\mu\}_{0 \le \mu \le D} \subset \mathcal{H}_0 \tag{85}$$

AF_0 is a subset of the linear readouts that can be generated by field models over bounded one-dimensional space.

Consider an arbitrary dynamical system with asymptotic exponential decay:

$$\frac{d}{dt}\kappa_\alpha(t) = -\frac{1}{\tau}\kappa_\alpha(t) + D_\alpha(\vec{\kappa}(t)) \quad (\alpha = 1, \dots, D)$$

where $D_\alpha \in \mathcal{C}(\mathbb{R}^D, \mathbb{R})$ vanish for $\|\vec{\kappa}\| \to \infty$. The approximation problem consists in finding functions g and $\{f_\mu\}$ so that each given D_μ is approximated by some $\hat{D} \in \mathrm{AF}_0$. As for FFNNs, it can be decomposed into two parts [77]. First, show that approximating functions can be linearly combined: for any $\hat{D}^{(1)}, \hat{D}^{(2)} \in \mathrm{AF}_0$ and $c_1, c_2 \in \mathbb{R}$, one can find $\hat{D}' \in \mathrm{AF}_0$ such that $\hat{D}' = c_1\hat{D}^{(1)} + c_2\hat{D}^{(2)}$. Second, show that elementary functions (like half-plane indicators) can be well approximated. Arbitrary functions can then be approximated using linear combinations of elementary functions.

Linear combination.

For any linear combination in AF_0, we can write:

$$\begin{aligned}
(c_1\hat{D}^{(1)} + c_2\hat{D}^{(2)})(\vec{\kappa}) &= c_1 \int_0^1 g^{(1)}(z)\phi\left(\vec{\kappa}\cdot\vec{f}^{(1)}(z) + f_0^{(1)}(z)\right) dz \\
&\quad + c_2 \int_0^1 g^{(2)}(z)\phi\left(\vec{\kappa}\cdot\vec{f}^{(2)}(z) + f_0^{(2)}(z)\right) dz \\
&= c_1 \int_0^{1/2} 2g^{(1)}(2z)\phi\left(\vec{\kappa}\cdot\vec{f}^{(1)}(2z) + f_0^{(1)}(2z)\right) dz \\
&\quad + c_2 \int_{1/2}^1 2g^{(2)}(2z-1)\phi\left(\vec{\kappa}\cdot\vec{f}^{(2)}(2z-1) + f_0^{(2)}(2z-1)\right) dz \\
&= \int_0^1 g'(z)\phi\left(\vec{\kappa}\cdot\vec{f}'(z) + f_0'(z)\right) dz \quad =: \hat{D}'(\vec{\kappa})
\end{aligned}$$

where the parameters of $\hat{D}'$ are

$$g'(z) = \begin{cases} 2c_1 g^{(1)}(2z) & \text{if } z \in [0, 1/2[\\ 2c_2 g^{(2)}(2z-1) & \text{if } z \in [1/2, 1] \end{cases} ; \; e'_\alpha(z) = \begin{cases} e_\alpha^{(1)}(2z) & \text{if } z \in [0, 1/2[\\ e_\alpha^{(2)}(2z-1) & \text{if } z \in [1/2, 1] \end{cases}$$

for $\alpha = 0, \dots, D$. Because of the vanishing boundary condition on $[0, 1]$, if $g^{(i)}, \{f_\mu^{(i)}\}$ are in $\mathcal{H}_0$, so are $g', \{f'_\mu\}$, and so $\hat{D}' \in \mathrm{AF}_0$. Thus, AF_0 is a vector subspace of $\mathcal{C}(\mathbb{R}^D, \mathbb{R})$.

Elementary function.

Following the line of [77], we can build approximating functions for the indicator of half-planes in $\mathbb{R}^D$, i.e. $\mathbb{1}\{\boldsymbol{\kappa} \cdot \boldsymbol{d} > b\}$ for arbitrary $\boldsymbol{d} \in \mathbb{R}^D, b \in \mathbb{R}$.

Let $\mathcal{H}_0 \ni T : z \mapsto 2(1 - |2z - 1|)$ be the normalised triangle function over $[0, 1]$. Given $\boldsymbol{d} \in \mathbb{R}^D$ and $b \in \mathbb{R}$, for $A > 0$, let $\boldsymbol{e}(z) = AT(z)\boldsymbol{d}$, $e_0(z) = -AT(z)b$ and $g(z) = T(z)$. Then for $A \to \infty$,

$$\hat{D}_A(\boldsymbol{\kappa}) := \int_0^1 T(z)\phi\left(AT(z)\big(\boldsymbol{\kappa} \cdot \boldsymbol{d} - b\big)\right) dz \to \begin{cases} \phi(+\infty) & (\boldsymbol{\kappa} \cdot \boldsymbol{d} > b) \\ \phi(-\infty) & (\boldsymbol{\kappa} \cdot \boldsymbol{d} < b) \\ \phi(0) & (\boldsymbol{\kappa} \cdot \boldsymbol{d} = b) \end{cases}$$

By assuming that ϕ is sigmoidal (i.e. $\phi(+\infty) = 1, \phi(-\infty) = 0$), we see that $\hat{D}_A$ approximates the indicator of the half-plane in direction $\boldsymbol{d}$, given A large enough. The constant function $\hat{D}(\kappa) = 1$ can also be trivially constructed.

Since linear combinations of indicators of half-planes are dense in $C(\mathbb{R}^D, \mathbb{R})$ with compact support [77], arbitrary functions D_α can be approximated by functions of the form of eq.(85).