Information content in continuous attractor neural networks is preserved in the presence of moderate disordered background connectivity

Tobias Kühn, Rémi Monasson

Information content in continuous attractor neural networks is preserved in the presence of moderate disordered background connectivity

Tobias Kühn[1,2*] and Rémi Monasson[1]
[1] *Laboratoire de Physique de l'Ecole Normale Supérieure, CNRS UMR8023 & PSL Research, Sorbonne Université, Université Paris Cité, F-75005 Paris, France*
[2] *Institut de la Vision, Sorbonne Université, CNRS, INSERM, F-75012 Paris, France*
(Dated: January 3, 2024)

Continuous attractor neural networks (CANN) form an appealing conceptual model for the storage of information in the brain. However a drawback of CANN is that they require finely tuned interactions. We here study the effect of quenched noise in the interactions on the coding of positional information within CANN. Using the replica method we compute the Fisher information for a network with position-dependent input and recurrent connections composed of a short-range (in space) and a disordered component. We find that the loss in positional information is small for not too large disorder strength, indicating that CANN have a regime in which the advantageous effects of local connectivity on information storage outweigh the detrimental ones. Furthermore, a substantial part of this information can be extracted with a simple linear readout.

I. INTRODUCTION

The ring attractor neural network was proposed by Amari in the 70's as a practical way to memorize a collective variable within a noisy neural population [1]. This work opened the way to various theoretical applications of the concept of continuous attractor neural networks (CANN), *e.g.* in the contexts of the orientational tuning [2] or hippocampal place cells [3], as well as to extensions, in particular to the case of multiple attractor embeddings [4–6]. While indirect evidence for the existence of CANN could be found in recordings of activity in the hippocampus [7], in the enthorinal [8] and the prefrontal cortex [9], a direct and beautiful observation of ring attractor coding for head direction was obtained only recently in the ellipsoid body of the fly [10].

From a theoretical point of view, CANN models rely on recurrent excitatory interactions between neurons active for similar values of the encoded variable, *e.g.* the position of the animal in physical space, together with a long-range inhibition preventing all cells to be active together. This combination of local positive interactions and global inhibition creates a localized bump of activity, whose center of mass reliably represents the collective variable. In this regard, a crucial condition is that the bump can be easily moved (under weak external, "sensory", inputs) to span the continuous set of values of the variable. This condition imposes that the short-range connections are finely tuned, so that the model be effectively translation invariant.

When the finite-tuning condition breaks down, *e.g.* due to random modulations of the interactions, the bump can get stuck in the absence of neural noise [3]. In practice, quenched noise in the interactions can come from imperfect learning of one environment, or from interferences resulting from other information encoded (maps, objects distorting the map locally, ...). Quantifying the loss in the accuracy of information storage resulting from heterogeneities in the interactions is an important issue.

We address this question here in the framework of decoding of information, based on analytical and numerical calculations. We propose an analytically tractable model of binary (active/silent) neurons receiving position-dependent inputs, and connected to each other through spatially coherent and short-range interactions, on top of a disordered and incoherent background. Using the replica method we compute the Fisher information in the high-dimensional neural activity about the encoded position as a function of the intensity of disordered interactions. This quantity was identified as a measure that is both relatively easy to compute for many systems and objectively quantifies the information contained in the neural activity about the stimulus (an orientation or a point in space) [11]. It is more appropriate for this quantification than, for example, the readout of the center of mass of a bump of activity [12]. Yet, the Fisher information is not an information measure in the sense of Shannon. From this point of view, the mutual information between the stimulus and the neural activity is the quantity we are eventually interested in [13]. However, the latter is a global quantity, integrated over all possible stimuli and its computation is generally more difficult than that of the Fisher information, which puts restrictions on the system it can be calculated for [14]. In the thermodynamic limit, the mutual information can be obtained from the Fisher information under the condition that

* tobias.kuhn@inserm.fr

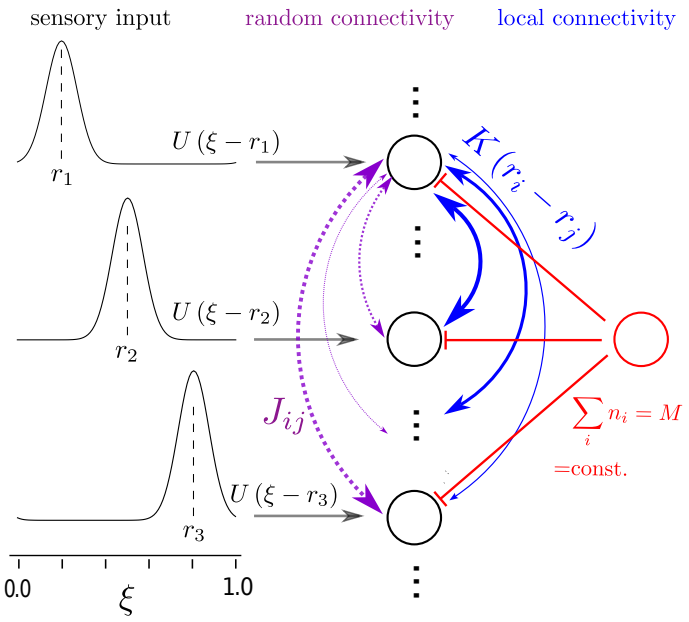

FIG. 1: Scheme of the network model employed in this work. The bell-shaped curves represent the space sensitive single-neuron (feedforward) input and the red circle a symbolic inhibitory neuron that ensures that the summed activity of all neurons is constant.

the correlations are not too strong [15]. We have explicitly checked that this prerequisite is fulfilled by our model, so that the computation of the Fisher information is actually sufficient. For a current review discussing both measures of information and their use in neuroscience, see e.g. [16].

The paper is organized as follows. In sec. II, our model is introduced. We establish the Fisher information as a means to quantify the information contained in the neural activity about the stimulus, together with its relation to other information-theoretic measures, compute it in the thermodynamic (mean-field) limit and derive its analytical properties in the limiting case of weak connection strengths. In sec. III, we validate our mean-field results by means of Monte-Carlo calulations, study the dependence of the Fisher information on changes in the recurrent and feed-forward connectivity and how precise a linear readout can be compared to the bound predicted by the Fisher information. In sec. IV, we put our results into context and give an outlook to possible future directions.

II. MODEL AND METHODS

A. Distribution of neural activities

We model neurons as binary units, taking the value 1 when active and 0 when inactive. Each neuron is receiving a 'sensory' input, whose value depends on the mismatch between the position r_i in physical space it is maximally responding to and the position of the 'animal', see fig. (1) for a scheme of our setup. The probability distribution of activities is governed by a Boltzmann law $P(\mathbf{n}|\xi) \sim e^{-E[\boldsymbol{n}]}$ and the energy

$$E[\boldsymbol{n}] = -\frac{1}{2}\sum_{i\neq j}(J_{ij} + K_{ij})\, n_i n_j - \sum_i n_i \,\mathrm{U}(\xi - r_i)\,, \tag{1}$$

where $K_{ij} = K(r_i - r_j)$ is the local part of the interaction that we assume to be decaying in space with a typical length scale w_{rec} and strength K_{rec}. J is the disordered part of the connectivity with the statistics

$$\langle J_{ij}\rangle = 0,\ \langle J_{ij}J_{i'j'}\rangle = \frac{g^2}{2N}\delta_{ii'}\delta_{jj'} \tag{2}$$

and U mimics the space-dependent input, which we model by

$$\mathrm{U}(\Delta x) = \mathrm{U}_{\mathrm{inp}} \exp\left(-\frac{\Delta x^2}{w_{\mathrm{inp}}^2}\right). \tag{3}$$

We are assuming periodic boundary conditions. As a simple way to model global inhibition, we impose the constraint that the summed activity is fixed to $\sum_i n_i = f \cdot N = M$, where $f \in [0, 1]$. Our model is closely related to that of [17], where the case of a proper continuous attractor neural network (CANN) is studied, so the connectivity matrix is composed of a sum of local connectivities in different environments. However, in this study we are interested in the effect of the disorder on the information content in the neural activity for a single environment. Also, the presence of other maps is not the only source of disorder as there is always some variability in the connectivity. Therefore, to simplify the setup, we content ourselves with approximating the disordered contribution to the connections as Gaussian, which also corresponds to the high-temperature behaviour of the disorder in [17].

B. Fisher information and mutual information

We now want to quantify the amount of information contained in the neural activity. One possibility to do so is to compute the Fisher information for a given stimulus ξ,

$$\mathcal{I}_{\boldsymbol{n}}(\xi) \coloneqq \left\langle -\frac{\partial^2}{\partial \xi^2} \ln P(\boldsymbol{n}|\xi) \right\rangle_{\boldsymbol{n}}, \tag{4}$$

a standard measure for the quantification of information in neural populations [12, 16]. According to the Cramér-Rao bound, its inverse gives a lower bound on the variance of any unbiased estimator of ξ [13]. We will discuss this relation in greater depth in section III C. Furthermore, in the thermodynamic limit, that we are interested in, it also determines the mutual information, a connection first established in [15] and later refined in [18]. The mutual information is given by the decrease of the entropy of the neural activity due to the knowledge of the stimulus, concretely

$$I_{\mathrm{MI}} \coloneqq -\sum_{\boldsymbol{n}} P(\boldsymbol{n}) \ln P(\boldsymbol{n}) + \int d\xi\, p(\xi) \sum_{\boldsymbol{n}} P(\boldsymbol{n}|\xi)\, \ln P(\boldsymbol{n}|\xi)$$
$$\text{where } P(\boldsymbol{n}) \coloneqq \int d\xi\, p(\xi)\, P(\boldsymbol{n}|\xi).$$

In [15] (their eq. (13)), the relation

$$I_{\mathrm{MI}} = -\int d\xi\, p(\xi) \ln p(\xi) + \int d\xi\, p(\xi) \ln\left(\frac{\mathcal{I}_{\boldsymbol{n}}(\xi)}{2\pi e}\right) + \mathcal{O}\left(\frac{1}{N}\right) \tag{5}$$

was derived for an ensemble of neurons with fixed covariances, without disorder. However, in this study we are limiting ourselves to the saddle-point approximation of the Fisher information, which is valid up to corrections of order $1/N$ as well. So the presence of disorder does not change much and we obtain that

$$\langle I_{\mathrm{MI}} \rangle_J = -\int d\xi\, p(\xi) \ln p(\xi) + \int d\xi\, p(\xi) \ln\left(\frac{\langle \mathcal{I}_{\boldsymbol{n}}(\xi) \rangle_J}{2\pi e}\right) + \mathcal{O}\left(\frac{1}{N}\right), \tag{6}$$

where $\langle \dots \rangle_J$ is indicating the average over the disordered connectivity J. In appendix D, we rederive this relation, for unconnected neurons, but more directly than in [15].
Hereafter, we will focus on the Fisher information, which is easier to obtain than the mutual information that we get for free due to eq. (6). Determining the Fisher information for our model, we obtain from eq. (4) after some lines of computation, detailed in appendix B,

$$\mathcal{I}_{\boldsymbol{n}}(\xi) = \sum_{i,j} \mathrm{U}'(\xi - r_i) \left\langle \left[\langle n_i n_j \rangle_{\boldsymbol{n}} - \langle n_i \rangle_{\boldsymbol{n}} \langle n_j \rangle_{\boldsymbol{n}} \right] \right\rangle_J \mathrm{U}'(\xi - r_j) \tag{7}$$

$$= \left[\mathbf{U}'(\xi - \boldsymbol{r}) \right]^{\mathrm{T}} C \mathbf{U}'(\xi - \boldsymbol{r}), \tag{8}$$

where C denotes the disorder-averaged covariance matrix of $\boldsymbol{n}$. Conditioned on one realization of the disorder furthermore, we have introduced the thermal average

$$\langle f(\boldsymbol{n}) \rangle_{\boldsymbol{n}} \coloneqq \frac{1}{\mathcal{Z}_J(\xi)} \sum_{\boldsymbol{n}} f(\boldsymbol{n})\, e^{-E[\boldsymbol{n}]}, \tag{9}$$

for some function f, together with the partition function

$$\mathcal{Z}_J(\xi) \coloneqq \sum_{\boldsymbol{n}} e^{-E[\boldsymbol{n}]}. \tag{10}$$

Eq. (7) can be brought into a more familiar form by noting that

$$\frac{\partial}{\partial \xi}\mathrm{T}(\xi - r_i) \coloneqq \frac{\partial}{\partial \xi}\langle n_i\rangle_{\boldsymbol{n}} = \left\langle n_i \left[\sum_j \mathrm{U}'(\xi - r_j)(n_j - \langle n_j\rangle)\right]\right\rangle_{\boldsymbol{n}} = \left(C\mathbf{U}'\right)_i \tag{11}$$

$$\Leftrightarrow \mathrm{U}'(\xi - r_i) = \left(C^{-1}\mathbf{T}'\right)_i, \tag{12}$$

where we have introduced the tuning curve T of neuron i indicating its average activity given the input ξ. With this, the Fisher information can be written as [14]

$$\mathcal{I}_{\boldsymbol{n}}(\xi) = \left[\mathbf{T}'(\xi)\right]^{\mathrm{T}} C^{-1}\mathbf{T}'(\xi). \tag{13}$$

This form is more handy when dealing with experimental data because the tuning curve is (in principle) a directly measurable quantity, whereas U′ is not. For our purposes, however, the form of eq. (7) is more practical because there, the only quantity depending on the disorder is the covariance matrix. We therefore only have to compute the disorder average of the covariance matrix, which we will tackle in the following.

C. Disorder-averaged statistics

As usual for disordered systems [19], we determine the statistics from the logarithm of the partition function, the cumulant-generating functional (or Gibbs free energy)

$$\langle W(\boldsymbol{h})\rangle_J = \int dJ\, P(J) \ln\left[\sum_{\boldsymbol{n},\,\sum_i n_i = M} e^{\frac{1}{2}\sum_{i\neq j}(J_{ij}+K_{ij})n_i n_j + \sum_i n_i[\mathrm{U}(\xi - r_i)+h_i]}\right]. \tag{14}$$

The computation of W proceeds along the classical lines [20], with the difference that we have a local connectivity. We therefore do not only introduce the Gaussian helping field q to decouple the four-point terms emerging from the disorder average, but also a space-dependent (also Gaussian) order parameter ϕ_x to decouple the local term $\boldsymbol{n}^{\mathrm{T}}K\boldsymbol{n}$. As apparent from the saddle-point equations (18) and (19), q quantifies the population-averaged variance of the activity and ϕ_x the population-averaged input to the neuron with place field at position x. Furthermore, due to the restriction on the summed activity, we introduce the Lagrange multiplier λ. As derived in appendix B, we obtain the disorder-averaged cumulant-generating function in the thermodynamic limit $N \to \infty$

$$\langle W(\boldsymbol{h})\rangle_J = \underset{q,\bar{q},\boldsymbol{\psi},\boldsymbol{\phi},\lambda}{\mathrm{extr}} \left\{\frac{1}{2}Ng^2q^2 - \frac{1}{2}Ng^2\bar{q}^2 - \frac{1}{2}\boldsymbol{\phi}^T K^{-1}\boldsymbol{\phi} - N\left(\lambda - g^2(\bar{q} - q)\right) f. \right. \tag{15}$$

$$\left. + \prod_y \frac{1}{\sqrt{2\pi}} \int dt_y\, e^{-\frac{t_y^2}{2}} \sum_x \ln\left[1 + e^{\phi_x + t_x g\sqrt{2q} + \mathrm{U}(\xi - r_x) + \lambda + h_x}\right]\right\} \tag{16}$$

$$=: G(\boldsymbol{h}, \boldsymbol{\phi}, q), \tag{17}$$

where the "extr." implies a supremum over q, $\bar{q}$, $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$ and an infimum over λ. We comment on the latter point in appendix A 1. As detailed in appendix A 2, we obtain the saddle-point equations

$$q = \int dx \int Dt_x \frac{1}{\left[1 + e^{-\left(\phi_x + t_x g\sqrt{2q} + \mathrm{U}(\xi - x) + \lambda\right)}\right]^2}, \tag{18}$$

$$\phi_x = \int dy\, K(x - y) \int Dt_y \frac{1}{1 + e^{-\left(\phi_y + t_y g\sqrt{2q} + \mathrm{U}(\xi - y) + \lambda\right)}}. \tag{19}$$

$$\text{and } f = \int dx \int \mathcal{D}t_x \frac{1}{1 + e^{-\left(\phi_x + t_x g\sqrt{2q} + \mathrm{U}(\xi - x) + \lambda\right)}}, \tag{20}$$

with the Gaussian measure

$$\int \mathcal{D}t = \frac{1}{\sqrt{2\pi}} \int dt e^{-\frac{t^2}{2}}. \tag{21}$$

The entire statistics of our system can now be determined by taking derivatives of G with respect to h, which is set to 0 afterwards. We have to calculate the total derivative, also taking into account the h-dependence of q, ϕ and λ, which in turn, by the implicit-function theorem, we obtain by taking the total derivative with respect to h of the their saddle-point-equations. This yields

$$\frac{1}{N}\frac{d^2}{d\boldsymbol{h}^2}\left\langle W_f\left(\boldsymbol{h}\right)\right\rangle_J = \frac{\partial^2 G}{\partial \boldsymbol{h}^2} - \begin{pmatrix} \frac{\partial^2 G}{\partial \boldsymbol{h}\partial \boldsymbol{\phi}} \\ \frac{\partial^2 G}{\partial \boldsymbol{h}\partial q} \\ \frac{\partial^2 G}{\partial \boldsymbol{h}\partial \lambda} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \frac{\partial^2 G}{\partial \boldsymbol{\phi}^2} & \frac{\partial^2 G}{\partial \boldsymbol{\phi}\partial q} & \frac{\partial^2 G}{\partial \boldsymbol{\phi}\partial \lambda} \\ \frac{\partial^2 G}{\partial q\partial \boldsymbol{\phi}} & \frac{\partial^2 G}{\partial q^2} & \frac{\partial^2 G}{\partial q\partial \lambda} \\ \frac{\partial^2 G}{\partial \lambda\partial \boldsymbol{\phi}} & \frac{\partial^2 G}{\partial \lambda\partial q} & \frac{\partial^2 G}{\partial \lambda^2} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial^2 G}{\partial \boldsymbol{\phi}\partial \boldsymbol{h}} \\ \frac{\partial^2 G}{\partial q\partial \boldsymbol{h}} \\ \frac{\partial^2 G}{\partial \lambda\partial \boldsymbol{h}} \end{pmatrix}. \tag{22}$$

Evaluating this expression, we obtain

$$C = V + V K_{\mathrm{eff}} V + C^{\mathrm{indirect}} \tag{23}$$

$$= V\left(\mathbb{1} - KV\right)^{-1} + C^{\mathrm{indirect}}, \tag{24}$$

where V is the diagonal matrix with the disorder-averaged single-neuron variances

$$V_{xy} = \delta_{xy} v_x, \tag{25}$$

$$\text{where } v_x \coloneqq \int \mathcal{D}t \frac{\partial m_x}{\partial \phi_x} = \int \mathcal{D}t\, m_x\left(1 - m_x\right), \tag{26}$$

m_x is the magnetization conditioned on the Gaussian helping variable t,

$$m_x \coloneqq \frac{1}{1 + e^{-\left[\phi_x + t\, g\sqrt{2q} + \mathrm{U}(\xi - x) + \lambda\right]}}, \tag{27}$$

the effective local connectivity K_{eff} is given by $\left[\left(K^{\mathrm{eff}}\right)^{-1}\right]_{xy} \coloneqq -\frac{\partial^2 G}{\partial \phi_x \partial \phi_y}$ and fulfills the Dyson equation

$$K^{\mathrm{eff}}_{xy} = K_{xy} + \int dz\, K_{xz} v_z K^{\mathrm{eff}}_{zy}, \tag{28}$$

and C^{indirect} emerges from the remaining part of the Hessian in eq. (22). It results in an subleading contribution to the Fisher information (see below), so we give its precise form only in appendix B.

D. Disorder-averaged Fisher information

The Fisher information per neuron averaged over the disorder now reads

$$\mathcal{I}_{\boldsymbol{n}}\left(\xi\right) = \sum_x \sum_y \mathrm{U}'\left(\xi - x\right)\left[v_x \delta_{x,y} + v_x K^{\mathrm{eff}}_{xy} v_y + C^{\mathrm{indirect}}_{xy}\right]\mathrm{U}'\left(\xi - y\right) \tag{29}$$

In fig. (6) in the appendix, we show these three contributions separately for the parameters used for fig. (3). The first term stems from the single-neuron variances and is therefore also present without network (if present the variances are effected by the network, though). The third is always negligible, which intuitively makes sense because it emerges from the indirect $\boldsymbol{h}$-dependence of the free energy via g and λ, which are both spatially unstructured. The second term emerges from the (positive) local interactions and also contributes positively. In order to gain a better intuition for where this term comes from, it is useful to re-derive the expression for the Fisher information using eq. (13), limiting ourselves to the case without disorder. By some lines of rearrangements, shown in appendix B 2, we derive that the vector of the derivatives of the tuning curves can be expressed as

$$\mathbf{T}' = V\left(1 + K_{\mathrm{eff}} V\right)\mathbf{U}'. \tag{30}$$

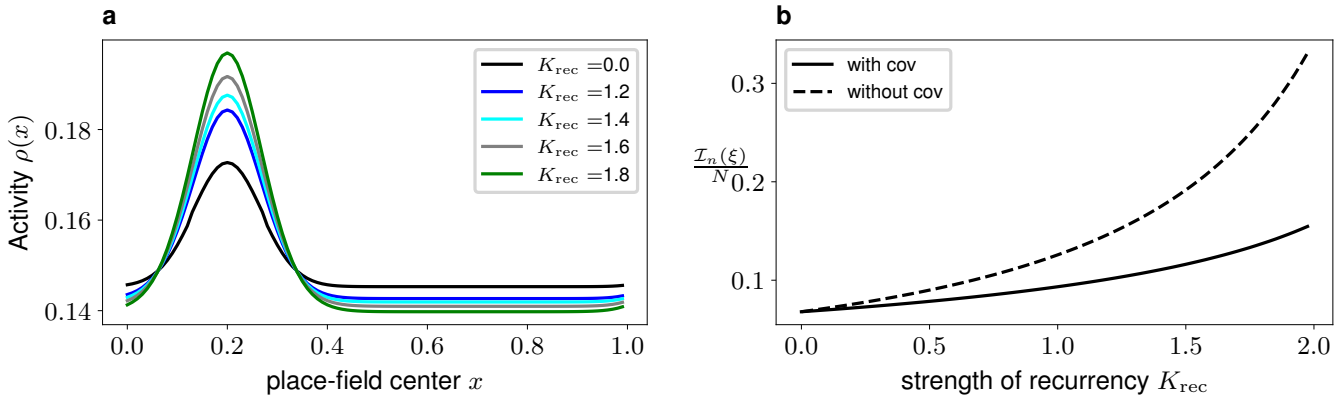

FIG. 2: Panel a: tuning curves for a network without disorder in the connectivity for different strengths of local interactions. Panel b: corresponding change in the Fisher information; the dotted lines show eq. (31) with the term $VK_{\mathrm{eff}}V$ in the middle part, constituting C^{-1}, removed. Parameters: $g = 0$, $w_{\text{two point}} = 0.1$, $\mathrm{U}_{\mathrm{inp}} = 0.2$, $w_{\text{one point}} = 0.07$, $f = 0.15$.

Combining this expression with eq. (13) and $C = V + VK_{\mathrm{eff}}V$, we are getting back eq. (29) (without the contribution of the disorder) after canceling a factor $(V + VK_{\mathrm{eff}}V)$, as expected. However, it is also insightful to write down the expression before this cancellation,

$$\mathcal{I}_{\boldsymbol{n}} = \overbrace{\left(\mathbf{U}'\right)^{\mathrm{T}}}^{=(\mathbf{T}')^{\mathrm{T}}} \left(V + VK_{\mathrm{eff}}V\right) \overbrace{\left[V + VK_{\mathrm{eff}}V\right]^{-1}}^{=C^{-1}} \overbrace{\left(V + VK_{\mathrm{eff}}V\right)\mathbf{U}'}^{=\mathbf{T}'}, \tag{31}$$

because it gives an intuition about how the local connectivity shapes the Fisher information: first, it modifies the tuning curves, which is captured by the term $VK_{\mathrm{eff}}V\mathrm{U}'$, second it introduces cross-covariances, which is captured by the term $VK_{\mathrm{eff}}V$ contributing to the covariance. As apparent from fig. (2), panel a, the tuning curves are sharpened with increasing K_{rec}, which is reflected by the fact that the direct contribution of the cross-covariances to the Fisher information is positive, see fig. (6). The cross cavariances, in turn, are detrimental in our case: they reduce the Fisher information, as apparent from panel b in fig. (2).

We can study further eq. (29) analytically, which amounts to examining the saddle-point equations (18) to (20). In particular, we can do this in the limiting case $g \to 0$. In this limit, the Gaussian integrals get trivial. As detailed in appendix A 3, we can use for the study of the derivatives of q, ϕ and λ that, in eqs. (18) to (20), these quantities are (implicitly) given by

$$0 = \frac{\partial}{\partial\left[\{\phi_x\}_x, \lambda, q\right]} G_g\left(g, q, \phi, \lambda\right), \tag{32}$$

with G as given in eq. (17) and therefore, by the implicit-function theorem,

$$\frac{\partial}{\partial g}\begin{pmatrix}\{\phi_x\}_x\\ \lambda\\ q\end{pmatrix} = -\left(\frac{\partial^2}{\partial\left[\{\phi_x\}_x, \lambda, q\right]^2} G_g\left[q, \boldsymbol{\phi}, \lambda\right]\right)^{-1} \frac{\partial^2}{\partial g \partial\left[\{\phi_x\}_x, \lambda, q\right]} G_g\left[q, \boldsymbol{\phi}, \lambda\right]. \tag{33}$$

For a reasonable choice of the parameters, also guaranteeing the stability of the saddle-point solution, $\frac{\partial^2}{\partial\left[\{\phi_x\}_x, \lambda, q\right]^2} G_g$ is invertible (which we as well check numerically by computing it explicitly in appendix B). The partial derivative of $\partial_{\{\phi_x\}_x, \lambda, q} G$ with respect to g vanishes in the limit $g \to 0$, as we show in appendix A 3. Therefore, the derivatives of q, ϕ and λ with respect to g go to 0 for vanishing g. This results carries over to higher-order cumulants and to the cumulant-generating functional itself. Because the Fisher information depends on g only via these quantities, its derivative vanishes in the limit of $g \to 0$:

$$\lim_{g \to 0} \frac{\partial}{\partial g} \mathcal{I}_{\boldsymbol{n}}\left(\xi\right) = 0. \tag{34}$$

The derivatives with respect to the strength of the local interaction, K_{rec}, however, in general do not vanish for vanishing connectivity. Therefore even small connection strengths will have a (beneficial) effect, as seen before.

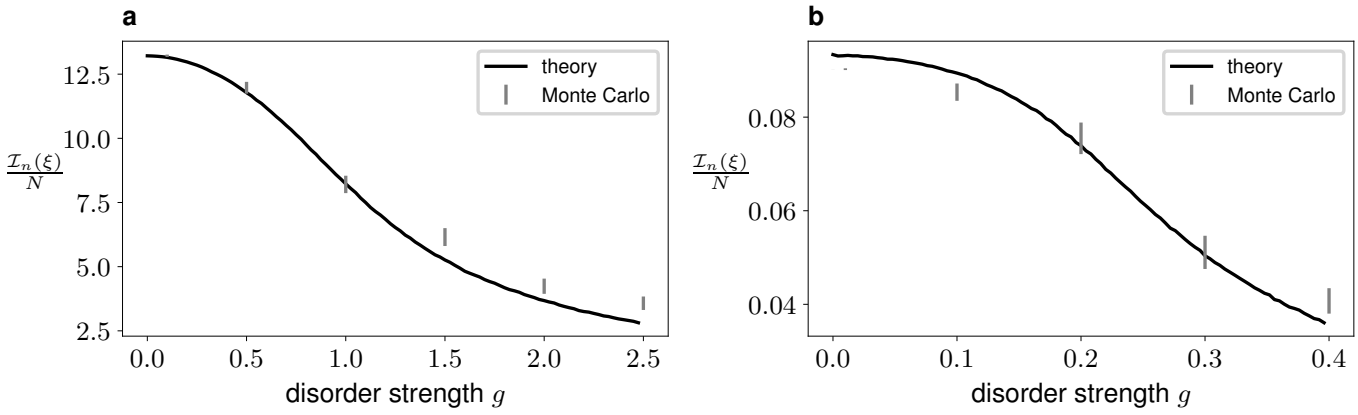

FIG. 3: The Fisher information per neuron in dependence of the disorder, comparison with Monte-Carlo simulations. In panel a, the network is strongly input driven, and in panel b only weakly. Parameters, panel a: $K_{\mathrm{rec}} = 5$, $w_{\mathrm{rec}} = 0.1$, $\mathrm{U}_{\mathrm{inp}} = 2.25$, $w_{\mathrm{inp}} = 0.07$, $f = 0.15$; panel b: $K_{\mathrm{rec}} = 20$, $\mathrm{U}_{\mathrm{inp}} = 0.2$, other parameters as in panel a.

III. NUMERICAL VALIDATION OF MEAN-FIELD RESULTS AND APPLICATIONS

A. Monte Carlo simulation

To validate our computation derived for the thermodynamic limit, we perform Monte-Carlo simulations for multiple sets of parameters, see fig. (3).We use a standard Metropolis algorithm, taking into account the condition of a fixed total activity by flipping always two spins, in opposite direction, as suggested in [17]. The code for these simulations, alongside with the implementation of the mean-field results is provided in [21]. The results confirm our mean-field result that the disorder diminishes the Fisher information, but quite slowly if the disorder is on a moderate level, as predicted by eq. (34) and visible in fig. (3).

B. Influence of the network on the Fisher information

In an attractor network, disorder may result from the presence of other maps stored in the same network. Therefore it scales in the same way as the spatially dependent part of the connectivity. It is thus interesting to examine the behavior of the Fisher information when both parts of the connectivity are scaled by the same factor r:

$$K \to r \cdot K, \quad J \to r \cdot J. \tag{35}$$

Because the derivative of the expression for the Fisher information with respect to the disorder strength g vanishes for $g = 0$, the effect of the local part of the connectivity dominates for small synaptic strength: the Fisher information initially increases. This can be understood from what we have derived before: increasing the local connectivity sharpens the tuning curves and therefore increases the signal. This effect is diminished (but not cancelled) by the introduction of covariances between the neurons (see also [12]). For larger scaling factors, however, this overall beneficial effect is wiped out by the disorder, whose detrimental effect eventually dominates, fig. (4), panel a.

Finally, we ask if we can keep the Fisher information constant by increasing the recurrent weights when the input gets weaker. We have plotted lines of constant Fisher information for varying strength of the input and the recurrent connections in fig. (4), panel b. We can indeed make up to a decrease in the input by strengthening local connections, even though of course only in a limited range.

C. Linear readout

To put our results into context and convey a more intuitive understanding, we briefly discuss what one can learn from the Fisher information about the accuracy of a linear readout.

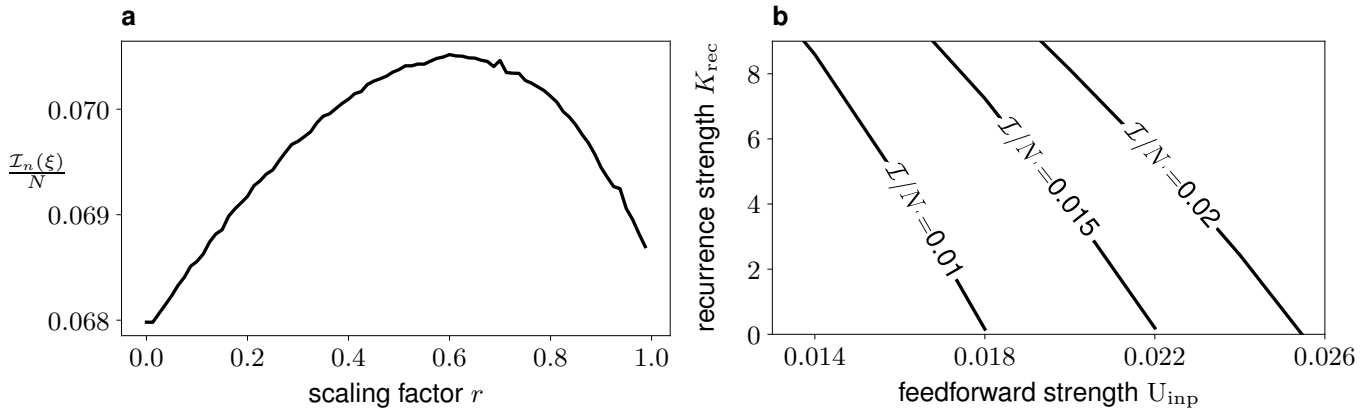

FIG. 4: Interplay of feedforward, local and disordered recurrent input shaping the Fisher information. For panel a, we scale the synapses according to eq. (35), keeping the other parameters fixed, for panel b, we keep the the Fisher information constant, varying $\mathrm{U_{inp}}$ and $K_{\rm rec}$ concertedly. Parameters, panel a: $K_{\rm rec}^{\rm max} = 8$, $w_{\rm rec} = 0.1$, $\mathrm{U_{inp}} = 0.2$, $w_{\rm inp} = 0.07$, $f = 0.15$, $g^{\rm max} = 0.16$ panel b: $g = 5$, other parameters as in panel a.

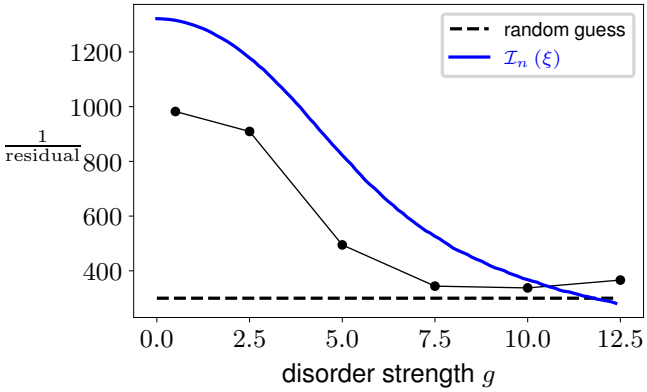

FIG. 5: Inverse residual of a linear fit of the neural activity to estimate ξ, averaged over $\xi \in [0.4, 0.6]$. Only for small g (approximately the first two data points at 0.1 and 0.5), this estimate is approximately unbiased and therefore only there, the Cramér-Rao bound guaranties that the Fisher information is an upper bound. Note that in this plot, different to all the others, we are plotting the total Fisher information for all 100 neurons, not the averaged, single-neuron analog. Parameters as in fig. 3, panel a.

We have fitted a readout vector $\boldsymbol{w}$ to the activity as measured after the thermalization process (so the random initial conditions should not play a role) and computed the squared residual error as

$$\langle \mathrm{res}\,(\xi)\rangle_\xi \coloneqq \min_{\boldsymbol{w}} \left\langle \left\langle \|\boldsymbol{w}\cdot\boldsymbol{n} - \xi\|_2^2 \right\rangle_{\boldsymbol{n}} \right\rangle_\xi = \langle \xi^2\rangle_\xi - 2\boldsymbol{w}_{\rm train}^{\rm T} C^{\rm test}_{\xi,\boldsymbol{n}} + \boldsymbol{w}_{\rm train}^{\rm T} C^{\rm test}_{\boldsymbol{n},\boldsymbol{n}} \boldsymbol{w}_{\rm train} \tag{36}$$

$$\text{where } \boldsymbol{w}_{\rm train} = \left[C^{\rm train}_{\boldsymbol{n},\boldsymbol{n}} + \lambda\cdot\mathbb{1}\right]^{-1} C^{\rm train}_{\xi,\boldsymbol{n}}, \tag{37}$$

where $\langle\ldots\rangle_{\boldsymbol{n}}$ and $\langle\ldots\rangle_\xi$ denote the thermal average over the configurations $\boldsymbol{n}$ and the average over the distribution of the stimulus ξ, respectively and λ denotes the strength of the L_2 regularization that we impose. We might expect to obtain an upper bound for the accuracy of the linear estimator by the Cramér-Rao bound. However, due to the periodic boundary conditions, the estimate from the linear readout $\xi_{\rm est.} = \boldsymbol{w}\cdot\boldsymbol{n}$ will be biased. This is particularly apparent at the borders 0 and 1, where the estimate will always be $\frac{1}{2}$, corresponding to random guessing. The farther away from the them the stimulus is situated, the less pronounced the effect becomes. We have therefore limited the fitted stimuli to $\xi \in [0.4, 0.6]$. However, even in this regime, the linear readout gets biased in the highly disordered

regime, so that there the Cramér-Rao bound only applies in its generalized form [13], their eq. (12.333):

$$\left\langle (\boldsymbol{w}\cdot\boldsymbol{n}-\xi)^2 \right\rangle_{\boldsymbol{n}} \geq \frac{(1+b'(\xi))^2}{\mathcal{I}_{\boldsymbol{n}}(\xi)} + b(\xi)^2\,, \tag{38}$$

where $b(\xi) \coloneqq \langle \boldsymbol{w}\cdot\boldsymbol{n}\rangle_\xi - \xi$ is the bias of the linear estimator. In case of random guessing, in particular, $b'(\xi) = -1$, so that the error is solely determined by the square of the bias. Therefore, the bound given by the Fisher information in fig. (5) is only meaningful for small disorder ($g \sim 0.1, 0.5$), whereas for greater disorder, it is invalidated by the bias, up to the point where the linear readout basically generates a random guess (for $g \sim 2$). In the low-disorder regime, however, we observe that the error of the linear readout is not far from the optimal case given by the Cramér-Rao bound.

IV. DISCUSSION AND OUTLOOK

In this work, we have studied an attractor-inspired neural network with a connectivity consisting of two parts: (1) excitatory couplings between neurons that are similarly tuned to stimuli, and (2) a quenched random background without local tuning. We have studied the influence of both of these contributions on the information about the stimulus contained in the neural activity through the analytical computation of the Fisher information. As expected the local part of the connectivity enhances the information content, whereas the disordered part degrades it. However, the latter effect is mild. By fitting a linear readout to estimate the driving stimulus we show that the Fisher information is not only a formal estimate of the information contained in the neural activity, but also gives a useful bound on how much of it can be extracted with simple decoders.

It has been recognized for a long time that the presence of disorder in the interactions could impact the information stored in attractor networks in the form of patterns. In the case of CANNs this translates into a breaking of the translational symmetry of bump-like solutions, see for instance [3] and box 3 in [22]. However, despite the loss of translation invariance, noise in the neural activity (controlled here by the inverse of the coupling strength, playing the role of temperature in statistical mechanics) may be sufficient to move the bump [23]. We here show that the disorder does not wipe out all information in the attractor network. In particular, the Fisher information is robust to the introduction of disorder, staying constant to first order in g. Consequently, globally enhancing the connectivity strength (the local and the disordered part by the same factor, as in eq. (35)) initially always has a beneficial effect, which is overtaken by the effect of the disorder only for larger connectivity. In real biological networks, the connectivity is not fixed, but builds up during development, partially through learning [24]. One might therefore speculate if this process is optimized for the synaptic strengths to eventually match this sweet spot.

Similarly, we can ask for the information-theoretical implication of the development of tuning curves as observed in the visual cortex. In young animals, recurrent connections between similarly tuned neurons get enhanced, other weakened [25], and the orientation selectivity is sharpened [26]. According to our analysis, these changes in recurrent connectivity could compensate a decrease in strength of sensory inputs [27], see fig. (4)(b).

While our model allows for this kind of qualitative considerations, it is minimal in the sense that it contains the ingredients needed to study the effects we are interested in in their simplest form. This of course limits biological plausibility and calls for enhancements. From a technical point of view, we did of course not consider all possible scenarios. We give below an outlook on possible further directions, starting with the technical aspects.

The results we have obtained here hold for binary-valued neurons. From a neuroscience perspective, this assumption can be interpreted as follows. Consider the neural population activity in a time-bin of duration Δt. If Δt is small enough (in practice, not larger than the inverse of the typical firing rate), it is likely that each neuron i has fired at most once in a time bin. We can therefore represent its activity through a binary variable n_i equal to 0 in the absence of spike, or to 1 if a spike has been emitted. This point of view was adopted in data-driven models of the hippocampal activity [7]. If the duration Δt is large, the binary hypothesis breaks down and continuous models, taking into account the real-value nature of firing rates, should be considered. Continuous-attractor models with real-valued neurons show similar - albeit for some parameters a bit richer [28] - behavior as models with binary neurons [3], so we do not expect our conclusions to be qualitatively modified in that setting. In addition, binary patterns represent an important limiting case: they may maximize the retrievable information, in the case of random memories, as compared to more complex ones [29].

Then, we have limited ourselves to a parameter regime far away from phase transitions. In particular we have not considered the low-temperature/high-connectivity regime. The analysis of this case (without the computation of the Fisher information) was carried out by one of us and collaborators in a series of papers [17, 23, 30], in the specific case of background noise due to a extensive number αN of alternative attractors embedded in the network. Although the quenched noise distribution in this case was not Gaussian we do not expect the phase diagram to significantly

change. Based on these previous works we can thus build educated guesses on what to expect in terms of information theory. Note that, in terms of scaling, the square of the disorder strength used here, g^2, roughly corresponds to the load α in [17] (compare, e.g. their saddle-point equations, eq. (28) to ours, eqs. (18) to (20)).

Analog to [17], we expect a glassy phase for large disorder and weak local connectivity, a ferromagnetic ("bump") phase for weak disorder and strong local connectivity and a paramagnetic phase in case both contributions are small (large-temperature regime). In the present study, we have basically stayed in the latter regime (however, the activity was still bump-like due to the feed-forward input). This limitation has also allowed us to stick to the replica-symmetric solution of the saddle-point equations - an assumption that might not be satisfied at very low temperature and strong enough disorder; however, the comparison with the results of Monte-Carlo simulations in fig. (3) led to reasonable results. This is expected; for spin-glass models, the effect of replica-symmetry breaking is typically rather limited, in particular close to the Almeida-Thouless line. In order to study the glassy regime, corresponding to a network load α beyond the critical value, we expect that replica symmetry breaking has to be considered. However, as for the Fisher information, we expect it to be approximately zero in the glassy state because the activity would not show any dependence on space anymore (besides a residual one due to the input), and the disorder-averaged single-neuron variances vanish for large disorder (and all other cumulants as well). As for the paramagnetic to ferromagnetic transition, we expect a qualitative change in the shape of the neural activities, with considerably sharper bumps on the ferromagnetic side of the transition. Due to the mechanisms discussed around eq. (13), we expect a corresponding steep increase in the Fisher information.

On a more biological side, we have made, for convenience, several unrealistic assumptions that could be waived. The receptive fields in our setup, for example, have all the same shape and size and are evenly spaced - in reality there is of course variability in shape and they are scattered in the environment. In particular, place fields may cluster near new objects [31], suggesting the importance of taking into account inhomogeneous densities. These features could rather easily be included in our framework, at least if the probability distributions for the single-neuron properties are independent. The additions suggested above only require another average over them (see also appendix D). Then, though we have here studied a one-dimensional stimulus, the features determining the Fisher information, such as the sharpness of the tuning curve, can also be defined in higher dimensions, and we expect qualitatively the same results in that case. Last of all, it would be interesting to better understand how much of the information we have estimated can be extracted from the neural population in practice, beyond the linear readout mechanism considered here.

ACKNOWLEDGMENTS

TK thanks Ulisse Ferrari and Gabriel Mahuas for many very insightful discussions. This work was partly funded by the Human Frontier Science Program RGP0057/2016 grant and TK by a short-term postdoc fellowship of the German Academic Exchange Service (DAAD).

Appendix A: Computing the cumulant-generating function

As indicated in the main text, computing the Fisher information basically amounts to calculating the covariance matrix of the activity, which we obtain from the disorder-average of the cumulant-generating functional defined in eq. (14), or in other words, of the logarithm of the partition function. In the following, we will explain step by step how to take into account its features, starting with the fixed total activity, then including the disordered part of the connectivity and finally deriving and analysing the saddle-point equations for this case in order to compute the cumulant-generating functional in the thermodynamic limit.

1. Effects of the fixed total activity and the space-dependent part of the coupling

As mentioned in the main text, we fix the summed activity of all neurons to a certain number M, mimicking the effect of a global inhibition. This determines the partition function, which reads

$$\mathcal{Z}_M\left[\boldsymbol{h}\right] = \sum_{\boldsymbol{n},\sum_i n_i = M} \exp\left(\sum_{i<j,} K_{ij} n_i n_j + \sum_{i=1}^{N} h_i n_i\right).$$

Explicitly performing the spin sums under this limitation is difficult, so we introduce the Fourier series with $\mathcal{Z}_M[\boldsymbol{j}]$ as coefficients:

$$\begin{aligned}
U_k[\boldsymbol{h}] &:= \sum_{M=0}^{N} e^{i2\pi kM} \mathcal{Z}_M[\boldsymbol{h}] \\
&= \sum_{M=0}^{N} e^{i2\pi kM} \sum_{\boldsymbol{n}} \delta_{M,\sum_i n_i} \exp\left(\sum_{i<j,} K_{ij} n_i n_j + \sum_{i=1}^{N} h_i n_i\right) \\
&= \sum_{\boldsymbol{n}} \exp\left(\sum_{i<j,} K_{ij} n_i n_j + \sum_{i=1}^{N} h_i n_i\right) \sum_{M=0}^{N} e^{i2\pi kM} \delta_{M,\sum_i n_i} \\
&= \sum_{\boldsymbol{n}} \exp\left(\sum_{i<j,} K_{ij} n_i n_j + \sum_{i=1}^{N} (h_i + i2\pi k) n_i\right).
\end{aligned}$$

Applying the transform to obtain the Fourier coefficients from a periodic function, we get

$$\begin{aligned}
\mathcal{Z}_M[\boldsymbol{h}] &= \int_0^1 dk\, e^{-i2\pi kM} U_k[\boldsymbol{j}] \\
&= \int_0^1 dk \sum_{\boldsymbol{s}} \exp\left(\sum_{i<j,} K_{ij} n_i n_j - i2\pi k \left(M - \sum_{i=1}^{N} n_i\right) + \sum_{i=1}^{N} h_i n_i\right).
\end{aligned}$$

For $N \gg 1$, we can evaluate the k-integral in saddle-point approximation, so that in this limit, we replace the partition function by its "grandcanonical" counterpart

$$\mathcal{Z}_{f,\mathrm{gc}}[\boldsymbol{h}] := \inf_{\lambda} \left[\sum_{\boldsymbol{n}} \exp\left(\sum_{i<j,} K_{ij} n_i n_j + \sum_{i=1}^{N} h_i n_i + \lambda\left(\sum_{i=1}^{N} n_i - Nf\right)\right)\right],$$

where we have introduced $f := \frac{M}{N}$ and $\lambda = i2\pi k$. Note that even though the stationary value of λ is real, we are integrating it along the imaginary axis (using what is known as Bromwich contour, see e.g. [32], appendix C), as always when the integration variable has been introduced as Lagrance multiplyer. Varying λ, we are therefore looking for the infimum, not the supremum. We obtain the "grand-canonical" cumlant-generating function that we will work with:

$$W_\lambda[\boldsymbol{h}] := \ln\left[\sum_{\boldsymbol{s}} \exp\left(\sum_{i<j,} K_{ij} n_i n_j + \sum_{i=1}^{N} \left[h_i n_i + \lambda(n_i - f)\right]\right)\right]$$

$$\lambda_{\boldsymbol{h},\boldsymbol{j},K}(f) \text{ such that } \frac{\partial \mathrm{W}_\lambda[\mathrm{h}]}{\partial \lambda} = Nf.$$

Now we decouple the interacting term by means of a Gaussian helping field

$$\begin{aligned}
&\exp\left(\frac{1}{2}\sum_{i\neq j,} K_{ij} n_i n_j\right) \\
=&\frac{1}{(2\pi)^{\frac{N}{2}}\sqrt{\det(K)}} \int d\boldsymbol{\phi}\, e^{-\frac{1}{2}\boldsymbol{\phi}^{\mathrm{T}} K^{-1} \boldsymbol{\phi} + \sum_i \phi_i n_i}
\end{aligned}$$

and replace the ϕ-integral by another saddle-point approximation:

$$W[\boldsymbol{h}] = \sup_{\boldsymbol{\phi}} \inf_{\lambda} \left[-\frac{1}{2}\boldsymbol{\phi}^{\mathrm{T}} K^{-1} \boldsymbol{\phi} + \sum_i \left(\ln\left(1 + e^{h_i + \lambda + \phi_i}\right) - \lambda f\right)\right]$$

For the examples shown in the figures, we are assuming a rectangular shape for K,

$$K(r_i - r_j) = \begin{cases} K_{\mathrm{rec}}, & \text{for } |r_i - r_j| \leq w_{\mathrm{rec}} \\ 0, & \text{else,} \end{cases} \tag{A1}$$

but this choice is only made for convenience, the theoretical results extend to general shapes.

2. Incorporating disorder

Drawing random connections in addition to the spatially ordered ones additionally modifies the extremizing probability distribution and introduces more contributions to the pairwise covariances. We would like to compute the quenched average of the cumulant-generating functional, so

$$\langle W(\boldsymbol{h})\rangle_J = \int dJ\, P(J) \ln\left[\sum_{\boldsymbol{n}} e^{\frac{1}{2}\sum_{i\neq j}(J_{ij}+K_{ij})n_i n_j + \sum_i (n_i \mathrm{U}(\xi - r_i) + h_i n_i)}\right] \tag{A2}$$

$$= \lim_{n\to 0} \int dJ\, P(J) \left[\frac{-1 + \sum_{\boldsymbol{n}^1,\ldots,\boldsymbol{n}^n} e^{\sum_{i\neq j}(J_{ij}+K_{ij})\sum_{\alpha=1}^n n_i^\alpha n_j^\alpha + \sum_i n_i^\alpha(\mathrm{U}(\xi-r_i)+h_i)}}{n}\right], \tag{A3}$$

where we have used the replica trick to represent the logarithm [19]. As indicated in the main text, eq. (2), we assume that the couplings are uncorrelated and Gaussian, so that, after the standard procedure of introducing appropriate helping fields, we obtain

$$\int dJ\, P(J)\; e^{\sum_{i\neq j} J_{ij}\sum_{\alpha=1}^n n_i^\alpha n_j^\alpha} \tag{A4}$$

$$= \exp\left(-\frac{1}{2}\frac{g^2}{N}\sum_i \sum_{\alpha,\beta} n_i^\alpha n_i^\beta\right) \prod_\alpha \left[\frac{\sqrt{N}}{g\sqrt{2\pi}} \int d\bar{q}_\alpha \exp\left(-\frac{1}{2}\frac{N}{g^2}\bar{q}_\alpha^2 + \bar{q}_\alpha \sum_i n_i^\alpha\right)\right] \tag{A5}$$

$$\times \prod_{\alpha\neq\beta}\left[\frac{\sqrt{N}}{g\sqrt{2\pi}} \int dq_{\alpha\beta}\; \exp\left(-\frac{1}{2}\frac{N}{g^2}q_{\alpha\beta}^2 + q_{\alpha\beta}\sum_i n_i^\alpha n_i^\beta\right)\right]. \tag{A6}$$

We combine this result with the contribution from the network without disorder, but local connectivity and solve the resulting integral, assuming replica-symmetry in q and $\boldsymbol{\phi}$. The validity of the assumption of replica symmetry is validated numerically by comparing our theoretical results with the outcomes of Monte-Carlo computations and discussed in sec. IV. Dropping subleading terms in N, we solve the integrals in saddle-point approximation and obtain for W:

$$\langle W(\boldsymbol{h})\rangle_J = \lim_{n\to 0} \underset{q,\bar{q},\boldsymbol{\phi}}{\mathrm{extr}} \left[-\frac{1}{n} + \frac{e^{-\frac{1}{2}Ng^2 n(n-1)q^2 - \frac{1}{2}Ng^2 n\bar{q}^2} e^{-\frac{1}{2}n\boldsymbol{\phi}^T K^{-1}\boldsymbol{\phi}}}{n}\right. \tag{A7}$$

$$\left. \times \left(\prod_{l,\gamma}\sum_{n_l^\gamma=0}^{1}\right)\left(\prod_k e^{g^2 q\sum_{\alpha\neq\beta} n_k^\alpha n_k^\beta}\right) e^{\sum_i g^2\bar{q}\sum_i\sum_\alpha n_i^\alpha + \sum_{\alpha=1}^n n_i^\alpha \phi^i + \sum_i\sum_{\alpha=1}^n n_i^\alpha(\mathrm{U}(\xi-r_i)+h_i)}\right] \tag{A8}$$

$$= \frac{1}{2}Ng^2q^2 - \frac{1}{2}Ng^2\bar{q}^2 - \frac{1}{2}\boldsymbol{\phi}^T K^{-1}\boldsymbol{\phi} \tag{A9}$$

$$+ \prod_k \frac{1}{\sqrt{2\pi}} \int dt_k\, e^{-\frac{t_k^2}{2}} \sum_i \ln\left[1 + e^{\phi^i + t_i g\sqrt{2q} + \mathrm{U}(\xi-r_i) + g^2(\bar{q}-q) + h_i}\right]. \tag{A10}$$

Taking now into account the restriction on the total activity in addition, the mean-field cumulant-generating functional reads

$$\langle W(\boldsymbol{h})\rangle_J = \underset{q,\bar{q},\boldsymbol{\phi},\lambda}{\mathrm{extr}} \left\{\frac{1}{2}Ng^2q^2 - \frac{1}{2}Ng^2\bar{q}^2 - \frac{1}{2}\boldsymbol{\phi}^T K^{-1}\boldsymbol{\phi} - N\lambda f\right. \tag{A11}$$

$$\left. + \prod_k \frac{1}{\sqrt{2\pi}} \int dt_k\, e^{-\frac{t_k^2}{2}} \sum_i \ln\left[1 + e^{\phi^i + t_i g\sqrt{2q} + (\mathrm{U}(\xi-r_i)) + g^2(\bar{q}-q) + \lambda + h_i}\right]\right\}. \tag{A12}$$

Because we are evaluating this quantity only at its extremal values, we are free to express it in shifted coordinates, $\lambda + g^2(\bar{q}-q) \to \lambda$, in order to simplify our expressions and to get rid of $\bar{q}$, so that we obtain

$$\langle W_I(\boldsymbol{h})\rangle_J = \underset{q,\bar{q},\boldsymbol{\psi},\boldsymbol{\phi},\lambda}{\mathrm{extr}} \left\{\frac{1}{2}Ng^2q^2 - \frac{1}{2}Ng^2\bar{q}^2 - \frac{1}{2}\boldsymbol{\phi}^T K^{-1}\boldsymbol{\phi} - N\left(\lambda - g^2(\bar{q}-q)\right) f.\right. \tag{A13}$$

$$\left. + \prod_k \frac{1}{\sqrt{2\pi}} \int dt_k\, e^{-\frac{t_k^2}{2}} \sum_i \ln\left[1 + e^{\phi^i + t_i g\sqrt{2q} + \mathrm{U}(\xi-r_i) + \lambda + h_i}\right]\right\} \tag{A14}$$

$$=: G_g(\boldsymbol{h}, \boldsymbol{\phi}, q, \lambda), \tag{A15}$$

which leads to the saddle-point equations

$$q = f + \int dx \int Dt \left\{ \frac{1}{\left[1 + e^{-\left(\phi_x + t_x g\sqrt{2q} + \mathrm{U}(\xi - x) + \lambda + h_x\right)}\right]^2} - \frac{1}{\left[1 + e^{-\left(\phi_x + t_x g\sqrt{2q} + \mathrm{U}(\xi - x) + \lambda + h_x\right)}\right]} \right\} \tag{A16}$$

$$\bar{q} = f \tag{A17}$$

$$f = \int dx \int Dt \, \frac{1}{1 + e^{-\left(\phi_x + t_x g\sqrt{2q} + \mathrm{U}(\xi - x) + \lambda + h_x\right)}} dx \tag{A18}$$

$$\phi\left(x\right) = \int K\left(x - y\right) \int Dt \, \frac{1}{1 + e^{-\left(\phi_y + t_y g\sqrt{2q} + \mathrm{U}(\xi - y) + \lambda + h_y\right)}} dy. \tag{A19}$$

Drawing the limit of $N \to \infty$, we have turned the sums over neuron sites into integrals over space, which we indicate by renaming the indices to x and y instead of i and j. Finally, setting $\boldsymbol{h} = 0$ and using eq. (A18) to simplify eq. (A16), we obtain the final saddle-point equations as given in the main text, eqs. (18) to (20). Note that this simplification is valid for the saddle-point values q, $\{\phi_x\}_x$ and λ - however, when taking further derivatives of G with respect to $\boldsymbol{h}$ (as necessary to determine covariances), we have to assume general q, $\{\phi_x\}_x$ and λ (not as given in the saddle point) and therefore have to use the right-hand side of eq. (A16), and not of eq. (18).

3. Analysis of the saddle-point equations in the limit $g \to 0$

The quantities q, ϕ_x and λ are implicitly given by

$$0 = \frac{\partial}{\partial\left[\{\phi_x\}_x, \lambda, q\right]} G_g\left[q, \boldsymbol{\phi}, \lambda\right]. \tag{A20}$$

For $g = 0$, the integrands in the saddle-point equations (18) to (20) get independent of t and we can perform the Gaussian integrals, so that we obtain

$$q = \int dx \, \frac{1}{\left[1 + e^{-(\phi_x + \mathrm{U}(\xi - x)t) + \lambda)}\right]^2}, \tag{A21}$$

$$\phi_x = \int dy \, K\left(x - y\right) \frac{1}{1 + e^{-(\phi_y + \mathrm{U}(\xi - y) + \lambda)}}. \tag{A22}$$

$$f = \int dx \, \frac{1}{1 + e^{-(\phi_x + \mathrm{U}(\xi - r_i) + \lambda)}}, \tag{A23}$$

the latter two corresponding to eqs. (12) - (13) in [17]. In particular, all auxiliary fields have a well-behaved limit for $g \to 0$. Furthermore, from eq. (A20), we obtain the derivatives of the auxiliary variables with repect to g to be given by

$$\frac{\partial}{\partial g} \begin{pmatrix} \{\phi_x\}_x \\ \lambda \\ q \end{pmatrix} = \left(\frac{\partial^2}{\partial\left[\{\phi_x\}_x, \lambda, q\right]^2} G\left[q, \boldsymbol{\phi}, \lambda\right] \right)^{-1} \frac{\partial^2}{\partial g \partial\left[\{\phi_x\}_x, \lambda, q\right]} G_g\left[q, \boldsymbol{\phi}, \lambda\right]. \tag{A24}$$

Further differentiating $\frac{\partial}{\partial\left[\{\phi_x\}_x, \lambda, q\right]} G_g\left[q, \boldsymbol{\phi}, \lambda\right]$ with respect to g yields

$$\frac{\partial^2}{\partial g \partial\left[\{\phi_x\}_x, \lambda, q\right]} G_g\left[q, \boldsymbol{\phi}, \lambda\right] = \begin{pmatrix} \left\{ \int dy \, K_{xy} \int Dt_y \, t_y \sqrt{2q} \cdot m_y \left(1 - 2m_y\right) \right\}_x \\ \int dy \int \mathcal{D}t_y \, t_y \sqrt{2q} \cdot m_y \left(1 - 2m_y\right) \\ \int dx \int Dt_y \, t_y \sqrt{2q} \cdot m_y \left(1 - 2m_y\right) \left(1 - m_y\right) \end{pmatrix} \overset{g=0}{=} 0, \tag{A25}$$

with m_x as introduced in eq. (27). The last equality in eq. (A25) holds because for $g = 0$, m_x is independent of t_x and the remaining t_x-integrand is antisymmetric. To obtain the derivatives of the order parameters at $g = 0$, we therefore only have to check that differentiating $\partial_{\{\phi_x\}_x, \lambda, q} G$ with respect to q, ϕ_x and λ once more yields a regular Hessian. We obtain

$$\frac{\partial^2}{\partial\left[\{\phi_x\}_x, \lambda, q\right]^2} G_g\left[q, \boldsymbol{\phi}, \lambda\right] = \begin{pmatrix} -\left(K^{-1}\right)_{xy} + \delta_{xy} v_y & v_x & g^2 \kappa_x^3 \\ v_y & \int dz \, v_z & g^2 \int dz \, \kappa_z^3 \\ g^2 \kappa_y^3 & g^2 \int dz \, \kappa_z^3 & g^2 + g^4 \int dz \, \kappa_z^4 \end{pmatrix}, \tag{A26}$$

where we have omitted the t_x-dependence of m_x for brevity and have introduced the higher-order cumulants

$$\kappa_x^3 \coloneqq \frac{\partial^2 m_x}{\partial \phi_x^2} = \int \mathcal{D}t\, m_x \left(1-2m_x\right)\left(1-m_x\right) \tag{A27}$$

$$\kappa_x^4 \coloneqq \frac{\partial^3 m_x}{\partial \phi_i^3} = \int \mathcal{D}t\, m_x \left(1-m_x\right)\left(1-6m_x+6m_x^2\right), \tag{A28}$$

(the fourth-order one for later use). It is not apparent why the Hessian should have a zero mode - indeed, this would mean in particular that the saddle-point approximation is not well-defined. So as long as we trust the saddle-point approximation, we also know that the derivatives of the order-parameters with respect to g vanish for $g = 0$. Also, in appendix B, we numerically confirm that the Hessian does not have zero modes for $g \to 0$.

Appendix B: Computing the Fisher information

First, we convince ourselves that computing the Fisher information simply amounts to computing the covariance matrix. Consider the probability distribution for the neural network state $\boldsymbol{n}$, conditioned on the stimulus ξ, $P\left(\boldsymbol{n}|\xi\right)$, given by

$$P\left(\boldsymbol{n}|\xi\right) = \frac{1}{\mathcal{Z}_J\left(\xi\right)} e^{\sum_i n_i \mathrm{U}(\xi - r_i) + \sum_{i<j}(J_{ij}+K_{ij})n_i n_j},$$

and

$$\mathcal{Z}_J\left(\xi\right) = \sum_{\boldsymbol{n}} e^{\sum_i n_i \mathrm{U}(\xi - r_i) + \sum_{i<j}(J_{ij}+K_{ij})n_i n_j}.$$

For the Fisher information, we need the second derivative of the logarithm of P with respect to ξ:

$$\begin{aligned}
-\frac{\partial^2}{\partial \xi^2}\ln\left(P\left(\boldsymbol{n}|\xi\right)\right) &= -\frac{\partial^2}{\partial \xi^2}\sum_i n_i \mathrm{U}\left(\xi - r_i\right) + \frac{\partial^2}{\partial \xi^2}\ln \mathcal{Z}_J\left(\xi\right) \\
&= -\frac{\partial}{\partial \xi}\sum_i n_i \mathrm{U}'\left(\xi - r_i\right) + \frac{\partial}{\partial \xi}\frac{\frac{\partial}{\partial \xi}\mathcal{Z}_J\left(\xi\right)}{\mathcal{Z}_J\left(\xi\right)} \\
&= -\sum_i n_i \mathrm{U}''\left(\xi - r_i\right) + \frac{\frac{\partial^2}{\partial \xi^2}\mathcal{Z}_J\left(\xi\right)}{\mathcal{Z}_J\left(\xi\right)} - \left(\frac{\frac{\partial}{\partial \xi}\mathcal{Z}_J\left(\xi\right)}{\mathcal{Z}_J\left(\xi\right)}\right)^2.
\end{aligned}$$

Upon averaging over the neurons states of the neuron n_i, we obtain

$$\begin{aligned}
\mathcal{I}_{\boldsymbol{n}}\left(\xi\right) = \Bigg\langle & -\frac{1}{\mathcal{Z}_J\left(\xi\right)}\sum_{\boldsymbol{n}} e^{\sum_i n_i \mathrm{U}(\xi - r_i) + \sum_{i<j} J_{ij} n_i n_j} \sum_i n_i \mathrm{U}''\left(\xi - r_i\right) \\
& + \frac{1}{\mathcal{Z}_J\left(\xi\right)}\sum_{\boldsymbol{n}} e^{\sum_i n_i \mathrm{U}(\xi - r_i) + \sum_{i<j} J_{ij} n_i n_j} \left[\sum_i n_i \mathrm{U}''\left(\xi - r_i\right) + \left(\sum_i n_i \mathrm{U}'\left(\xi - r_i\right)\right)^2\right] \\
& - \left(\frac{1}{\mathcal{Z}_J\left(\xi\right)}\sum_{\boldsymbol{n}} e^{\sum_i n_i \mathrm{U}(\xi - r_i) + \sum_{i<j} J_{ij} n_i n_j} \sum_i n_i \mathrm{U}'\left(\xi - r_i\right)\right)^2 \Bigg\rangle_J
\end{aligned} \tag{B1}$$

$$= \sum_{i,j} \mathrm{U}'\left(\xi - r_i\right) \left\langle \left[\left\langle n_i n_j \right\rangle_{\boldsymbol{n}} - \left\langle n_i \right\rangle_{\boldsymbol{n}} \left\langle n_j \right\rangle_{\boldsymbol{n}}\right] \right\rangle_J \mathrm{U}'\left(\xi - r_j\right), \tag{B2}$$

where we have used the usual thermal average

$$\left\langle f\left(\boldsymbol{n}\right)\right\rangle_{\boldsymbol{n}} \coloneqq \frac{1}{\mathcal{Z}_J\left(\xi\right)}\sum_{\boldsymbol{n}} f\left(\boldsymbol{n}\right) e^{\sum_i n_i \mathrm{U}(\xi - r_i) + \sum_{i<j} J_{ij} n_i n_j}. \tag{B3}$$

As indicated before, to determine the Fisher information, we therefore just have to compute the covariance matrix, which we achieve by differentiating the cumulant-generating functional twice with respect to $\boldsymbol{h}$, considering all indirect

dependencies via the auxiliary fields (evaluated at their respective saddle-point values). Taking into account both the fixed total activity and the disorder, the cumulant-generating functional is given by eq. (A15). Formally differentiating this expression yields

$$\frac{d^2}{d\boldsymbol{h}^2}\left\langle W_f\left(\boldsymbol{h}\right)\right\rangle_J = \frac{\partial^2 G}{\partial \boldsymbol{h}^2} + 2\frac{\partial^2 G}{\partial \boldsymbol{h}\partial \boldsymbol{\phi}}\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{h}} + 2\frac{\partial^2 G}{\partial \boldsymbol{h}\partial q}\frac{\partial q}{\partial \boldsymbol{h}} + 2\frac{\partial^2 G}{\partial \boldsymbol{h}\partial \lambda}\frac{\partial \lambda}{\partial \boldsymbol{h}} \tag{B4}$$

$$+ \frac{\partial^2 G}{\partial q^2}\left(\frac{\partial q}{\partial \boldsymbol{h}}\right)^2 + \frac{\partial^2 G}{\partial \lambda^2}\left(\frac{\partial \lambda}{\partial \boldsymbol{h}}\right)^2 + \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{h}}\frac{\partial^2 G}{\partial \boldsymbol{\phi}^2}\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{h}} \tag{B5}$$

$$+ 2\frac{\partial^2 G}{\partial q\partial \lambda}\frac{\partial q}{\partial \boldsymbol{h}}\frac{\partial \lambda}{\partial \boldsymbol{h}} + 2\frac{\partial^2 G}{\partial q\partial \boldsymbol{\phi}}\frac{\partial q}{\partial \boldsymbol{h}}\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{h}} + 2\frac{\partial^2 G}{\partial \lambda\partial \boldsymbol{\phi}}\frac{\partial \lambda}{\partial \boldsymbol{h}}\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{h}}. \tag{B6}$$

We obtain the derivatives of q and $\boldsymbol{\phi}$ by taking the total derivatives of their defining equations, i.e. (A16) and (A19), which yields

$$0 = \frac{d}{d\boldsymbol{h}}\frac{\partial}{\partial q}G\left(\boldsymbol{h},\boldsymbol{\phi},q,\lambda\right) = \frac{\partial^2 G}{\partial q^2}\frac{\partial q}{\partial \boldsymbol{h}} + \frac{\partial^2 G}{\partial \lambda\partial q}\frac{\partial \lambda}{\partial \boldsymbol{h}} + \frac{\partial^2 G}{\partial \boldsymbol{\phi}\partial q}\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{h}} + \frac{\partial^2 G}{\partial q\partial \boldsymbol{h}} \tag{B7}$$

$$0 = \frac{d}{d\boldsymbol{h}}\frac{\partial}{\partial \lambda}G\left(\boldsymbol{h},\boldsymbol{\phi},q,\lambda\right) = \frac{\partial^2 G}{\partial q\partial \lambda}\frac{\partial q}{\partial \boldsymbol{h}} + \frac{\partial^2 G}{\partial \lambda^2}\frac{\partial \lambda}{\partial \boldsymbol{h}} + \frac{\partial^2 G}{\partial \boldsymbol{\phi}\partial \lambda}\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{h}} + \frac{\partial^2 G}{\partial \lambda\partial \boldsymbol{h}} \tag{B8}$$

$$0 = \frac{d}{d\boldsymbol{h}}\frac{\partial}{\partial \boldsymbol{\phi}}G\left(\boldsymbol{h},\boldsymbol{\phi},q,\lambda\right) = \frac{\partial^2 G}{\partial q\partial \boldsymbol{\phi}}\frac{\partial q}{\partial \boldsymbol{h}} + \frac{\partial^2 G}{\partial \lambda\partial \boldsymbol{\phi}}\frac{\partial \lambda}{\partial \boldsymbol{h}} + \frac{\partial^2 G}{\partial \boldsymbol{\phi}^2}\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{h}} + \frac{\partial^2 G}{\partial \boldsymbol{\phi}\partial \boldsymbol{h}}, \tag{B9}$$

so that we obtain after inserting into (B6)

$$\frac{d^2}{d\boldsymbol{h}^2}\left\langle W_f\left(\boldsymbol{h}\right)\right\rangle_J = \frac{\partial^2 G}{\partial \boldsymbol{h}^2} - \begin{pmatrix} \frac{\partial^2 G}{\partial \boldsymbol{h}\partial \boldsymbol{\phi}} \\ \frac{\partial^2 G}{\partial \boldsymbol{h}\partial q} \\ \frac{\partial^2 G}{\partial \boldsymbol{h}\partial \lambda} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \frac{\partial^2 G}{\partial \boldsymbol{\phi}^2} & \frac{\partial^2 G}{\partial \boldsymbol{\phi}\partial q} & \frac{\partial^2 G}{\partial \boldsymbol{\phi}\partial \lambda} \\ \frac{\partial^2 G}{\partial q\partial \boldsymbol{\phi}} & \frac{\partial^2 G}{\partial q^2} & \frac{\partial^2 G}{\partial q\partial \lambda} \\ \frac{\partial^2 G}{\partial \lambda\partial \boldsymbol{\phi}} & \frac{\partial^2 G}{\partial \lambda\partial q} & \frac{\partial^2 G}{\partial \lambda^2} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial^2 G}{\partial \boldsymbol{\phi}\partial \boldsymbol{h}} \\ \frac{\partial^2 G}{\partial q\partial \boldsymbol{h}} \\ \frac{\partial^2 G}{\partial \lambda\partial \boldsymbol{h}} \end{pmatrix}. \tag{B10}$$

In order to compactly write down the entries of the matrix and the vectors above, we introduce the effective local connectivity

$$\left(K_{\mathrm{eff}}^{-1}\right)_{xy} \coloneqq -\frac{\partial^2 G}{\partial \phi_x \partial \phi_y} \Leftrightarrow \left[\left(\frac{\partial^2 G}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}}\right)^{-1}\right]_{xy} = -\left(K_{\mathrm{eff}}\right)_{xy}, \tag{B11}$$

which fulfills the Dyson equation whose concrete form we obtain by performing the derivatives of G explicitly:

$$\left(K_{\mathrm{eff}}^{-1}\right)_{xy} = \left(K^{-1}\right)_{xy} - \delta_{xy}v_x \tag{B12}$$

$$\Leftrightarrow K_{xy}^{\mathrm{eff}} = K_{xy} + \int K_{xz}v_z K_{zy}^{\mathrm{eff}}. \tag{B13}$$

Using the identities derived in appendix C, in particular eq. (C9), we can note the final form of the covariance matrix:

$$C = V + VK_{\mathrm{eff}}V - \left(\mathbb{1}_N + VK_{\mathrm{eff}}\right)\left(g\boldsymbol{\kappa}^3,\ \boldsymbol{v}\right)S^{-1}\begin{pmatrix} g\boldsymbol{\kappa}^3 \\ \boldsymbol{v} \end{pmatrix}\left(\mathbb{1}_N + K_{\mathrm{eff}}V\right), \tag{B14}$$

where V is the diagonal matrix with the disorder-averaged variances v_i and

$$S = \begin{pmatrix} N + g^2\sum_i \kappa_i^4 & g\sum_i \kappa_i^3 \\ g\sum_i \kappa_i^3 & \sum_i v_i \end{pmatrix} + \begin{pmatrix} g\boldsymbol{\kappa}^3 \\ \boldsymbol{v} \end{pmatrix} K_{\mathrm{eff}}\left(g\boldsymbol{\kappa}^3,\ \boldsymbol{v}\right) \tag{B15}$$

The Fisher information, finally is then given by

$$\mathcal{I}_{\boldsymbol{n}}\left(\xi\right) = \sum_{x,y}\mathrm{U}'\left(\xi - x\right)C_{xy}\mathrm{U}'\left(\xi - y\right). \tag{B16}$$

The last term from B14 contributes to the Fisher information with a term containing (twice) the expression

$$\begin{pmatrix} g\boldsymbol{\kappa}^3 \\ \boldsymbol{v} \end{pmatrix}\left(\mathbb{1}_N + K_{\mathrm{eff}}V\right)\boldsymbol{U}' \tag{B17}$$

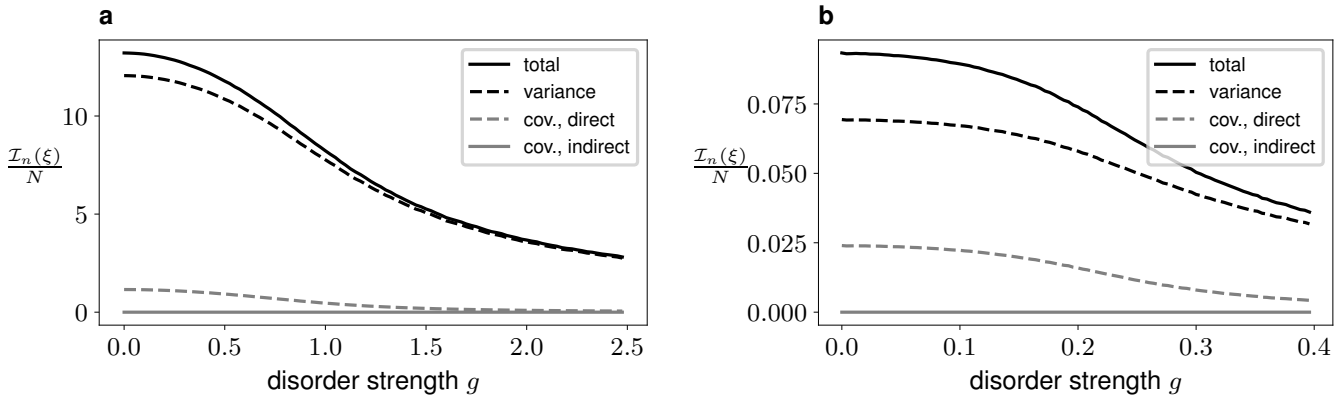

FIG. 6: The Fisher information per neuron in dependence of the disorder, contributions from different parts of covariance as defined in eq. (23). Parameters as in fig. (3).

The space-dependence of the contribution from the covariance is mostly determined by the shape of U (multiplications by K or K_{eff} merely smear it out), so that multiplication with its derivative with U' is well approximated by a spatial derivative and summation over space and therefore yields a contribution close to 0. This part of the covariance therefore only yields subleading contributions to the Fisher information (see fig. (6)) and we can neglect it in the analysis. This makes sense because it emerges from the source-dependence of q and λ, the auxiliary variables representing the disorder and the global inhibition, which are global quantities. It is therefore expected that their contribution to the spatial information is negligible.

1. Analysis of the covariance matrix

Having an analytical expression for the covariance matrix at hand, we can investigate its behavior for special cases, in particular around $g = 0$. Because it depends on g only via the cumulants v, κ^3 and κ^4, we primarily have to examine their behavior near $g = 0$. We observe that

$$\frac{d}{dg}v_x = \frac{\partial}{\partial g}v_x + \frac{\partial v_x}{\partial q}\frac{\partial q}{\partial g} + \int dy\, \frac{\partial v_x}{\partial \phi_y}\frac{\partial \phi_y}{\partial g} + \frac{\partial v_x}{\partial \lambda}\frac{\partial \lambda}{\partial g} \tag{B18}$$

$$= \frac{\partial}{\partial g}\int \mathcal{D}t\, m_x\,(1-m_x) = \int \mathcal{D}t\, t\sqrt{2q}\, m_x\left(1-3m_x+2m_x^2\right) \overset{g=0}{=} 0, \tag{B19}$$

where we have used the result from appendix (A 3) that the derivatives of the auxiliary variables with respect to g vanish as g goes to 0. Again because m_x does not depend on t_x for $g = 0$ and the remaining integral over t_x is antisymmetric, it also yields 0. With the same argument, the derivatives of the other cumulants vanish as well. Therefore, the linear orders of all g-dependent quantities, that the covariance C depends on, vanish. Thus, the derivatives of the covariances and of the Fisher information equal 0 for $g = 0$ as well, as apparent from the plots in figure (3).

2. Relating inputs and tuning curves by means of K_{eff}

Without disorder, the tuning curve T in the thermodynamic limit is given by

$$\mathrm{T}_x = \int dy\, K\,(x-y)\,\frac{1}{1+e^{-(\phi_y+\mathrm{U}(\xi-y)+\lambda)}} \tag{B20}$$

$$\phi_y = \int dx\, K\,(y-z)\,\mathrm{T}_z \tag{B21}$$

and therefore, we can write for its derivative

$$\mathbf{T}' = \mathbf{T}\,(1-\mathbf{T})\left(\mathbf{U}' + K\mathbf{T}'\right) \tag{B22}$$
$$\Leftrightarrow \left[1 - \mathbf{T}\,(1-\mathbf{T})\,K\right]\mathbf{T}' = \mathbf{T}\,(1-\mathbf{T})\,\mathbf{U}' \tag{B23}$$
$$\mathbf{T}' = (1-VK)^{-1}\,V\mathbf{U}' = \left(V^{-1}-K\right)^{-1}\mathbf{U}' \tag{B24}$$
$$\Leftrightarrow \mathbf{T}' = V\left(1 + K\left(V^{-1}-K\right)^{-1}\right)\mathbf{U}', \tag{B25}$$

where we have abbreviated $V_{ij} = \delta_{ij} f_i\,(1-f_i)$. We furthermore have

$$K_{\text{eff}} = K + KVK_{\text{eff}} \tag{B26}$$
$$\Leftrightarrow (1-KV)\,K_{\text{eff}} = K \tag{B27}$$
$$\Leftrightarrow K_{\text{eff}} = (1-KV)^{-1}\,K = V^{-1}\left(V^{-1}-K\right)^{-1}K \tag{B28}$$
$$\Leftrightarrow K_{\text{eff}} = K\left(V^{-1}-K\right)^{-1}V^{-1} \tag{B29}$$
$$\Leftrightarrow K_{\text{eff}}V = K\left(V^{-1}-K\right)^{-1}, \tag{B30}$$

where we obtained the second-to-last equivalence by transposing. Inserting this expression into eq. (B25), we arrive at eq. (30).

Appendix C: Matrix-vector calculus

1. Inversion of a matrix with blocks on the diagonal of the sizes N and M

Assume we have a matrix of the form

$$U := \begin{pmatrix} A & b \\ b^{\mathrm{T}} & a \end{pmatrix}, \tag{C1}$$

where

$$A \in \mathbb{R}^{N\times N},\ b \in \mathbb{R}^{N\times M},\ a \in \mathbb{R}^{M\times M}, \tag{C2}$$

a and A are symmetric and A is invertible. To invert it, we make the ansatz

$$V := \begin{pmatrix} C & d \\ d^{\mathrm{T}} & c \end{pmatrix}. \tag{C3}$$

Multiplying U and V, we obtain the conditions

$$AC + bd^{\mathrm{T}} = \mathbb{1}_N \tag{C4}$$
$$Ad + bc = 0 \tag{C5}$$
$$b^{\mathrm{T}}C + ad^{\mathrm{T}} = 0 \tag{C6}$$
$$b^{\mathrm{T}}d + ac = \mathbb{1}_M \tag{C7}$$

Solving (C5) for d and inserting into (C7), we obtain

$$c = \left(a - b^{\mathrm{T}}A^{-1}b\right)^{-1}$$
$$\text{and } d = -A^{-1}b\left(a - b^{\mathrm{T}}A^{-1}b\right)^{-1}.$$

Solving (C4) for C and inserting the results gained until here, we obtain

$$C = A^{-1} + A^{-1}b\left(a - b^{\mathrm{T}}A^{-1}b\right)^{-1}\left(A^{-1}b\right)^{\mathrm{T}}.$$

Plugging these results into the left-hand side of (C6), which we did not use so far, we obtain

$$\begin{aligned}
&b^{\mathrm{T}}\left(A^{-1}+A^{-1}b\left(a-b^{\mathrm{T}}A^{-1}b\right)^{-1}\left(A^{-1}b\right)^{\mathrm{T}}\right)-a\left[A^{-1}b\left(a-b^{\mathrm{T}}A^{-1}b\right)^{-1}\right]^{\mathrm{T}}\\
=&\left(\left(A^{-1}b\right)^{\mathrm{T}}+\left(b^{\mathrm{T}}A^{-1}b-a+a\right)\left(a-b^{\mathrm{T}}A^{-1}b\right)^{-1}\left(A^{-1}b\right)^{\mathrm{T}}\right)-a\left(a-b^{\mathrm{T}}A^{-1}b\right)^{-1}\left(A^{-1}b\right)^{\mathrm{T}}\\
=&a\left(a-b^{\mathrm{T}}A^{-1}b\right)^{-1}\left(A^{-1}b\right)^{\mathrm{T}}-a\left(a-b^{\mathrm{T}}A^{-1}b\right)^{-1}\left(A^{-1}b\right)^{\mathrm{T}}=0,
\end{aligned}$$

therefore our ansatz is consistent. Summarizing, we can write the inverse of U as

$$U^{-1}=\begin{pmatrix}A^{-1} & 0\\ 0 & 0\end{pmatrix}+\begin{pmatrix}-A^{-1}b\\ \mathbb{1}_M\end{pmatrix}\left(a-b^{\mathrm{T}}A^{-1}b\right)^{-1}\left(-\left(A^{-1}b\right)^{\mathrm{T}}\ ,\mathbb{1}_M\right). \tag{C8}$$

2. Vector-matrix-vectors multiplication

Calculating cross-covariances, we are interested in calculating objects of the type

$$\begin{pmatrix}B & b\end{pmatrix}\begin{pmatrix}A & b\\ b^{\mathrm{T}} & a\end{pmatrix}^{-1}\begin{pmatrix}B\\ b^{\mathrm{T}}\end{pmatrix}.$$

Making use of (C8), we then obtain

$$\begin{aligned}
&\begin{pmatrix}B & b\end{pmatrix}\begin{pmatrix}A & b\\ b^{\mathrm{T}} & a\end{pmatrix}^{-1}\begin{pmatrix}B\\ b^{\mathrm{T}}\end{pmatrix}\\
=&BA^{-1}B+\left(\mathbb{1}_N-BA^{-1}\right)b\left(a-b^{\mathrm{T}}A^{-1}b\right)^{-1}b^{\mathrm{T}}\left(\mathbb{1}_N-A^{-1}B\right)
\end{aligned} \tag{C9}$$

Appendix D: Relating Fisher and mutual information for uncoupled neurons with inhomogeneously distributed place fields

Here, we consider independent neurons, but allow variability in the tuning curves T. The probability distribution of the neural population is then given by

$$P\left(\boldsymbol{n}\right)=\prod_{i=1}^{N}\left[n_i\mathrm{T}_i\left(\xi\right)+\left(1-n_i\right)\left(1-\mathrm{T}_i\left(\xi\right)\right)\right]P_\xi\left(\xi\right). \tag{D1}$$

To compute the mutual information, we first need to compute the entropy of this distribution, which is given by

$$h_{\mathrm{uncond}}=-\lim_{N\to\infty}\frac{1}{N}\sum_{\boldsymbol{n}}\left\langle P\left(\boldsymbol{n}\right)\ln\left(P\left(\boldsymbol{n}\right)\right)\right\rangle_{\mathrm{T}},$$

where we denote by $\langle\ldots\rangle_{\mathrm{T}}$ the average over the variability of the tuning curves. The tricky part here is that we average over a logarithm, a complication that we deal with by introducing replicas ($n+1$ in this case because of the prefactor $P\left(\boldsymbol{n}\right)$, compare [33]), which leads to

$$h_{\mathrm{uncond}}=-\lim_{N\to\infty}\frac{1}{N}\lim_{k\to0}\frac{1}{k}\left\{\int\prod_{\alpha=0}^{k}\left(d\xi_\alpha P_\xi\left(\xi_\alpha\right)\right)\left(\left\langle\left[\prod_{\alpha=0}^{k}\mathrm{T}\left(\xi_\alpha\right)+\prod_{\alpha=0}^{k}\left(1-\mathrm{T}\left(\xi_\alpha\right)\right)\right]\right\rangle_{\mathrm{T}}\right)^{N}-1\right\} \tag{D2}$$

$$=-\lim_{N\to\infty}\frac{1}{N}\lim_{k\to0}\frac{1}{k}\left\{\int\prod_{\alpha=0}^{k}\left(d\xi_\alpha P_\xi\left(\xi_\alpha\right)\right)\left(G_{\mathrm{T}}\left(\boldsymbol{\xi}\right)\right)^{N}-1\right\}, \tag{D3}$$

where we have introduced

$$G_{\mathrm{T}}\left(\boldsymbol{\xi}\right)\coloneqq\left\langle\left[\prod_{\alpha=0}^{n}\mathrm{T}\left(\xi_\alpha\right)+\prod_{\alpha=0}^{n}\left(1-\mathrm{T}\left(\xi_\alpha\right)\right)\right]\right\rangle_{\mathrm{T}}.$$

An obvious idea is now to evaluate eq. (D3) in saddle-point approximation, as also shown in [14, 15]. This is indeed what we will do, but with a small twist because one of the eigenvalues of the Hessian of G_{T} vanishes for $n \to 0$. However, this replicon mode can be identified to be the one corresponding to the replica-symmetric direction. This allows us to transform the $n+1$-dimensional integral over the ξ_α such that the first coordinate corresponds to the replica-symmetric direction $(1, \ldots, 1)$ and the other n are orthogonal to it. Like this, we can perform the integral over the first coordinate exactly and only the orthogonal directions are evaluated in saddle-point approximation. Having determined the unconditioned entropy in this way, we obtain the mutual information $\mathrm{MI} = h_{\mathrm{uncond}} - h_{\mathrm{cond}}$ by subtracting

$$h_{\mathrm{cond}} = -\lim_{N\to\infty} \frac{1}{N} \sum_{\boldsymbol{n}} \int d\xi \ \left\langle P\left(\boldsymbol{n}|\xi\right) \ln\left(P\left(\boldsymbol{n}|\xi\right)\right)\right\rangle_{\mathrm{T}} \tag{D4}$$

from the unconditioned entropy. Performing the limit of $k \to 0$ in eq. (D3), we see that, to zeroth order, h_{uncond} equals h_{cond}, so that the mutual information is, to first order, given by the one-loop correction

$$I_{\mathrm{MI}} = \int d\xi P_\xi\left(\xi\right) \left[\frac{1}{2} \ln\left(-\frac{N\lambda_{\mathrm{T}}^{1,k=0}\left(\xi\right)}{2\pi}\right) - \frac{1}{2} - \ln\left(P_\xi\left(\xi\right)\right)\right] + \mathcal{O}\left(\frac{1}{N}\right), \tag{D5}$$

where $\lambda_{\mathrm{T}}^{1,k=0}$ is the n-fold degenerate eigenvalue of G_{T} at $k=0$ (see below). Note that our computation neither requires the introduction of helping fields to perform the average over the state space of the neural population, as [15], nor do we assume it to be normally distributed, as in [14]. However, in return, we are assuming the neurons to be independent, which limits the applicability of our approach.
What is left to do is the computation of the Hessian of G_{T}. On the replica-symmetric line, we only have two values for its entries, the diagonal and the off-diagonal. We calculate

$$\left.\frac{\partial^2 G_{\mathrm{T}}}{\partial \xi_\alpha^2}\right|_{\xi_0=\cdots=\xi_n=\xi} = \sum_{n=0,1} \left\langle \prod_{\gamma=0,\gamma\neq\alpha}^{k} \left[n\mathrm{T}\left(\xi_\gamma\right)\left(1-n\right)\left(1-\mathrm{T}\left(\xi_\gamma\right)\right)\right] \left(2n-1\right)\mathrm{T}''\left(\xi_\alpha\right)\right\rangle_r \Bigg|_{\xi_0=\cdots=\xi_n=\xi} \tag{D6}$$

$$= \sum_{n=0,1} \left(2n-1\right) \left\langle \left[n\mathrm{T}\left(\xi\right) + \left(1-n\right)\left(1-\mathrm{T}\left(\xi\right)\right)\right]^k \mathrm{T}''\left(\xi\right)\right\rangle_r \tag{D7}$$

$$\overset{k=0}{=} 0. \tag{D8}$$

and

$$\left.\frac{\partial^2 G_{\mathrm{T}}}{\partial \xi_\alpha \partial \xi_\beta}\right|_{\xi_0=\cdots=\xi_n=\xi} \tag{D9}$$

$$= \sum_{n=0,1} \left\langle \prod_{\gamma=0,\gamma\neq\alpha,\beta}^{k} \left[n\mathrm{T}\left(\xi_\gamma\right) + \left(1-n\right)\left(1-\mathrm{T}\left(\xi_\gamma\right)\right)\right] \left(2n-1\right)^2 \mathrm{T}'\left(\xi_\alpha\right)\mathrm{T}'\left(\xi_\beta\right)\right\rangle_r \Bigg|_{\xi_0=\cdots=\xi_n=\xi} \tag{D10}$$

$$= \sum_{n=0,1} \left\langle \left[n\mathrm{T}\left(\xi\right) + \left(1-n\right)\left(1-\mathrm{T}\left(\xi\right)\right)\right]^{k-1} \left[\mathrm{T}'\left(\xi\right)\right]^2\right\rangle_r \tag{D11}$$

$$\overset{k=0}{=} \left\langle \left[\frac{1}{\mathrm{T}\left(\xi\right)} + \frac{1}{1-\mathrm{T}\left(\xi\right)}\right] \left[\mathrm{T}'\left(\xi\right)\right]^2\right\rangle_r = \left\langle \frac{\left[\mathrm{T}'\left(\xi\right)\right]^2}{\mathrm{T}\left(\xi\right)\left(1-\mathrm{T}\left(\xi\right)\right)}\right\rangle_r. \tag{D12}$$

The eigenvalues of the Hessian of G_{T} are given by

$$\lambda_0\left(\xi\right) = \left.\frac{\partial^2 G}{\partial \xi_\alpha^2}\right|_{\xi_0=\cdots=\xi_n=\xi} + n \left.\frac{\partial^2 G}{\partial \xi_\alpha \partial \xi_\beta}\right|_{\xi_0=\cdots=\xi_n=\xi}$$

$$\lambda_1\left(\xi\right) = \left.\frac{\partial^2 G}{\partial \xi_\alpha^2}\right|_{\xi_0=\cdots=\xi_n=\xi} - \left.\frac{\partial^2 G}{\partial \xi_\alpha \partial \xi_\beta}\right|_{\xi_0=\cdots=\xi_n=\xi},$$

where the first one is non-degenerate, whereas the second one is n-fold degenerate. Inserting eqs. (D8) and (D12), we obtain that, for $k = 0$,

$$\lambda_0(\xi) \overset{k=0}{=} 0 \tag{D13}$$

$$\lambda_1(\xi) \overset{k=0}{=} -\left\langle \frac{\left[\mathrm{T}'(\xi)\right]^2}{\mathrm{T}(\xi)\left(1-\mathrm{T}(\xi)\right)} \right\rangle_r , \tag{D14}$$

$$\tag{D15}$$

where the latter expression equals minus the Fisher information $\mathcal{I}_{\boldsymbol{n}}(\xi)$ for the stimulus ξ. Therefore, inserting this result into eq. (D5), we finally obtain

$$I_{\mathrm{MI}} = \frac{1}{2}\left\langle \ln\left(\frac{N\mathcal{I}_{\boldsymbol{n}}(\xi)}{2\pi}\right)\right\rangle_{\xi\sim P_\xi} - \frac{1}{2} - \left\langle \ln\left(P_\xi(\xi)\right)\right\rangle_{\xi\sim P_\xi} + \mathcal{O}\left(\frac{1}{N}\right), \tag{D16}$$

as expected according to [15].

[1] S. Amari, Dynamics of pattern formation in lateral-inhibition type neural fields, Biological Cybernetics **27**, 77 (1977).
[2] R. Ben-Yishai, R. L. Bar-Or, and H. Sompolinsky, Theory of orientation tuning in visual cortex, Proceedings of the National Academy of Sciences of the United States of America **92** (1995).
[3] M. Tsodyks and T. Sejnowski, Associative memory and hippocampal place cells, International journal of neural systems **6**, 81 (1995).
[4] F. P. Battaglia and A. Treves, Attractor neural networks storing multiple space representations: a model for hippocampal place fields, Physical Review E **58**, 7738 (1998).
[5] S. Romani and M. Tsodyks, Continuous attractors with morphed/correlated maps, PLoS computational biology **6**, e1000869 (2010).
[6] A. Battista and R. Monasson, Capacity-resolution trade-off in the optimal learning of multiple low-dimensional manifolds by attractor neural networks, Phys. Rev. Lett. **124**, 048302 (2020).
[7] L. Posani, S. Cocco, and R. Monasson, Integration and multiplexing of positional and contextual information by the hippocampal network, PLoS computational biology **14**, e1006320 (2018).
[8] K. Yoon, M. A. Buice, C. Barry, R. Hayman, N. Burgess, and I. R. Fiete, Specific evidence of low-dimensional continuous attractor dynamics in grid cells, Nature Neuroscience **16**, 1077 (2013).
[9] K. Wimmer, D. Q. Nykamp, C. Constantinidis, and A. Compte, Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory, Nature neuroscience **17**, 431 (2014).
[10] S. S. Kim, H. Rouault, D. Shaul, and V. Jayaraman, Ring attractor dynamics in the drosophila central brain, Science **356**, 849 (2017).
[11] H. S. Seung and H. Sompolinsky, Simple models for reading neuronal population codes, Proceedings of the National Academy of Sciences **90**, 10749 (1993), https://www.pnas.org/content/90/22/10749.full.pdf.
[12] A. Pouget, S. Deneve, J.-C. Ducom, and P. E. Latham, Narrow Versus Wide Tuning Curves: What's Best for a Population Code?, Neural Computation **11**, 85 (1999), https://direct.mit.edu/neco/article-pdf/11/1/85/814038/089976699300016818.pdf.
[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
[14] H. Sompolinsky, H. Yoon, K. Kang, and M. Shamir, Population coding in neuronal systems with correlated noise, Physical Review E **64**, 051904 (2001).
[15] N. Brunel and J.-P. Nadal, Mutual information, fisher information, and population coding, Neural Computation **10**, 1731 (1998), https://doi.org/10.1162/089976698300017115.
[16] N. Kriegeskorte and X.-X. Wei, Neural tuning and representational geometry, Nature Reviews Neuroscience **22**, 703 (2021).
[17] R. Monasson and S. Rosay, Crosstalk and transitions between multiple spatial maps in an attractor neural network model of the hippocampus: Phase diagram, Phys. Rev. E **87**, 062813 (2013).
[18] X.-X. Wei and A. A. Stocker, Mutual Information, Fisher Information, and Efficient Coding, Neural Computation **28**, 305 (2016), https://direct.mit.edu/neco/article-pdf/28/2/305/955081/neco_a_00804.pdf.
[19] M. Mézard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond (World Scientific Lecture Notes in Physics, Vol 9)* (World Scientific Publishing Company, 1987).
[20] D. Sherrington and S. Kirkpatrick, Solvable model of a spin-glass, Physical Review Letters **35**, 1792 (1975).
[21] T. Kühn, *https://github.com/tobikausk/spacentation_code* (2023).
[22] E. I. Moser, Y. Roudi, M. P. Witter, C. Kentros, T. Bonhoeffer, and M.-B. Moser, Grid cells and cortical representation, Nature Reviews Neuroscience **15**, 466 (2014).
[23] R. Monasson and S. Rosay, Crosstalk and transitions between multiple spatial maps in an attractor neural network model of the hippocampus: Collective motion of the activity, Phys. Rev. E **89**, 032803 (2014).

[24] L. M. Richter and J. Gjorgjieva, Understanding neural circuit development through theory and models, Current Opinion in Neurobiology **46**, 39 (2017).
[25] H. Ko, L. Cossell, C. Baragli, J. Antolik, C. Clopath, S. B. Hofer, and T. D. Mrsic-Flogel, The emergence of functional microcircuits in visual cortex, Nature **496**, 96 (2013).
[26] S. Sadeh, C. Clopath, and S. Rotter, Emergence of functional specificity in balanced networks with synaptic plasticity, PLOS Computational Biology **11**, 1 (2015).
[27] S. Lim, Mechanisms underlying sharpening of visual response dynamics with familiarity, eLife **8**, e44098 (2019).
[28] A. Treves, Threshold-linear formal neurons in auto-associative nets, Journal of Physics A: Mathematical and General **23**, 2631 (1990).
[29] A. Treves, Graded-response neurons and information encodings in autoassociative memories, Phys. Rev. A **42**, 2418 (1990).
[30] R. Monasson and S. Rosay, Transitions between spatial attractors in place-cell models, Phys. Rev. Lett. **115**, 098101 (2015).
[31] R. Bourboulou, G. Marti, F.-X. Michon, E. El Feghaly, M. Nouguier, D. Robbe, J. Koenig, and J. Epsztein, Dynamic control of hippocampal spatial coding resolution by local visual cues, eLife **8**, e44487 (2019).
[32] H. Touchette, The large deviation approach to statistical mechanics, Physics Reports **478**, 1 (2009).
[33] J.-P. Nadal and N. Parga, Information processing by a perceptron in an unsupervised learning task, Network: Computation in Neural Systems **4**, 295 (1993), https://doi.org/10.1088/0954-898X_4_3_004.