A geometric framework to predict structure from function in neural networks

Tirthabir Biswas[1, 2, *] and James E. Fitzgerald[1, †]

[1] *Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA 20147, USA.*
[2] *Department of Physics, Loyola University, New Orleans, LA 70118, USA.*

Neural computation in biological and artificial networks relies on nonlinear synaptic integration. The structural connectivity matrix of synaptic weights between neurons is a critical determinant of overall network function, but quantitative links between neural network structure and function are complex and subtle. For example, many networks can give rise to similar functional responses, and the same network can function differently depending on context. Whether certain patterns of synaptic connectivity are required to generate specific network-level computations is largely unknown. Here we introduce a geometric framework for identifying synaptic connections required by steady-state responses in recurrent networks of rectified-linear neurons. Assuming that the number of specified response patterns does not exceed the number of input synapses, we analytically calculate the solution space of all feedforward and recurrent connectivity matrices that can generate the specified responses from the network inputs. A generalization accounting for noise further reveals that the solution space geometry can undergo topological transitions as the allowed error increases, which could provide insight into both neuroscience and machine learning. We ultimately use this geometric characterization to derive certainty conditions guaranteeing a non-zero synapse between neurons. Our theoretical framework could thus be applied to neural activity data to make rigorous anatomical predictions that follow generally from the model architecture.

I. INTRODUCTION

Structure-function relationships are fundamental to biology [1–3]. In neural networks, the structure of synaptic connectivity critically shapes the functional responses of neurons [4, 5], and large-scale techniques for measuring neural network structure and function provide exciting opportunities for examining this link quantitatively [6–14]. The ellipsoid body in the central complex of *Drosophila* is a beautiful example where modeling showed how the structural pattern of excitatory and inhibitory connections enables a persistent representation of heading direction [15–18]. Lucid structure-function links have also been found in several other neural networks [19–21]. However, it is generally hard to predict either neural network structure or function from the other [5, 22]. For example, functionally inferred connectivity can capture neuronal response correlations without matching structural connectivity [23–26], and network simulations with structural constraints do not automatically reproduce function [27–29]. Two broad modeling difficulties hinder the establishment of robust structure-function links. First, models with too much detail are difficult to adequately constrain and analyze. Second, models with too little detail may poorly match biological mechanisms, the model mismatch problem. Here we propose a rigorous theoretical framework that attempts to balance these competing factors to predict components of network structure required for function.

Neural network function probably does not depend on the exact strength of every synapse. Indeed, multiple network connectivity structures can generate the same functional responses [30, 31], as illustrated by structural variability across individual animals [22, 32] and artificial neural networks [27, 33–35]. Such redundancy may be a general feature of emergent phenomena in physics, biology, and neuroscience [36–38]. Despite this potential variability, here we find well-constrained structure-function links by characterizing all connectivity structures that are consistent with the desired functional responses [22]. We also account for ambiguities caused by measurement noise. Our goal is not to find degenerate networks that perform equivalently in all possible scenarios. We instead seek a framework that finds connectivity required for specific functional responses, independently of whatever else the network might do.

The model mismatch problem has at least two facets. First, neurons and synapses are incredibly complex [39–42], but which complexities are needed to elucidate specific structure-function relationships is unclear [5, 43, 44]. This issue is very hard to address in full generality, and here we seek a theoretical framework that makes clear experimental predictions that can adjudicate candidate models empirically. In particular, we predict neural network structure only when it occurs in all networks generating the functional responses. This high bar precludes the analysis of biophysically detailed network models, which require numerical exploration of the connectivity space that is typically incomplete [22, 30, 45–47]. We instead focus on recurrent firing rate networks of threshold-linear neurons, which are growing in popularity because they strike an appealing balance between biological realism, computational power, and mathematical tractability [12, 15, 17, 19, 21, 27, 28, 35, 48–53].

The second facet of the model mismatch problem is hidden variables, such as missing neurons, neuromodulator levels, and physiological states [5, 54–56]. Here we

*Electronic address: biswast@janelia.hhmi.org
†Electronic address: fitzgeraldj@janelia.hhmi.org

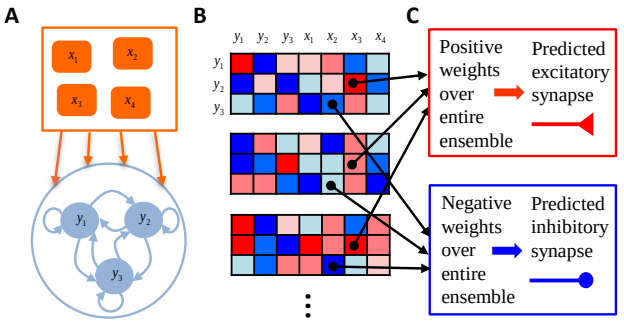

FIG. 1: **Cartoon of theoretical framework.** (**A**) We first specify some steady-state responses of a recurrent threshold-linear neural network receiving feedforward input. (**B**) We then find all synaptic weight matrices that have fixed points at the specified responses. Red (blue) matrix elements are positive (negative) synaptic weights. (**C**) When a weight is consistently positive (or consistently negative) across all possibilities, then the model needs a nonzero synaptic connection to generate the responses. We therefore make the experimental prediction that this synapse must exist. We also predict whether the synapse is excitatory or inhibitory.

take inspiration from whole-brain imaging in small organisms, such as *C. elegans* [9], larval zebrafish [8, 12, 55], and larval *Drosophila* [11], and assume access to all relevant neurons. Our model neglects neuromodulators and other state-variables, which would be interesting to consider in the future. Furthermore, many experiments indirectly assess neuronal spiking activity, such as by calcium florescence [56–59] or hemodynamic responses [23, 60–62]. We restrict our analysis to steady-state responses to mitigate mismatch between fast firing rate changes and these inherently slow measurement techniques.

Our analysis begins with an analytical characterization of synaptic weight matrices that realize specified steady-state responses as fixed points of neural network dynamics (Figs. 1A-B). A key insight is that asymmetrically constrained dimensions appear as a consequence of the threshold nonlinearity. Synaptic weight components in these semi-constrained dimensions are completely uncertain in one half of the dimension but well-constrained in the other. We then compute error surfaces by finding weight matrices with fixed points near the desired ones. This error landscape has a continuum of local and global minima, and constant-error surfaces exhibit topological transitions that add semi-constrained dimensions as the error increases. This may help explain the importance of weight initialization in machine learning, as poorly initialized models can get stuck in semi-constrained dimensions that abruptly vanish at nonzero error. By studying the geometric structure of the neural network ensemble that can approximate the functional responses, we derive analytical formulas that pinpoint a subset of connections that must exist for the model to work (Fig. 1C). These analytical results are especially useful for studying high-dimensional synaptic weight spaces that are otherwise intractable. Since the presence of a synapse is readily measurable, our theory generates accessible experimental predictions (Fig. 1C). Tests of these predictions assess the utility of the modeling framework itself, as the predictions hold across model parameters. Their successes and failures can thus move us forward towards identifying the mechanistic principles governing how neural networks implement brain computations.

II. SOLUTION SPACE GEOMETRY

Neural network structure and dynamics: Consider a neural network of $\mathcal{I}$ input neurons that send signals to a recurrently connected population of $\mathcal{D}$ driven neurons (Fig. 2A). We compactly represent the network connectivity with a matrix of synaptic weights, w_{im}, where $i = 1, \cdots, \mathcal{D}$ indexes the driven neurons, and $m = 1, \cdots, \mathcal{D}+\mathcal{I}$ indexes presynaptic neurons from both the driven and input populations. We suppose that activity in the population of driven neurons dynamically evolves according to

$$\tau_i \frac{dy_i}{dt} = -y_i + \Phi\left(\sum_{m=1}^{\mathcal{D}} w_{im} y_m + \sum_{m=\mathcal{D}+1}^{\mathcal{D}+\mathcal{I}} w_{im} x_{m-\mathcal{D}}\right), \tag{1}$$

where y_i is the firing rate of the i^{th} driven neuron, x_m is the firing rate of the m^{th} input neuron, τ_i is the time constant that determines how long the i^{th} driven neuron integrates its presynaptic signals, and $\Phi(x) = \max(0, x)$ is a threshold-linear transfer function that relates firing rates to input currents. We denote the number of synapses onto each driven neuron as $\mathcal{N}$. Note that $\mathcal{N} = \mathcal{I}+\mathcal{D}$ for a general recurrent network, $\mathcal{N} = \mathcal{I}+\mathcal{D}-1$ for recurrent networks without self-synapses, and $\mathcal{N} = \mathcal{I}$ for feedforward networks. We suppose that the network functionally maps input patterns, $x_{\mu m}$, to steady-state driven signals, $y_{\mu i} \geq 0$, where $\mu = 1, \cdots, \mathcal{P}$ labels the patterns (Fig. 2B). We assume throughout that $\mathcal{P} \leq \mathcal{N}$, as the number of known response patterns is typically small.

Parametrizing synaptic weight matrices that implement specified stimulus transformations: We aim to find features of the synaptic weight matrix that are required for this stimulus transformation. Since all time-derivatives are zero at steady-state, these response properties provide $\mathcal{D} \times \mathcal{P}$ nonlinear equations for $\mathcal{D} \times \mathcal{N}$ unknown parameters. However, each neuron's steady-state activity depends only on a single row of the connectivity matrix (Fig. 2C), so these equations separate into $\mathcal{D}$ independent sets of $\mathcal{P}$ equations for $\mathcal{N}$ unknowns. In particular,

$$y_\mu = \Phi\left(\sum_{m=1}^{\mathcal{N}} z_{\mu m} w_m\right), \tag{2}$$

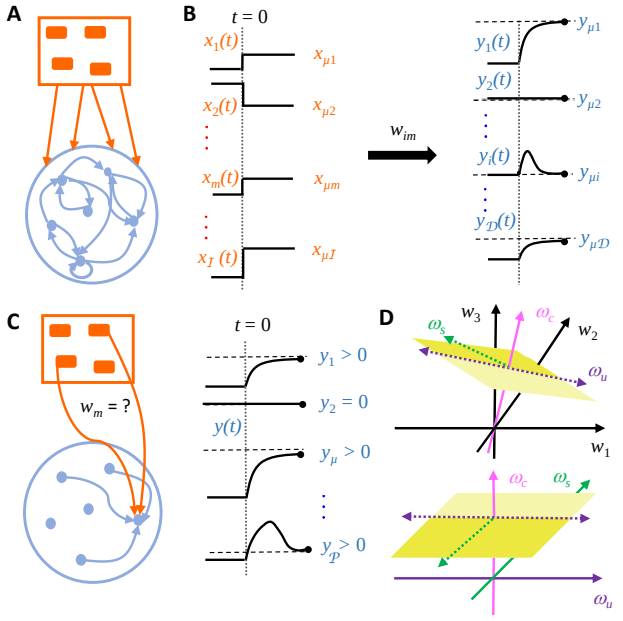

FIG. 2: **Finding network structure that implements functional responses.** (**A**) Cartoon depicting a recurrent network of *driven* neurons (blue) receiving feedforward input from a population of *input* neurons (red). (**B**) The μ^{th} pattern of input neuron activity ($x_{\mu m}$) appears at $t = 0$ and drives the recurrent neurons to approach the steady-state response pattern ($y_{\mu i}$) via feedforward and recurrent network connectivity (w_{im}). Here $m = 1, \ldots, \mathcal{I}$ indexes input neurons and $i = 1, \ldots, \mathcal{D}$ indexes driven neurons. (**C**) (*Left*) We focus on one *target* neuron at a time to determine its possible incoming synaptic weights. (*Right*) These weights must reproduce the target neuron's $\mathcal{P}$ steady-state responses from the steady-state activity patterns of all $\mathcal{N}$ presynaptic neurons. (**D**) The yellow planes depict the subspace of incoming weights that can exactly reproduce all non-zero responses of the target neuron, and the subregion shaded dark yellow indicates weights that also reproduce the target neuron's zero responses. The top graph depicts the weight space parametrized by physically meaningful w-coordinates, but the solution space is more simply parametrized by abstract ω-coordinates (bottom). The ω-coordinates depend on the specified stimulus transformation ($x_{\mu m} \to y_{\mu i}$), and ω_c, ω_s, and ω_u are coordinates in $\mathcal{C}$-dimensional constrained, $\mathcal{S}$-dimensional semi-constrained, and $\mathcal{U}$-dimensional unconstrained subspaces, respectively.

where y_μ now stands for the activity of one of the driven neuron (the target neuron) in the μ^{th} stimulus condition, z groups together the steady-state activity of all input and driven neurons, and $\vec{w}$ is the $\mathcal{N}$-vector of synaptic weights onto the target neuron. Assuming that z is full rank, we let Z be a full rank $\mathcal{N} \times \mathcal{N}$ matrix with $Z_{\mu m} = z_{\mu m}$ for $\mu = 1, \ldots, \mathcal{P}$. This implies that the last $\mathcal{N} - \mathcal{P}$ rows of Z span the null space of z, and Z defines a basis transformation on the weight space,

$$\omega_\mu = \sum_{m=1}^{\mathcal{N}} Z_{\mu m} w_m. \tag{3}$$

The $\mathcal{N}$ linearly-independent columns of Z^{-1} define the basis vectors corresponding to the ω-coordinates,

$$Z^{-1} = \left(\vec{\epsilon}_1 \ \cdots \ \vec{\epsilon}_\mu \ \cdots \ \vec{\epsilon}_{\mathcal{N}}\right), \tag{4}$$

and we can write any vector of incoming weights as

$$\vec{w} = \sum_{m=1}^{\mathcal{N}} w_m \widehat{e}_m = \sum_{\mu=1}^{\mathcal{N}} \omega_\mu \vec{\epsilon}_\mu, \tag{5}$$

where $\{\widehat{e}_m\}$ are orthonormal basis vectors with coordinates corresponding to the synaptic connection strengths between neurons. Note that $\{\widehat{e}_m\}$ is the physical basis whose coordinates correspond to the material substrates of network connectivity. The nonlinear constraint equations become

$$y_\mu = \Phi\left(\omega_\mu\right) \ \text{ for } \mu = 1, \cdots, \mathcal{P} \ . \tag{6}$$

It will be convenient to extend y_μ to an $\mathcal{N}$-vector, $\vec{y}$, by defining $y_\mu = 0$ for $\mu = \mathcal{P}+1, \cdots, \mathcal{N}$.

These ω-coordinates succinctly parametrize the solution space of all weight matrices that support the specified fixed points (Fig. 2D). Each ω-dimension can be neatly categorized into one of three types. First, for each stimulus condition μ where $y_\mu > 0$, we must have $\omega_\mu > 0$. This in turn implies that $\Phi(\omega_\mu) = \omega_\mu = y_\mu$. Because the coordinate ω_μ must adopt a specific value to generate the transformation, we say that μ defines a *constrained* dimension. We denote the number of constrained dimensions as $\mathcal{C} \leq \mathcal{P}$. Second, note that the threshold in the transfer function implies that $\Phi(x) = 0$ for all $x \leq 0$. Therefore, for any stimulus condition such that $y_\mu = 0$, we have a solution whenever $\omega_\mu \leq 0$. Because positive values of ω_μ are excluded but all negative values are equally consistent with the transformation, we say that μ defines a *semi-constrained* dimension. We denote the number of semi-constrained dimensions as $\mathcal{S} = \mathcal{P} - \mathcal{C}$. Finally, we have no constraint equations for ω_μ if $\mu = \mathcal{P}+1, \cdots, \mathcal{N}$. Because all positive or negative values of ω_μ are equally consistent with the stimulus transformation, we say that μ defines an *unconstrained* dimension. We denote the number of unconstrained dimensions as $\mathcal{U} = \mathcal{N} - \mathcal{P}$. Altogether, the stimulus transformation is consistent with every incoming weight vector that satisfies

$$\begin{aligned} \omega_\mu &= y_\mu \ \text{ if } y_\mu > 0 \\ -\infty < \omega_\mu &\leq 0 \ \text{ if } y_\mu = 0, \mu \leq \mathcal{P} \\ -\infty < \omega_\mu &< \infty \ \text{ if } \mu > \mathcal{P} \ . \end{aligned} \tag{7}$$

Note that one can enumerate the solutions in the physically meaningful w-coordinates by simply applying the inverse basis transformation to any solution found in ω-coordinates, $w = Z^{-1}\omega$. One applies this procedure to each target neuron to find full synaptic weight matrices.

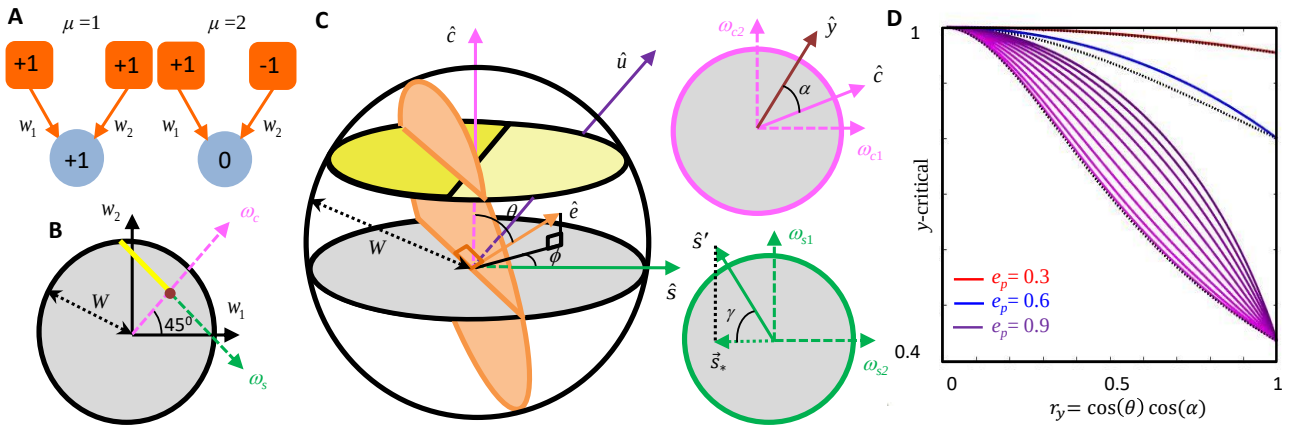

FIG. 3: **A certainty condition to determine whether neurons must be synaptically connected.** (**A**) Cartoon depicting two stimulus response patterns in a simple feedforward network with two input neurons and one driven neuron. (**B**) Since the driven neuron in (A) responds in one condition but not the other, we have one constrained dimension (magenta axis) and one semi-constrained dimension (green axis). The yellow ray depicts the space of weights, (w_1, w_2), that generate the stimulus transformation. The weight vector $(\frac{1}{2}, \frac{1}{2})$ (brown dot) would uniquely generate the neural responses in a linear network. We assume that the magnitude of the weight vector is bounded by W, such that all candidate weight vectors lie within a circle of that radius. A nonzero synapse $x_2 \to y$ exists in all solutions, but the $x_1 \to y$ synapse can be zero because the yellow ray intersects the $w_1 = 0$ axis. (**C**) Geometrically determining whether a synapse is nonzero throughout a high-dimensional solution space. (*Left*) A synapse can only vanish if the $w = 0$ hyperplane (orange circle) intersects the solution space (dark yellow wedge) within the weight bounds (bounding sphere). For example, here this intersection occurs, so the synapse is not required for the responses. For orthonormal neural responses, several parameters determine whether this intersection occurs (Appendix A). First, the magnitude of the weight bound, W, controls the extent of the solution space. Second, each synapse corresponds to a direction in synaptic weight space (orange arrow), which we represent using spherical coordinates (θ, ϕ) with respect to orthogonal axes in the constrained ($\hat{c}$, magenta), semi-constrained ($\hat{s}$, green), and unconstrained ($\hat{u}$, purple) subspaces. These angles, θ and ϕ, reflect the orientation of the hyperplane, and they are important determinants of the certainty condition. Note that the shown example would not have had an intersection if the solution space (dark yellow wedge) were moved up (along $\hat{c}$) to lie above the hyperplane (orange circle). The solution space's height is proportional to the magnitude of the postsynaptic responses, y. Thus, the solution space does not intersect the hyperplane if y exceeds a critical value, y_{cr}. (*Right top*) The angle between $\hat{c}$ and $\vec{y}$, denoted α, also matters, because it determines how much a change in y affects the height of the solution space. (*Right bottom*) The final parameter that matters is the minimal angle, γ, between the solution space and $\hat{s}' = -\text{sgn}(\cos\alpha)\hat{s}$. This angle pertains to the orientation of the hyperplane within the semi-constrained subspace. (**D**) Plots of the certainty condition, Eq. (18), for $W = 1$. The red, blue, and purple curves plot y_{cr} as a function of $r_y = e_y/e_p$ for $e_p = 0.3$, 0.6, and 0.9, respectively. Different purple shades correspond to different values of $r_{s*} = e_{s*}/\sqrt{e_p^2 - e_y^2}$. As this ratio increases, nonlinear effects increase y_{cr} and make the sign harder to determine. The red and blue curves are for the maximally nonlinear case when $r_{s*} = 1 \Rightarrow e_{s*} = \sqrt{e_p^2 - e_y^2}$. The dashed black curves represent y_{cr} in a linear model, which cannot exceed the nonlinear y_{cr}. Note that e_y, e_p, and e_{s*} are projections of the synaptic weight direction, $\hat{e}$. They quantitatively relate to correlations between pre- and postsynaptic neural activities.

III. WHEN A SYNAPSE MUST EXIST

Consistent network structure despite weight uncertainty: Although we've found infinitely many weight matrices that solve the problem, it's nevertheless possible that the solutions imply firm anatomical constraints. For illustrative purposes, we first consider the simple scenario where a driven neuron integrates feedforward input from two input neurons to selectively respond to one stimulus condition over another (Fig. 3A). This problem would have a unique solution if the network were linear, but the threshold nonlinearity introduces weight uncertainty (Fig. 3B). Nevertheless, all solutions to the problem have a synaptic connection $x_2 \to y$, and the connection is always excitatory. Positive, negative, or zero connection weights are all possible for $x_1 \to y$. However, biological synapses cannot be arbitrarily strong, and synaptic weight bounds have important implications for the solution space. For example, all solutions in Fig. 3B with $|\vec{w}| < 1$ have $w_1 > 0$, whereas larger magnitude weight vectors have $w_1 \leq 0$. Therefore, one would be certain that an excitatory $x_1 \to y$ synapse exists if prior knowledge bounded the weight vector's magnitude above by 1. Looser weight bounds raise the possibility that the synapse is absent or inhibitory. Note that too tight weight bounds, here less than $1/\sqrt{2}$, can exclude all solutions.

Finding network connectivity necessary to implement stimulus transformations: One can similarly ask when a synapse is required in high-dimensional networks. Geometrically, a synaptic connection between two neurons must exist whenever the hyperplane sepa-

rating its positive and negative weight values does not intersect the solution space within the weight bounds (Fig. 3C), because this condition guarantees that the weight is nonzero, and fixed sign, for all solutions.

Although the rigorous derivation is intricate, this certainty condition is remarkably simple for orthonormal Z (Appendix A). The sign of the synapse, when it is certain, is equal to the sign of the correlation between the presynaptic and postsynaptic activity, as intuition from linear neural network theory would suggest. However, the true synapse sign can violate this linear intuition, because the more relevant correlation is between presynaptic activity and postsynaptic drive (*i.e* the response that would have been evoked in the absence of a threshold), and the postsynaptic neuron's threshold masks suppressive inputs that could be useful for determining the weight. Accordingly, semi-constrained dimensions introduce sign ambiguity when large postsynaptic suppression, if left unmasked, would flip the sign of the activity correlation. Note that since the postsynaptic drive is negative for zero responses, this sign flip can only occur if the activity correlation and presynaptic activity have the same sign. Thus, only a subset of semi-constrained stimulus conditions contribute to sign ambiguity. In contrast, unconstrained dimensions consistently contribute to sign ambiguity, because they essentially encode presynaptic activity patterns for which the postsynaptic drive was not measured. The postsynaptic drive is effectively masked for both positive and negative values.

Quantitatively, orthonormal Z imply that only a few parameters matter for the certainty condition (Appendix A). For any given synapse, the physical basis vector is a sum of components in the constrained, semi-constrained, and unconstrained subspaces,

$$\widehat{e}_m = \sum_{\mu=1}^{\mathcal{N}} Z_{\mu m}\vec{\epsilon}_\mu = \vec{c} + \vec{s} + \vec{u}, \tag{8}$$

where we've noted that $Z_{\mu m}$ is the μ^{th} ω-coordinate of $\widehat{e}_m$, and $\vec{c}$, $\vec{s}$, and $\vec{u}$ denote the partial sums over μ in the constrained, semi-constrained, and unconstrained subspaces, respectively. Note that $\{\vec{\epsilon}_\mu\}$ are orthogonal unit vectors if and only if Z is an orthogonal matrix. In this case, the decomposition of $\widehat{e} = \widehat{e}_m$ is a sum of three orthogonal vectors that are parameterized by two angles,

$$\widehat{e} = \cos\theta\ \widehat{c} + \sin\theta\cos\phi\ \widehat{s} + \sin\theta\sin\phi\ \widehat{u}, \tag{9}$$

where $\widehat{c}$, $\widehat{s}$, and $\widehat{u}$ are unit vectors in the constrained, semi-constrained, and unconstrained subspaces, and (θ, ϕ) are spherical coordinates specifying the orientation of $\widehat{e}$ with respect to these subspaces (Fig. 3C, left). In particular,

$$\begin{aligned}
\cos\theta &= \sqrt{\sum_{\{\mu|y_\mu>0\}}^{\mathcal{P}} Z_{\mu m}^2}, \\
\sin\theta\cos\phi &= \sqrt{\sum_{\{\mu|y_\mu=0\}}^{\mathcal{P}} Z_{\mu m}^2}, \\
\sin\theta\sin\phi &= \sqrt{\sum_{\mu=\mathcal{P}+1}^{\mathcal{N}} Z_{\mu m}^2}.
\end{aligned} \tag{10}$$

These two orientation angles heavily influence whether the synapse is certain.

The certainty condition also depends on a few parameters of the $w = w_m = 0$ hyperplane that divides the positive and negative synaptic regions in the solution space (Appendix A). First, the hyperplane equation within the solution space depends on

$$\vec{y}\cdot\widehat{c} = y\cos\alpha, \tag{11}$$

where y is the length of $\vec{y}$ and α is the angle between $\vec{y}$ and $\widehat{c}$ (Fig. 3C, top right). Second, the hyperplane's normal vector can be conveniently parameterized as $\widehat{s}' = -\mathrm{Sgn}(\cos\alpha)\widehat{s}$ within the semi-constrained subspace, and whether the hyperplane intersects with the solution space depends on the size of

$$\cos\gamma = \sqrt{\sum_{\{\mu|\widehat{s}'_\mu<0\}} \widehat{s}_\mu'^2}, \tag{12}$$

where γ corresponds to the minimal angle between $\widehat{s}'$ and the solution space (Fig. 3C, bottom right).

Putting these pieces together, the synapse must be present, and its sign is unambiguous, if and only if y exceeds a critical value

$$y_{\mathtt{cr}} = W\sqrt{\frac{\cos^2\gamma\sin^2\theta\cos^2\phi + \sin^2\theta\sin^2\phi}{\cos^2\alpha\cos^2\theta + \cos^2\gamma\sin^2\theta\cos^2\phi + \sin^2\theta\sin^2\phi}} \tag{13}$$

(Appendix A). We say that the synapse is identifiable if $y > y_{\mathtt{cr}}$. Intuitively, W bounds the magnitude of weight vectors (Fig. 3C), and large W increase $y_{\mathtt{cr}}$ by admitting more solutions. Note that $y > y_{\mathtt{cr}} = W\Omega(\theta, \phi, \alpha, \gamma)$ also means that a synapse is identifiable, for a given y, when the weight bound is less than a critical value, $W < W_{\mathtt{cr}} = y/\Omega(\theta, \phi, \alpha, \gamma)$. Finally, we note that we must have $W \geq y$ for any solutions to exist.

The geometric description of Eq. 13 can be written more intuitively as

$$y_{\mathtt{cr}} = W\sqrt{\frac{e_{s*}^2 + e_u^2}{e_y^2 + e_{s*}^2 + e_u^2}} = W\sqrt{\frac{1}{1 + e_y^2/(e_{s*}^2 + e_u^2)}} \tag{14}$$

(Appendix A), where $\widehat{s}_*$ is the unit vector in the solution space that is most aligned with $\widehat{s}'$ (Fig. 3C, bottom

right), and e_y, e_{s*}, and e_u are the projections of $\widehat{e}$ onto $\widehat{y} = \vec{y}/y$, $\widehat{s}_*$, and $\widehat{u}$. Each of these projections is interpretable. Most simply,

$$e_y = \widehat{e} \cdot \widehat{y} = \frac{\sum_{\mu=1}^{\mathcal{P}} y_\mu z_{\mu m}}{\sqrt{\sum_{\nu=1}^{\mathcal{P}} y_\nu^2}} \tag{15}$$

is a normalized correlation of the pre- and postsynaptic activity (note that $\sum_{\rho=1}^{\mathcal{N}} Z_{\rho m}^2 = 1$). As expected, synapse identifiability is aided by large magnitudes of e_y. Moreover, the sign of an identifiable synapse is the sign of e_y. Identifiability is hindered by large values of

$$e_u = \widehat{e} \cdot \widehat{u} = \sqrt{1 - \sum_{\mu=1}^{\mathcal{P}} z_{\mu m}^2}, \tag{16}$$

which effectively measures the weakness of the presynaptic neuron's activity, as it is the amount of presynaptic drive for which we do not have any information on the target neuron's response. The more subtle quantity is

$$e_{s*} = \widehat{e} \cdot \widehat{s}_* = -\operatorname{Sgn}(\cos\alpha) \sqrt{\sum_{\{\mu | s'_\mu < 0\}}^{\mathcal{P}} z_{\mu m}^2},$$
$$\implies e_{s*}^2 = \sum_{\{\mu | s'_\mu < 0\}}^{\mathcal{P}} z_{\mu m}^2. \tag{17}$$

The condition that $s'_\mu < 0$ selects for patterns where the sign of the presynaptic activity is $\operatorname{Sgn}(\cos\alpha) = \operatorname{Sgn}(e_y)$, but the postsynaptic neuron does not respond. In other words, presynaptic activity should have promoted a response in the target neuron according to the observed activity correlation. That it doesn't generates uncertainty in the sign of the synapse. See Appendix A for a heuristic derivation of y_{cr} based on this argument.

Since the parameters e_y, e_{s*}, and e_u cannot be set independently, it is convenient to reparameterize Eq. 14 as

$$y_{\mathrm{cr}} = W \sqrt{\frac{1}{1 + r_y^2 e_p^2 / \left(1 - e_p^2\left(1 - r_{s*}^2\left(1 - r_y^2\right)\right)\right)}}, \tag{18}$$

where $e_p^2 = 1 - e_u^2$, $r_y^2 = e_y^2/e_p^2$, $r_{s*}^2 = e_{s*}^2/(e_p^2 - e_y^2)$, and all three composite parameters can be independently set between 0 and 1. Conceptually, r_y and r_{s*} merely normalize e_y and e_{s*} by their maximal values, and e_p is the projection of $\widehat{e}$ into the activity-constrained subspace spanned by both constrained and semi-constrained dimensions. As expected, y_{cr} is a decreasing function of r_y^2 and e_p^2 and an increasing function of r_{s*}^2 (Fig. 3D).

Numerical illustration of the certainty condition: To illustrate and verify the theory numerically, we first simulated a small neural network of three input neurons and three driven neurons across many weights (Fig. 4A).

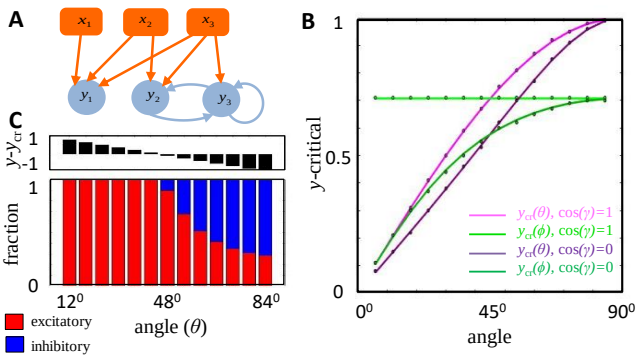

FIG. 4: **Testing the certainty condition with exhaustive low-dimensional simulations.** (**A**) A simple recurrent network with three input neurons and three driven neurons (Appendix D). (**B**) We plot the theoretically derived y_{cr} for feedforward synapses to y_1 as we vary θ (green curves) or ϕ (magenta curves), keeping the other angle fixed at 45°. The lighter shades correspond to $\cos\gamma = 1 \Rightarrow r_{s*} = 1$. The darker shades correspond to $\cos\gamma = 0 \Rightarrow r_{s*} = 0$, where the predictions from the nonlinear network match those of a linear network. The dots represent y_{cr} estimated through simulations, and they agree well with the theory. (**C**) (*Bottom*) Bar graph of the fraction of solutions with positive (red) and negative (blue) self-couplings ($y_3 \to y_3$) as a function of θ. (*Top*) As predicted, all solutions have positive w_{y_3,y_3} when $y - y_{\mathrm{cr}} > 0$.

We supposed that each driven neuron has three inputs, and we constrained weights with two orthonormal stimulus responses. Appendix D contains complete simulation details for Fig. 4. In brief, we set $W = 1$ for all simulations and randomly screened weights to comprehensively characterize the connectivity parameters that can implement a two-parameter family of responses,

$$x_{\mu m} = \begin{pmatrix} \cos\chi & \sin\psi\sin\chi & \cos\psi\sin\chi \\ 0 & \cos\psi & -\sin\psi \end{pmatrix},$$
$$y_{\mu m} = \begin{pmatrix} 0 & \sin\psi\sin\chi & \cos\chi \\ y_1 & \cos\psi & 0 \end{pmatrix}. \tag{19}$$

Appendix D enumerates the functional relationships between the (ψ, χ) variables and the (θ, ϕ, γ) angles.

The first driven neuron, y_1, receives only feedforward drive and responds to one stimulus condition. Its synapses thus have one constrained, one semi-constrained, and one unconstrained dimension. We confirmed that stronger responses were needed to make synapses fixed sign when the synaptic direction, $\widehat{e}$, was less aligned with the constrained dimension, $\widehat{c}$ (Fig. 4B, purple). Furthermore, smaller y-critical values occurred when the synaptic direction anti-aligned with the semi-constrained dimension, $\widehat{s}'$ (Fig. 4B, dark green). All simulation results (Fig. 4B, dots) were consistent with the theoretical certainty condition (Fig. 4B, curves).

Although y_2 has both feedforward and recurrent input, we can analyze its connectivity in exactly the same way as y_1. Recurrence only complicates the analysis for neurons

that synapse onto themselves, like y_3. Here changing the output activity also changes the input drive, so y and $y_{\mathtt{cr}}$ are not independent. We supposed that y_3 responded to one condition, evaluated y and $y_{\mathtt{cr}}$ (for w_{y_3,y_3}) as functions of θ, and confirmed the theory's prediction that the sign of w_{y_3,y_3} is unambiguous when $y - y_{\mathtt{cr}} > 0$ (Fig. 4C).

IV. ACCOUNTING FOR NOISE

Predicting connectivity in the presence of noise: So far we have only considered exact solutions to the fixed point equations. However, it's also important to determine weights that lead to fixed points near to the specified ones. For example, biological variability and measurement noise generally make it infeasible to specify exact biological responses. Furthermore, numerical optimization typically produces model networks that only approximate the specified computation. We therefore define the $\mathcal{E}$-*error surface* as those weights that generate fixed points a distance $\mathcal{E}$ from the specified ones,

$$\mathcal{V}_{\mathcal{E}} = \left\{ w \Big| \sum_{\mu=1}^{\mathcal{P}} \sum_{i=1}^{\mathcal{D}} \left(y_{\mu i} - \widetilde{y}_{\mu i}(w)\right)^2 = \mathcal{E}^2 \right\}, \qquad (20)$$

where $y_{\mu i}$ is the activity of the i^{th} driven neuron in the μ^{th} fixed point, and $\widetilde{y}_{\mu i}$ is the corresponding activity level in the fixed point approached by the model network when it's initialized as $y_i(t=0) = y_{\mu i}$. If the network dynamics do not approach a fixed point, perhaps oscillating or diverging instead [52], we say $\mathcal{E} = \infty$.

Each $\mathcal{E}$-error surface can be found exactly for feedforward networks. For illustrative purposes, assume that the target neuron responses satisfy $0 < y_1 < y_2 < \cdots < y_{\mathcal{P}}$. Then $\mathcal{V}_0$ is an $\mathcal{U}$-dimensional linear subspace, and $\mathcal{V}_0$ is a point in the $\mathcal{P}$-dimensional activity-constrained subspace. For $0 < \mathcal{E} < y_1$, we must have $\widetilde{y}_\mu > 0$ for all μ. Therefore, the nonlinearity is irrelevant, and $\mathcal{E}$-error surfaces are spherical in the activity-constrained ω-coordinates (Fig. 5Ai). However, once $\mathcal{E} = y_1$ it becomes possible that $\widetilde{y}_1 = 0$, and suddenly a semi-infinite line of solutions appears with $\omega_1 \leq 0$. As $\mathcal{E}$ further increases, this line dilates to a high-dimensional cylinder (Fig. 5Aii). A similar transition happens at $\mathcal{E} = y_2$, whereafter two cylinders cap the sphere (Fig. 5Aiii). Things get more interesting as $\mathcal{E}$ increases further because two transitions are possible. A third cylinder appears at $\mathcal{E}' = y_3$. However, at $\mathcal{E}'' = \sqrt{(y_1)^2 + (y_2)^2}$ it's possible for both $\widetilde{y}_1$ and $\widetilde{y}_2$ to be zero, and the two cylindrical axes merge into a semi-infinite hyperplane defined by $\omega_1 \leq 0, \omega_2 \leq 0$. Thus, when $\mathcal{E}' < \mathcal{E}''$ the error surface grows to attach a third cylinder (Fig. 5Aiv), and when $\mathcal{E}'' < \mathcal{E}'$ the error surface grows to attach a generalized cylinder with two cylindrical axes (Fig. 5Av). These topological transitions continue by adding new cylinders and merging existing ones, and the sequence is easily calculable from $\{y_\mu\}$. This geometry only approximates

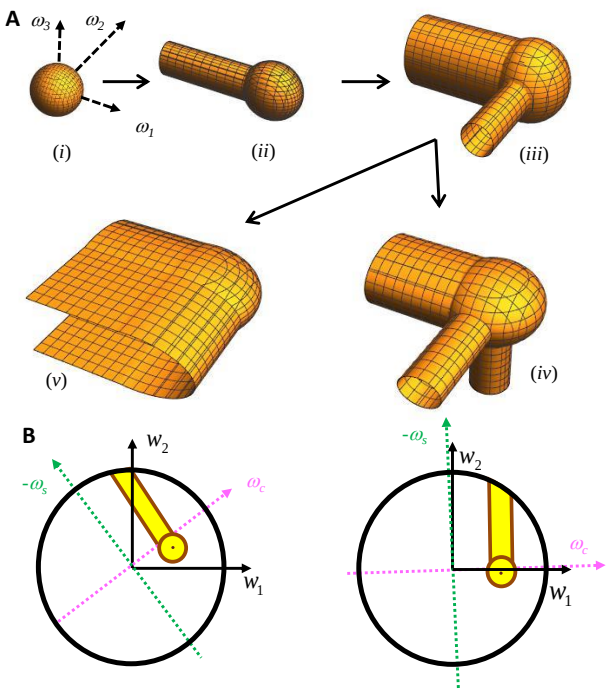

FIG. 5: **The solution space geometry changes as the allowed error increases.** (**A**) Error surface contours in a three-dimensional subspace corresponding to ω_1, ω_2, and ω_3. Several topological transitions occur as the error increases. (*i*) We consider the case where all responses are positive, so the contours are spherical for small errors, just like in a linear neural network. (*ii*)-(*iii*) Two cylindrical dimensions sequentially open up when the error is large enough for some ω-coordinates to become negative. (*iv*)-(*v*) After that, either a third cylindrical dimension can open up, or the two cylindrical axes can join to form a plane. Which transition occurs at lower error depends on the pattern of neural responses. (**B**) (*Left*) We illustrate a case where there is a unique exact solution to the problem (brown dot). Allowing error but neglecting topological transitions would expand the solution space to an ellipse (here, brown circle), but the signs of w_1 and w_2 remains positive. Including topological transitions in the error surface can cap the ellipse with a cylinder (full yellow solution space). Now we can say with certainty that the sign of w_2 is positive, but negative values of w_1 become possible. (*Right*) Graphical conventions are the same. However, in this case all solutions inside the cylinder have $w_2 > 0$. Therefore the topological transition breaks a near symmetry between positive and negative weights.

$\mathcal{E}$-error surfaces for recurrent networks, but it will nevertheless prove useful.

The threshold nonlinearity and error-induced topological transitions can have a major impact on synapse identifiability (Fig. 5B). For example, one might model a neuronal dataset with a linear neural network and find that models with acceptably low error consistently have positive signs for some synapses. However, if measured neural activity was sometimes comparable to the noise level, then semi-constained dimensions could open up

that suddenly make some of these synapse signs ambiguous (Fig. 5B, left). Although semi-constrained dimensions can never make an ambiguous synapse fully unambiguous, semi-constrained dimensions can heavily affect the distribution of synapse signs across the model ensemble by providing a large number of solutions that have consistent anatomical features (Fig. 5B, right).

We therefore generalized the certainty condition to include the effects of error, including topological transitions in the error surface (Appendix C). We usually focus on a theoretical upper bound for y-critical, $y_{\mathtt{cr},\max}$, that determines when a synapse is nonzero across all models with $\mathcal{E} \leq \epsilon$. Note that this upper bound suffices for making rigorous predictions for certain synapses, because $y > y_{\mathtt{cr},\max} \implies y > y$-critical. In the absence of topological transitions, this formula is

$$y_{\mathtt{cr},\max} = W\left[\sqrt{\left(\frac{e_{s*}^2+e_u^2}{e_y^2+e_{s*}^2+e_u^2}\right)+\frac{\epsilon^2}{W^2}\left(1+\frac{e_y^2(e_p^2-e_y^2)}{(e_y^2+e_{s*}^2+e_u^2)^2}\right)}+\frac{\epsilon}{W}\sqrt{1+\frac{e_y^2(e_p^2-e_y^2)}{(e_y^2+e_{s*}^2+e_u^2)^2}}\right]. \tag{21}$$

The effect of topological transitions is that $y_{\mathtt{cr},\max}$ becomes the maximum of several terms, each corresponding to a way that constrained dimensions could behave as semi-constrained within the error bound (Appendix C). We compute each term from a variant of Eq. 21 that reduces ϵ by the amount of error needed to open up the semi-constrained dimensions. We also computed a lower bound, $y_{\mathtt{cr},\min}$, to assess the tightness of the upper bound. This bound is

$$y_{\mathtt{cr},\min} = W\left[\sqrt{\frac{e_{s*}^2+e_u^2}{e_y^2+e_{s*}^2+e_u^2}}+\frac{\epsilon}{W}\right] \tag{22}$$

without topological transitions, and we include topological transitions by again computing the maximum of several terms. Both bounds increase with error and should be considered to be bounded above by W. As expected, both expressions reduce to Eq. (14) as $\epsilon/W \to 0$. We also note that the two bounds coincide, to leading order in ϵ/W, if $e_y \ll \max(e_{s*}, e_u)$ and $e_p/\max(e_{s*}, e_u) = \mathcal{O}(1)$, and we argue in Appendix B that this is typical when the network size is large.

Testing the theory with neural network simulations: To examine our theory's applicability, we assessed its predictions with numerical simulations of feedforward and recurrent networks (Fig. 6A). Each assessment used gradient descent learning to find neural networks whose late time activity approximated some specified orthogonal configuration of input neuron activity and driven neuron activity (Appendix E). We then used our analytically-derived certainty conditions with noise to identify a subset of synapses that were predicted to not vary in sign across the model ensemble (Appendix C), and we checked these predictions using the numerical ensemble. We similarly checked predictions from simpler certainty conditions that ignored the nonlinearity or neglected topological transitions in the error surface.

We first considered feedforward network architectures, for which our analytical treatment of noise is exact. To illustrate how non-linearity and noise affect synapse identifiability, we calculated the magnitude of postsynaptic activity needed to make a particular synapse sign certain (Fig. 6B). We specifically considered 101 random input-output configurations of a small feedforward network with 6 input neurons, which were tailored to generate one topological error surface transition at small errors (Appendix E). We then systematically varied the magnitude of target neuron activity, y, finding 10^5 synaptic weight matrices with moderate error ($\mathcal{E} \approx 0.1$) for each magnitude y. As predicted, the maximum value of y that produced numerical solutions with mixed synapse signs (Fig. 6B, black dots) was always below the theoretical upper bound for y-critical (Fig. 6B, black line). In contrast, mixed-sign numerical ensembles were often found above theoretical y-critical values that neglected topological transitions in the error surface (Fig. 6B, yellow line) or that neglected the nonlinearity entirely (Fig. 6B, cyan line). Therefore, accurately assessing synapse identifiability generally required us to include both the nonlinearity and noise-induced topological transitions in the error surface.

We next considered synapse identifiability in larger networks. For this purpose, we generated 25 random input-output configurations in a feed-forward setting (Appendix E), but this time we increased the number of input neurons from 4 to 100 across the configurations (Fig. 6C). As we increased the size of the network, we kept $\mathcal{P}/\mathcal{N}$ fixed at 1 (Fig. 6C, brown) or 0.5 (Fig. 6C, purple), kept $\mathcal{C}/\mathcal{N}$ fixed at 0.25, and numerically found solutions with $\mathcal{E}^2/\mathcal{P} \approx 10^{-6}$. The number of certain synapses predicted by the theory (Fig. 6C, light lines) empirically scaled with the network size linearly (Fig. 6C, dark lines), and our simulations validated all theoretical predictions (Fig. 6C, circles). Interestingly, a simple heuristic argument predicts the linear scaling of Fig. 6C (Appendix B). These results suggest that the theory will predict many synapses to be certain in realistically large neural systems.

Finally, we empirically tested our theory for a recurrent network (Figs. 6D, 6E), where our treatment of noise is only approximate. Here we constructed a single configuration with non-negative driven neuron responses and

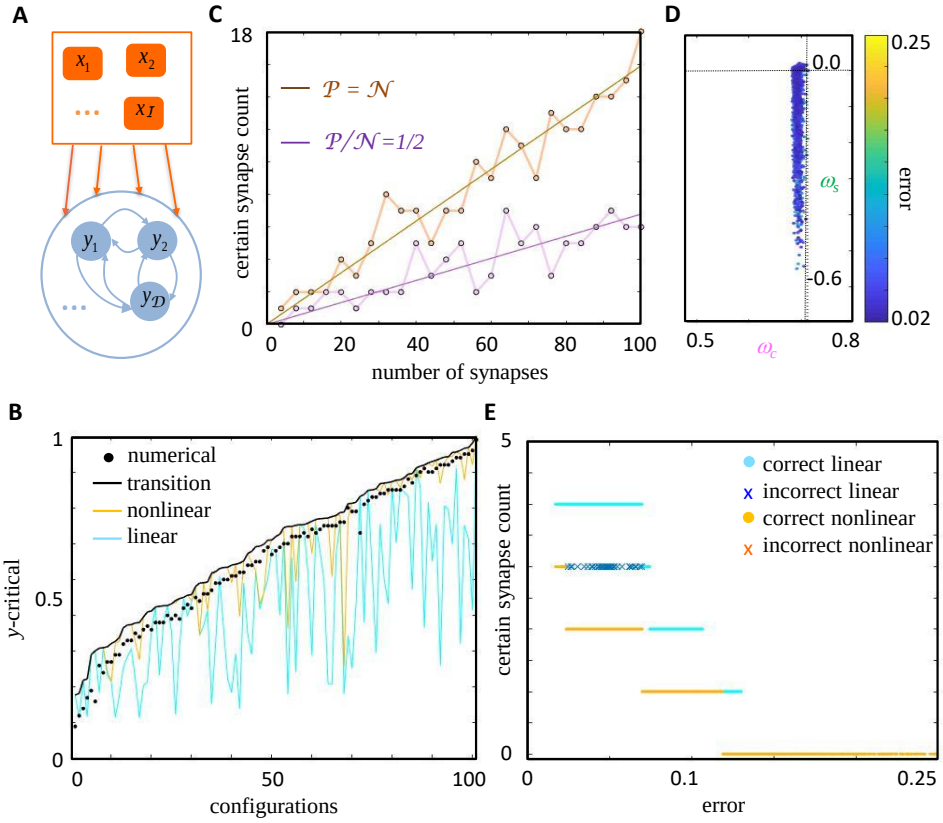

FIG. 6: **The theory accounting for error explains numerical ensembles of feedforward and recurrent networks.** (**A**) Cartoon of a recurrent neural network. We disallow recurrent connectivity of neurons onto themselves throughout this figure. $\mathcal{D} = 1$ corresponds to the feedforward case. (**B**) Comparison of numerical and theoretical y-critical values for 101 random configurations of input-output activity (Appendix E). We considered a feedforward network with $\mathcal{I} = 6$, $\mathcal{P} = 5$, $\mathcal{C} = 2$, and $W = 1$. For each configuration and postsynaptic activity level y, we used gradient descent learning to numerically find many solutions to the problem with $\mathcal{E} \approx 0.1$. The black dots correspond to the maximal value of y in our simulations that resulted in an inconsistent sign for the synaptic weight under consideration. The continuous curves show theoretical values for y-critical that upper bound the true y-critical ($y_{\mathrm{cr,max}}$, black), that neglect topological transitions in the error surface (yellow), or that neglect the threshold nonlinearity (cyan). Only the black curves successfully upper bounded the numerical points. Configurations were sorted by the $y_{\mathrm{cr,max}}$ value predicted by the black curve. (**C**) The number of certain synapses increased with the total number of synapses in feedforward networks. Purple and brown correspond to $\mathcal{N} = 2\mathcal{P} = 4\mathcal{C}$ and $\mathcal{N} = \mathcal{P} = 4\mathcal{C}$, respectively. The thin lines plot the predicted number of certain synapses. The circles represent the number of correctly predicted synapse signs in the simulations. The dark brown and purple lines are best-fit linear curves. (**D**)-(**E**) Testing the theory in a recurrent neural network with $\mathcal{N} = 10, \mathcal{I} = 7, \mathcal{D} = 4, \mathcal{P} = 8$ and $\mathcal{C} = 3$. Each dot shows a model found with gradient descent learning. (D) x- and y-axes show two ω-coordinates predicted to be constrained and semi-constrained, respectively, and the color axis shows the model's root mean square error over neurons, $\mathcal{E}/\sqrt{\mathcal{D}}$. Although our theory for error surfaces is approximate for recurrent networks, the solution space was well explained by the constrained and semi-constrained dimensions. Note that the numerical solutions tend to have constrained coordinates smaller than the theoretical value (vertical line) because the learning procedure is initialized with small weights and stops at nonzero error. (E) The x-axis shows the model's error, and the y-axis shows the number of synapse signs correctly predicted by the nonlinear theory (yellow dots or red crosses) or linear theory (cyan dots or blue crosses). Dots denote models for which every model prediction was accurate, and crosses denote models for which some predictions failed.

orthogonal presynaptic patterns for the target neuron (Appendix E). Driven neurons were not allowed to make synapses onto themselves. We then used gradient descent learning to find around 4500 networks that approximated the desired fixed points with variable accuracy. For technical simplicity, we first found connectivity matrices using a proxy cost function that treated the network as if it were feedforward. We then simulated the neural network dynamics with these weights and correctly evaluated the model's error as prescribed by Eq. (20). This network ensemble revealed that constrained and semi-constrained dimensions accurately explained the structure of the solution space for non-zero errors as well (Fig. 6D). Moreover, the theory correctly predicted how the number of certain

synapses would decrease as a function of $\mathcal{E}$ (Fig. 6E, yellow circles), and we never found a numerical violation of the theoretical certainty condition that included nonlinearity and noise (Fig. 6E, no red crosses). Here accurate predictions did not require us to account for topological error surface transitions. In contrast, although our simulations usually agreed with the predictions of the linear theory (Fig. 6E, cyan circles), they could also disagree (Fig. 6E, blue crosses). More simulations and theory are needed to fully understand when the certainty conditions with error apply to recurrent networks, but the current simulations are clearly encouraging.

V. DISCUSSION

In summary, we enumerated all threshold-linear recurrent neural networks that generate specified sets of fixed points, under the assumption that the number of candidate synapses onto a neuron exceeds the number of fixed points. We found that the geometry of the solution space was elegantly simple, and we described a coordinate transformation that permits easy classification of weight-space dimensions into constrained, semi-constrained, and unconstrained varieties. This geometric approach also generalized to approximate error-surfaces of model parameters that imprecisely generate the fixed points. We used this geometric description of the error surface to analyze structure-function links in neural networks. In particular, we found that it is often possible to identify synapses that must be present for the network to perform its task, and we verified the theory with simulations of feedforward and recurrent neural networks.

Rectified-linear units are also popular in state of the art machine learning models [27, 63–65], so the fundamental insights we provide into the effects of neuronal thresholds on neural network error landscapes may have practical significance. For example, machine learning often works by taking a model that initially has high error and gradually improving it by modifying its parameters in the gradient-direction [66]. However, error surfaces with high error can have semi-constrained dimensions that abruptly vanish at lower errors (Fig. 5). Local parameter changes typically cannot move the model through these topological transitions, because models that wander deeply into semi-constrained dimensions are far from where they must be to move down the error surface. The model has continua of local and global minima, and the network needs to be initialized correctly to reach its lowest possible errors. This could provide insight into deep learning theories that view its success as a consequence of weight subspaces that happen to be initialized well [67, 68].

Our theory could be extended in several important ways. First, we only analyzed the identifiability of critical synapses from orthonormal sets of fixed points. A more general analysis will be needed to pinpoint synapses in realistic biological settings, as activity patterns elicited under different stimulus conditions are typically correlated. Since our error surface description made no orthonormality assumptions, this analysis will only require more complicated geometrical calculations to discern whether the synapse sign is consistent across the space of low-error models. Furthermore, we could use the error surfaces to identify multi-synapse anatomical motifs that are required for function, or to estimate the fraction of models in which an uncertain synapse is excitatory versus inhibitory. It will also be interesting to relax the assumption that the number of fixed points is small. This would allow us to consider scenarios where the fixed points can only be generated nonlinearly. We could also consider cases where no exact solution exists at all. Here we assumed that we knew the activity level of every neuron in the circuit. This is not always the case, and it will be important to determine how unobserved neurons alter the error landscape for synaptic weights connecting the observed neurons. The error landscape geometry will also be affected by recurrent network effects that we ignored here (Appendix C). It will be interesting to see whether the geometric toolbox of theoretical physics can provide insights into the nontrivial effects of unobserved neurons and recurrent network dynamics. Finally, we note that our analysis could be trivially extended to nonlinear recurrent networks of capped rectified linear units [64], which saturate above a second threshold. In particular, semi-constrained dimensions would emerge from any condition where the target neuron is inactive or saturated.

The geometric simplicity of the zero-error solution space provides several insights into neural network computation. Every time a neuron has a vanishing response, half of a dimension remains part of the solution space, which the network could explore to perform other tasks. In other words, by replacing an equality constraint with an inequality constraint, simple thresholding nonlinearities effectively increase the computational capacity of the network [69, 70]. The flexibility afforded by vanishing neuronal responses thereby provides an intuitive way to understand the impressive computational power of sparse neural representations [48, 71–73]. Furthermore, the brain could potentially use this flexibility to set some synaptic strengths to zero, thereby improving wiring efficiency. This would link sparse connectivity to sparse response patterns, both of which are observed ubiquitously in neural systems.

Our primary motivation for undertaking this study was to find rigorous theoretical methods for predicting neural circuit structure from its functional responses. This identification can be used to corroborate or broaden circuit models that posit specific connectivity patterns, such as center-surround excitation-inhibition in ring attractors [15–17] or contralateral relay neuron connectivity in zebrafish binocular vision [12, 74]. More generally, if an experimental test violates the certainty conditions we derived using our ensemble modeling approach, it will suggest that some aspect of model mismatch is important. We could then move on to the development of qualitatively improved models that might modify neuronal non-

linearities, relax weight bounds, incorporate sub-cellular processes or neuromodulation, or hypothesize hidden cell populations. On the other hand, we hope that our focus on predictions that follow with certainty from simple network assumptions will enable predictions that are relatively insensitive to minor mismatches between our abstract model and the real biological brain. More nuanced predictions may require more nuanced models.

An exciting prospect is to explore how our ensemble modeling framework can be combined with other theoretical principles and biological constraints to obtain more refined structure-function links. For instance, we could refine our ensemble by restricting to stable fixed points. Alternatively, once the sign of a given synapse is identified, Dale's principle might allow us to fix the signs of all other synapses from this neuron [75]. This would restrict the solution space and could make other synapses certain. Utilizing limited connectomic data to impose similar restrictions might also be a fruitful way to benefit from large-scale anatomical efforts [7, 10, 13, 14]. Finally, rather than restricting the magnitude of the incoming synaptic weight vector, we could consider alternate biologically relevant constraints, such as limiting the number of synapses, minimizing the total wiring length, or positing that the network operates at capacity [76, 77]. These changes would modify the certainty conditions in our framework, as well as our experimental predictions. We could therefore assess candidate optimization principles and biological priors experimentally. While the base framework developed here was designed to identify crucial network connections required for function, we hope that our approach will eventually allow us to assess theoretical principles that determine how neural network structure follows from function.

ACKNOWLEDGMENTS

The authors thank Tianzhi (Lambus) Li, Srini Turaga, Andrew Saxe, Ran Darshan, and Larry Abbott for helpful discussions and comments on the manuscript. This work was supported by the Howard Hughes Medical Institute and Janelia's Visiting Scientist Program.

APPENDIX

A. A Certainty Condition to Pinpoint Synapses Required for Specified Response Patterns

Preliminaries

For completeness, we begin by briefly reviewing a few central concepts from the main manuscript.

From recurrent to feedforward networks: Let us consider a neural network of $\mathcal{I}$ input neurons that send signals to an interconnected population of $\mathcal{D}$ driven neurons governed by dynamical equations (1), as described in the main manuscript. At steady-state, since all time-derivatives are zero, (1) yields

$$y_{\mu i} = \Phi\left(\sum_{m=1}^{\mathcal{D}} w_{im} y_{\mu m} + \sum_{m=\mathcal{D}+1}^{\mathcal{D}+\mathcal{I}} w_{im} x_{\mu,m-\mathcal{D}}\right) = \Phi\left(\sum_{m=1}^{\mathcal{N}} w_{im} z_{\mu m}\right) \ , \tag{A.1}$$

where, as prescribed in the main manuscript, $y_{\mu i}$ and $x_{\mu m}$ denote steady-state activity levels of the driven and input neurons to the μ^{th} stimulus, which we have combined into $z_{\mu m}$, and $\mathcal{N}$ is the number of incoming synapses onto each of the driven neurons. (A.1) provides $\mathcal{D} \times \mathcal{P}$ nonlinear equations for $\mathcal{D} \times \mathcal{N}$ unknown parameters. However, we immediately notice that the steady-state activity of neuron i depends only on the i^{th} row of the connectivity matrix, so these equations separate into $\mathcal{D}$ independents sets of $\mathcal{P}$ equations with $\mathcal{N}$ unknowns, the weights onto a given driven neuron. In other words, the recurrent network involving $\mathcal{D}$ driven and $\mathcal{I}$ input neurons decomposes into $\mathcal{D}$ feedforward networks with $\mathcal{N} = \mathcal{D} + \mathcal{I}$ feedforward inputs. The steady-state equations for these feedforward networks are given by,

$$y_{\mu} = \Phi\left(\sum_{m=1}^{\mathcal{N}} z_{\mu m} w_m\right) \ , \tag{A.2}$$

where we have now suppressed the i index in $y_{\mu i}$ and in w_{im}. For this feedforward network we will refer the i^{th} neuron as the target neuron, and it is as if that all the neurons (driven and input) are providing feedforward inputs to it. As long as we only consider exact solutions to the fixed point equations, the problem of identifying synaptic connectivity in a recurrent network reduces to solving the problem for feedforward networks. Thus in the rest of this appendix we will focus on identifying w_m's satisfying (A.2).

A convenient set of variables: In all our discussions in this section the input neuronal response matrix, $z_{\mu m}$, will be assumed to be fixed. Note that $z_{\mu m}$ connects synaptic weight vectors to the target response vector and can be used to

define $\mathcal{P}$ weight combinations, the ω-coordinates. Each ω-coordinate controls the target response to a single stimulus condition:

$$y_\mu = \Phi(\omega_\mu) \ , \ \text{where} \ \ \omega_\mu \equiv \sum_{m=1}^{\mathcal{N}} z_{\mu m} w_m \ . \tag{A.3}$$

It is rather convenient to extend this set of $\mathcal{P}$ ω-coordinates to a basis set of $\mathcal{N}$ ω-coordinates, such that all synaptic weights can be uniquely expressed as a linear combination of these ω-coordinates, and vice-versa. To see how this can be done, we will henceforth make the simplifying assumption that the $\mathcal{P} \times \mathcal{N}$ matrix has the maximal rank, $\mathcal{P}$, although we anticipate that much of our framework, results, and insights will apply more generally. If z has maximal rank, its kernel will be an $(\mathcal{N}-\mathcal{P})$-dimensional linear subspace spanned by $(\mathcal{N}-\mathcal{P})$ orthogonal basis vectors, denoted $\vec{\epsilon}_\mu$ for $\mu = \mathcal{P}+1 \ldots \mathcal{N}$. We can now extend z to an $\mathcal{N} \times \mathcal{N}$ matrix, Z, as follows

$$\begin{aligned} Z_{\mu m} &= z_{\mu m} \ \ \text{for} \ \ \mu = 1 \ldots \mathcal{P}, \ \text{and} \ \ \forall \ m, \\ Z_{\mu m} &= \epsilon_{\mu m} \ \ \text{for} \ \ \mu = \mathcal{P}+1 \ldots \mathcal{N}, \ \text{and} \ \ \forall \ m, \end{aligned} \tag{A.4}$$

where $\epsilon_{\mu m}$ is the m^{th} component of the null vector $\vec{\epsilon}_\mu$. With this construction, it is easy to see that the new ω-coordinates,

$$\omega_\mu \equiv \sum_{m=1}^{\mathcal{N}} Z_{\mu m} w_m \ , \ \text{for} \ \ \mu = \mathcal{P}+1, \ldots, \mathcal{N} \ , \tag{A.5}$$

remain completely unconstrained by the specified response patterns, as these linear combinations do not contribute to any of the target responses. In contrast, the original ω-coordinates,

$$\omega_\mu = \sum_{m=1}^{\mathcal{N}} Z_{\mu m} w_m = \sum_{m=1}^{\mathcal{N}} z_{\mu m} w_m \ , \ \text{for} \ \ \mu = 1, \ldots, \mathcal{P} \ , \tag{A.6}$$

are all constrained by the data:

$$\omega_\mu \begin{cases} = y_\mu & \text{for} \ \ \mu = 1, \ldots, \mathcal{C}, \text{the constrained dimensions} \\ \leq 0 & \text{for} \ \ \mu = \mathcal{C}+1, \ldots, \mathcal{P}, \text{the semi-constrained dimensions.} \end{cases} \ , \tag{A.7}$$

where for notational simplicity we have ordered the response patterns such that $y_\mu \neq 0$ for $\mu = 1, \ldots, \mathcal{C}$. Also, we extend the y_μ's to an $\mathcal{N}$-dimensional vector, $\vec{y}$, by assigning $y_\mu = 0$ for $\mu = \mathcal{P}+1 \ldots \mathcal{N}$.

The extended response matrix Z defines a basis transformation connecting synaptic directions, $\widehat{e}_m$, with directions

$$\vec{\epsilon}_\mu \equiv \sum_{m=1}^{\mathcal{N}} \widehat{e}_m Z^{-1}_{m\mu} \ , \tag{A.8}$$

along which the ω-coordinates change. These $\vec{\epsilon}$ vectors clearly differentiate directions in the weight space that are activity-constrained by neuronal responses ($\mu = 1, \ldots, \mathcal{P}$) from those that are not ($\mu = \mathcal{P}+1, \ldots, \mathcal{I}$). We can express any weight vector in either the $\{\widehat{e}_m\}$ basis or the $\{\vec{\epsilon}_\mu\}$ basis:

$$\vec{w} = \sum_{m=1}^{\mathcal{N}} w_m \widehat{e}_m = \sum_{\mu=1}^{\mathcal{N}} \omega_\mu \vec{\epsilon}_\mu \ , \ \text{where} \ w_m = \sum_{\mu=1}^{\mathcal{N}} Z^{-1}_{m\mu} \omega_\mu \ , \ \omega_\mu = \sum_{\mu=1}^{\mathcal{N}} Z_{\mu m} w_m \ , \vec{\epsilon}_\mu \equiv \sum_{m=1}^{\mathcal{N}} \widehat{e}_m Z^{-1}_{m\mu} \ , \ \ \widehat{e}_m = \sum_{\mu=1}^{\mathcal{N}} \vec{\epsilon}_\mu Z_{\mu m} \ . \tag{A.9}$$

For later convenience we also define the number of semi-constrained and unconstrained dimensions as, $\mathcal{S} = \mathcal{P} - \mathcal{C}$, and $\mathcal{U} = \mathcal{N} - \mathcal{P}$, respectively.

Derivation of the certainty condition for orthogonal inputs:

Our goal here is to use the solution space (*i.e.* ensemble of weights that are precisely able to recover the specified target responses) to derive a condition for when we can be certain that a given synapse must be nonzero. For technical simplicity, we will specialize to the case when all the response patterns are orthonormal, *i.e.*

$$\sum_{m=1}^{\mathcal{N}} z_{\mu m} z_{\nu m} = \delta_{\mu\nu} \Leftrightarrow z z^T = I \ . \tag{A.10}$$

Then, we can always choose the extended Z matrix to be an $\mathcal{N} \times \mathcal{N}$ orthogonal matrix, such that $Z^{-1} = Z^T$ and the $\vec{\epsilon}_\mu$ vectors now form an orthonormal basis. Motivated by biological constraints, we will impose a bound on the magnitude of the synaptic weight vector. For orthonormal responses, this translates into a spherical bound on ω coordinates as well (Fig. 3C)

$$|\vec{w}|^2 = \sum_{m=1}^{\mathcal{N}} w_m^2 = \sum_{\mu=1}^{\mathcal{N}} \omega_\mu^2 \leq W^2 \ . \tag{A.11}$$

We refer to this $\mathcal{N}$-dimensional ball, in which all admissable synaptic weights reside, as the weight-space.

A heuristic argument for y-critical: Before diving into the rigorous and technical derivation, in this subsection we first try to intuitively understand how the certainty condition (14) can arise. For this purpose, let us start with a linear theory with no unconstrained dimension, so $\mathcal{S} = \mathcal{U} = 0$. In this case, there is a unique set of weights that can precisely reproduce the observed responses:

$$w_m = \sum_{\mu=1}^{\mathcal{N}} Z^{-1}_{m\mu} y_\mu = \sum_{\mu=1}^{\mathcal{N}} Z_{\mu m} y_\mu \ . \tag{A.12}$$

Since $Z_{\mu m} = z_{\mu m}$ represents the responses of the m^{th} presynaptic neuron, the solution for the m^{th} synaptic weight (A.12) is simply the correlation between the pre and post synaptic activity. In a linear theory, the sign of the synapse is thus dictated by the sign of the correlation between the pre and post synaptic neuron.

Let us now allow a single ($\mathcal{N}^{th}$) unconstrained direction. One can think of this situation as if we do not have the information on how the target neuron would respond to the unconstrained stimulus pattern. If we knew that this response was say, y_u, then we would have been able to determine the sign of w_m:

$$\mathrm{Sgn}(w_m) = \mathrm{Sgn}\left(\sum_{\mu=1}^{\mathcal{N}-1} Z_{\mu m} y_\mu + Z_{\mathcal{N} m} y_u \right) \ . \tag{A.13}$$

However, since we do not know what the last term is, if it can cancel the first term for some allowed value of y_u then the overall sign becomes ambiguous. Conversely, w_m becomes identifiable if

$$\left| \sum_{\mu=1}^{\mathcal{N}-1} Z_{\mu m} y_\mu \right| > |Z_{\mathcal{N} m} y_u| \ \forall \ y_u \ . \tag{A.14}$$

Now, it is easy to recognize that the first term is just $\widehat{e} \cdot \vec{y} = y e_y$, where we have suppressed the m index on $\widehat{e}_m$ and will continue to do so while referring to the synapse direction whose sign we are considering. $e_y = \widehat{e} \cdot \widehat{y}$ refers to the projection of $\widehat{e}$ along $\widehat{y}$. Also, note that in this simple case with one unconstrained direction, the projection of $\widehat{e}$ along the unconstrained subspace is just given by $e_u = \widehat{e} \cdot \vec{\epsilon}_{\mathcal{N}} = Z_{\mathcal{N} m}$. Further, since Z is orthogonal, in order to have any solution at all

$$y^2 + y_u^2 \leq W^2 \ . \tag{A.15}$$

Substituting the maximum $|y_u|$ from (A.15) into (A.14), after some algebra we get the condition for sign identifiability as

$$y > y_{\mathtt{cr}} = W \sqrt{\frac{e_u^2}{e_u^2 + e_y^2}} \ , \tag{A.16}$$

where the subscript on $y_{\mathtt{cr}}$ abbreviates "critical."

The same argument applies if the $\mathcal{N}^{th}$ direction is semi-constrained instead of unconstrained, with one notable difference. If the $\mathcal{N}^{th}$ pattern was semi-constrained that means $y_{\mathcal{N}} = 0$, and the nonlinear thresholding is masking how the target neuron would have responded in a linear model [1]. However, the ambiguity in sign can only arise if the second term has a sign opposite to the first term, or a sign opposite to $\mathrm{Sgn}(e_y)$. Moreover, for the thresholding to

[1] Note that our relation between the sign of the synapse and the sign of the correlation is based on a linear response.

act, the target response for the semi-constrained pattern must be negative in the linear theory, so $Z_{\mathcal{N}m}$ has to have the same sign as e_y to generate the ambiguity. And, if it is indeed so, then we obtain a certainty condition that is identical to (A.16) except that $e_u \to e_s$, the projection of $\widehat{e}$ along the semi-constrained direction:

$$y > y_{\mathtt{cr}} = W\sqrt{\frac{e_s^2}{e_s^2+e_y^2}} \ . \tag{A.17}$$

If $Z_{\mathcal{N}m}$ and e_y have opposite signs, then the synapse always has the same sign throughout the solution space.

While this derivation of $y_{\mathtt{cr}}$ is heuristic and only deals with a single semi-constrained or unconstrained dimension, it provides intuition for the general result (14). Essentially, whether the sign of a given synapse is constant across the solution space depends on two competing quantities: the correlation between the pre- and postsynaptic responses; and the strength of the postsynaptic drive for patterns where the target response is either unknown or masked by the thresholding nonlinearity.

Hyperplane dividing excitatory and inhibitory synaptic regions: Having gained some intuition about the certainty condition, let us now proceed to a rigorous derivation of the result. Since the constrained coordinates are fixed for the weight vectors that belong to the solution space (the deep yellow wedge in Fig. 3C), we must have

$$y^2 \equiv \sum_{\mu=1}^{\mathcal{C}} y_\mu^2 = \sum_{\mu=1}^{\mathcal{C}} \omega_\mu^2 \ , \tag{A.18}$$

so that the solution space resides within an $(\mathcal{U}+\mathcal{S})$-dimensional ball with radius

$$\overline{W} \equiv \sqrt{W^2-y^2} \ , \tag{A.19}$$

as depicted in Fig. 3C by the light yellow region. We refer to this semi-constrained plus unconstrained subspace as the flexible subspace.

Now, the synaptic direction of interest, $\widehat{e}$, can be decomposed into its projections along constrained, semi-constrained and unconstrained subspaces. Let us define $e_\mu \equiv (\widehat{e}\cdot\vec{\epsilon}_\mu) = Z_{\mu m}$ to be the component of $\widehat{e}$ along $\vec{\epsilon}_\mu$. Note that the second equality follows from the orthogonality assumption and (A.9). In general, in this manuscript we will use subscripts on e to denote projections of $\widehat{e}$ along different directions or subspaces. We can now write

$$\widehat{e} = \sum_{\mu=1}^{\mathcal{N}} e_\mu\vec{\epsilon}_\mu = \sum_{\mu=1}^{\mathcal{C}} e_\mu\vec{\epsilon}_\mu + \sum_{\mu=\mathcal{C}+1}^{\mathcal{P}} e_\mu\vec{\epsilon}_\mu + \sum_{\mu=\mathcal{P}+1}^{\mathcal{N}} e_\mu\vec{\epsilon}_\mu = \cos\theta\widehat{c} + \sin\theta\cos\phi\widehat{s} + \sin\theta\sin\phi\widehat{u} \ , \tag{A.20}$$

where

$$\widehat{c} \equiv \frac{\sum_{\mu=1}^{\mathcal{C}} e_\mu\vec{\epsilon}_\mu}{\sqrt{\sum_{\mu=1}^{\mathcal{C}} e_\mu^2}} \ , \ \widehat{s} \equiv \frac{\sum_{\mu=\mathcal{C}+1}^{\mathcal{P}} e_\mu\vec{\epsilon}_\mu}{\sqrt{\sum_{\mu=\mathcal{C}+1}^{\mathcal{P}} e_\mu^2}} \ , \widehat{u} \equiv \frac{\sum_{\mu=\mathcal{P}+1}^{\mathcal{N}} e_\mu\vec{\epsilon}_\mu}{\sqrt{\sum_{\mu=\mathcal{P}+1}^{\mathcal{N}} e_\mu^2}} \ , \tag{A.21}$$

are unit vectors that lie within the constrained, semi-constrained and unconstrained subspaces, and

$$\begin{aligned} e_c \equiv \widehat{c}\cdot\widehat{e} = \cos\theta &= \sqrt{\sum_{\mu=1}^{\mathcal{C}} e_\mu^2} \geq 0 \ , \\ e_s \equiv \widehat{s}\cdot\widehat{e} = \sin\theta\cos\phi &= \sqrt{\sum_{\mu=\mathcal{C}+1}^{\mathcal{P}} e_\mu^2} \geq 0 \ , \\ e_u \equiv \widehat{u}\cdot\widehat{e} = \sin\theta\sin\phi &= \sqrt{\sum_{\mu=\mathcal{P}+1}^{\mathcal{N}} e_\mu^2} \geq 0 \end{aligned} \tag{A.22}$$

are the projections of $\widehat{e}$ along these directions. One could think of θ, ϕ as representing a spherical coordinate system where the role of x, y and z axes are played by $\widehat{s}, \widehat{u}$ and $\widehat{c}$ respectively, and our definitions (A.20 - A.22) imply the convention, $0 \leq \theta, \phi < \pi/2$. For later convenience, let us also introduce the projection of $\widehat{e}$ onto the activity-constrained subspace (constrained plus semi-constrained):

$$e_p \equiv \widehat{p}\cdot\widehat{e} = \sqrt{\cos^2\theta + \sin^2\theta\cos^2\phi} = \sqrt{\sum_{\mu=1}^{\mathcal{P}} e_\mu^2} \geq 0 \ , \ \text{where } \widehat{p} \equiv \frac{\sum_{\mu=1}^{\mathcal{P}} e_\mu\vec{\epsilon}_\mu}{\sqrt{\sum_{\mu=1}^{\mathcal{P}} e_\mu^2}} \ . \tag{A.23}$$

We would also like to emphasize that we can compute θ, ϕ just from the knowledge of the neuronal responses, $z_{\mu m} = e_\mu$, which is particularly useful for numerical calculations:

$$\theta \equiv \cos^{-1}\left(\sqrt{\sum_{\mu=1}^{\mathcal{C}} z_{\mu m}^2}\right) \ , \text{ and } \ \phi \equiv \cos^{-1}\left(\sqrt{\sum_{\mu=\mathcal{C}+1}^{\mathcal{P}} z_{\mu m}^2}\Big/\sqrt{1-\sum_{\mu=1}^{\mathcal{C}} z_{\mu m}^2}\right) \ . \tag{A.24}$$

Now, any weight vector in the solution space can be written as

$$\vec{w} = \vec{y} + \vec{w}_s + \vec{w}_u \ , \tag{A.25}$$

where $\vec{w}_s$ and $\vec{w}_u$ are the projections of $\vec{w}$ onto the semi-constrained and unconstrained subspaces, and the constrained part of $\vec{w}$ is fixed at $\vec{y}$. Using (A.20) and (A.25), one then finds that the $w_m = 0$ hyperplane dividing the excitatory and inhibitory regions in the flexible subspace satisfies the equation

$$w = \vec{w} \cdot \widehat{e} = y \cos\theta \cos\alpha + \sin\theta \cos\phi \ \widehat{s} \cdot \vec{w}_s + \sin\theta \sin\phi \ \widehat{u} \cdot \vec{w}_u = 0 \ , \tag{A.26}$$

where we have now also suppressed the index m in w_m. Also, we have defined, $\alpha \in [0, \pi]$, to be the angle between $\vec{y}$ and $\widehat{c}$. We now notice that the origin of the flexible subspace, $\vec{w}_s = \vec{w}_u = 0$, is in the solution space and the sign of w for this solution point is given by

$$\mathrm{Sgn}(w) = \mathrm{Sgn}(\cos\alpha) = \mathrm{Sgn}(e_y) = \mathrm{Sgn}\left(\sum_{\mu=1}^{\mathcal{C}} z_{\mu m} y_\mu\right) \ . \tag{A.27}$$

In other words, if the sign of the synapse is certain, this certain sign must be $\mathrm{Sgn}(\cos\alpha)$, which corresponds to the sign of the correlation between the target neuron and the presynaptic neuron. Intuitively, positive correlations point to an excitatory connection, and negative correlations point to an inhibitory connection.

Special case without unconstrained dimensions: To derive the certainty condition, let's start by looking at the case when $\mathcal{P} = \mathcal{N}$, so that there are no unconstrained directions, or equivalently, $\phi = 0$. In this case, the solution space is just the all-negative orthant in the $\mathcal{S}$-dimensional semi-constrained hypersphere (Fig. 7), and the equation for the $w = 0$ hyperplane can be written as

$$\sin\theta \ (\widehat{s}\,' \cdot \vec{w}_s) = y \cos\theta |\cos\alpha| \ , \tag{A.28}$$

where the right hand side is positive, and we have introduced

$$\widehat{s}\,' \equiv -\mathrm{Sgn}(\cos\alpha)\widehat{s} \ , \tag{A.29}$$

which flips the direction of $\widehat{s}$ if $\cos\alpha > 0$, or equivalently if $e_y > 0$. Now, if the $w = 0$ hyperplane (orange lines in Fig. 7) is far enough along $\widehat{s}\,'$ from the origin that it does not intersect with the all-negative orthant within the weight bounds, then we can be certain that w is nonzero and always has a consistent sign. To check this, we need to compare the cone angle that the orange hyperplane subtends at the center, φ, with the minimum angle, γ, that the $\widehat{s}\,'$ vector makes with the all-negative orthant.

First, φ can easily be inferred from trigonometry:

$$\cos\varphi = \frac{d_s}{W} = \frac{y \cot\theta |\cos\alpha|}{\sqrt{W^2 - y^2}} \ , \tag{A.30}$$

where $d_s = y \cot\theta |\cos\alpha|$ represents the distance from the center of the semi-constrained sphere to the hyperplane. The expression for d_s follows from the general mathematical result that if,

$$\vec{\beta} \cdot \vec{x} + \beta_0 = 0 \ , \tag{A.31}$$

is an equation for a hyperplane, where $\vec{x}$ denotes the coordinate vector and $\vec{\beta}, \beta_0$ are constants, then the perpendicular distance, $d_\perp$, to it from a point $\vec{x} = \vec{\zeta}$ is given by

$$d_\perp = \frac{|\vec{\beta} \cdot \vec{\zeta} + \beta_0|}{|\vec{\beta}|} \ . \tag{A.32}$$

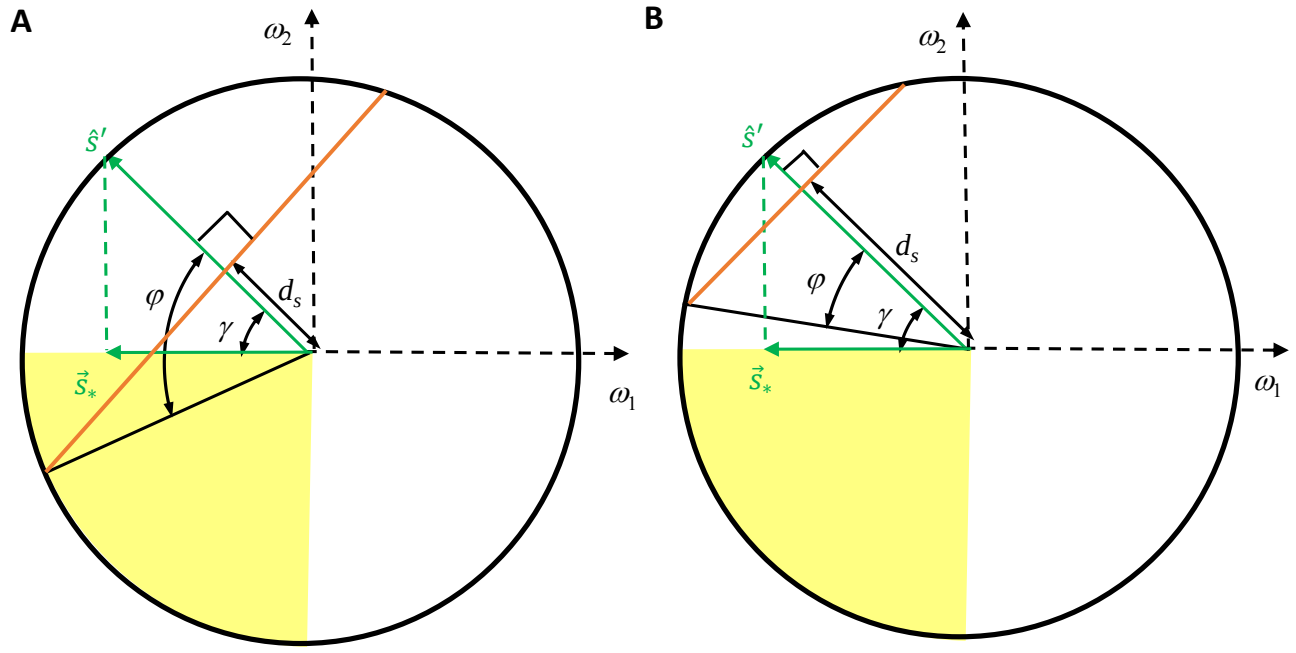

FIG. 7: **Cartoons depicting the orientation of the semi-constrained projection of a given synaptic weight direction ($\widehat{s}'$) within the semi-constrained subspace and its impact on determining the sign of the given weight.** In these plots, the yellow cones represent the solution space, $\omega_1, \omega_2 \leq 0$. d_s is the distance of the $w_m = 0$ orange line (hyperplane in higher dimension) from the origin. If d_s is small, as in the left plot (**A**), the projection angle γ is smaller than φ, half of the angle subtended by the orange line to the origin, and therefore the orange line and the yellow cone intersect. This means that solutions with both positive and negative w's are present. In the right plot (**B**), d_s is sufficiently large such that $\gamma > \varphi$ and consequently, all the solutions must have consistent sign.

Note, we are interested in the distance from the origin, $\vec{\zeta} = 0$, to the hyperplane satisfying the equation (A.28), so $\vec{\beta} = \sin\theta\ \widehat{s}' \Rightarrow |\vec{\beta}| = \sin\theta$ and $\beta_0 = -y\cos\theta|\cos\alpha|$.

To provide a geometric intuition for γ, let us first assume that $\widehat{s}'$ doesn't point into the all-negative orthant. If we can find the projection of $\widehat{s}'$ on the correct boundary of the solution space, then γ will be given by the angle between $\widehat{s}'$ and the appropriate semi-constrained boundary vector, $\vec{s}_*$ (Fig. 7). Since all the components in the solution space (all-negative orthant) have to be negative or zero, to find the appropriate projection vector of $\widehat{s}'$ onto the boundary of solution space, we essentially have to set all the positive components to zero:

$$\vec{s}_* = \sum_{\mu=\mathcal{C}+1}^{\mathcal{P}} s'_\mu \Theta(-s'_\mu)\vec{\epsilon}_\mu = \mp \sum_{\mu=\mathcal{C}+1}^{\mathcal{P}} s_\mu\ \Theta(\pm s_\mu)\vec{\epsilon}_\mu\ , \tag{A.33}$$

depending upon whether $\mathrm{Sgn}(e_y) = \pm$. Here s'_μ, s_μ are just the μ^{th} components of $\widehat{s}'$ and $\widehat{s}$ vectors, and $\Theta(x)$ is the Heaviside step function, which is one if x is positive and zero otherwise. Then, γ is given by

$$\cos\gamma = \widehat{s}' \cdot \widehat{s}_* = \frac{\widehat{s}' \cdot \vec{s}_*}{|\vec{s}_*|} = \sqrt{\sum_{\mu=\mathcal{C}+1}^{\mathcal{P}} s'^2_\mu\ \Theta(-s'_\mu)} = \sqrt{\sum_{\mu=\mathcal{C}+1}^{\mathcal{P}} s^2_\mu\ \Theta(\pm s_\mu)} = |\vec{s}_*|\ , \tag{A.34}$$

where again the sign is determined by the sign of e_y.

A formal way to see that γ is indeed given by (A.34) is to start with any unit vector, $\vec{w}_s$, lying in the solution space. Then, the angle, γ, between $\widehat{s}'$ and $\vec{w}_s$ is given by

$$\cos\gamma = \sum_{\mu=\mathcal{C}+1}^{\mathcal{P}} s'_\mu w_{s\mu} = \sum_{\mu\in A_+} s'_\mu w_{s\mu} + \sum_{\mu\in A_-} s'_\mu w_{s\mu}\ , \tag{A.35}$$

where we have defined $A_\pm$ as the set of all μ indices for which s'_μ is positive/negative, respectively. Since $\vec{w}_s$ is in the

solution space, $w_{s\mu} \leq 0$, and therefore the second term sums positive quantities while the first term subtracts. Thus,

$$\cos\gamma \leq \sum_{\mu\in A_-} s'_\mu w_{s\mu} = \vec{s}_* \cdot \vec{w}_s \leq |\vec{s}_*| \ , \tag{A.36}$$

where both the equalities are achieved when $\vec{w}_s$ is aligned with the boundary semi-constrained vector, $\vec{s}_*$, or $\vec{w}_s = \widehat{s}_*$, as argued previously. Note also, that this formal proof didn't assume any restrictions on $\widehat{s}'$ direction and thus (A.34) turns out to be a general result that also holds if $\widehat{s}'$ points into the all-negative quadrant.

Combining (A.34) and (A.30), the certainty condition now reads

$$\varphi < \gamma \Rightarrow \frac{y^2\cot^2\theta\cos^2\alpha}{W^2-y^2} > \cos^2\gamma \Rightarrow y > y_{\mathtt{cr}} \equiv W\sqrt{\frac{\cos^2\gamma\sin^2\theta}{\cos^2\alpha\cos^2\theta+\cos^2\gamma\sin^2\theta}} \ . \tag{A.37}$$

General case with unconstrained dimensions: We can extend the above analysis to the case when we have unconstrained dimensions by noting that, for a given set of unconstrained coordinates, the solution space is again the all-negative orthant in a semi-constrained hypersphere. Isometry along unconstrained dimensions ensures that it is always possible to make one of the null directions, lets say $\vec{\epsilon}_{\mathcal{N}}$, align with $\widehat{u}$. Then, the $w = 0$ hyperplane equation (A.26) reads

$$w = \sin\theta\cos\phi\ \widehat{s}\cdot\vec{w}_s + y\cos\theta\cos\alpha + \omega_u\sin\theta\sin\phi = 0 \ , \tag{A.38}$$

which can be rewritten as

$$\sin\theta\cos\phi\ \widehat{s}'\cdot\vec{w}_s = y\cos\theta|\cos\alpha| - \omega'_u\sin\theta\sin\phi \ , \tag{A.39}$$

where we have introduced $\omega'_u = -\mathrm{Sgn}(\cos\alpha)\omega_u$. To have a certain synapse, the $w = 0$ hyperplane cannot intersect the solution space for any allowed value of ω'_u.

The direction of $\widehat{s}'$ is independent of the unconstrained coordinates and hence the value of γ remains unchanged. However, the cone-angle, φ, does depend on the unconstrained coordinates in two ways. Firstly, the radius, $\widetilde{W}$, of the $\mathcal{S}$-dimensional spherical subspace containing admissible solutions is now:

$$\widetilde{W} = \sqrt{W^2 - y^2 - \omega_\perp^2 - \omega_u'^2} \tag{A.40}$$

where $\omega_\perp$ is the magnitude of the weight-vector in the $(\mathcal{U}-1)$ dimensional subspace that is perpendicular to $\vec{\epsilon}_{\mathcal{N}} = \widehat{u}$. We note in passing that (A.40) implies, $\omega_\perp, \omega'_u \leq \sqrt{W^2-y^2} = \overline{W}$. Secondly, the distance of the hyperplane from the origin that follows from (A.39) is now a function of ω'_u:

$$d_s = \frac{y|\cos\theta\cos\alpha| - \omega'_u\sin\theta\sin\phi}{\sin\theta\cos\phi} \ . \tag{A.41}$$

Strictly speaking, this expression for the distance is only valid as long as the numerator in the d_s expression stays positive. However, if there exists an allowed $\omega'_u \leq \overline{W}$ (let's call it ω_{u0}) for which the numerator can vanish, that would mean that the synapse cannot have a certain sign, because at that point $d_s = 0$, the hyperplane intersects the origin, and the weight can vanish even for a linear theory. In fact, the $d_s = 0$ condition provides us with the y-critical value below which the synapse sign becomes uncertain in a linear theory:

$$\omega_{u0} = \frac{y\cos\theta|\cos\alpha|}{\sin\theta\sin\phi} \leq \overline{W} \Rightarrow y_{\mathtt{cr,lin}} = W\sqrt{\frac{\sin^2\theta\sin^2\phi}{\cos^2\theta\cos^2\alpha+\sin^2\theta\sin^2\phi}} \ . \tag{A.42}$$

So we will now look into cases when $y \geq y_{\mathtt{cr,lin}}$ which also means that (A.41) will remain valid.

Combining (A.40) and (A.39) we get

$$\cos\varphi = \frac{d_s}{\widetilde{W}} = \frac{y|\cos\theta\cos\alpha| - \omega'_u\sin\theta\sin\phi}{\sin\theta\cos\phi\sqrt{W^2-y^2-\omega_\perp^2-\omega_u'^2}} \ . \tag{A.43}$$

In order for us to be certain that w is nonzero, we have to make sure that even the largest φ that one can obtain by varying $\omega_\perp$ and ω'_u is still smaller than γ. Clearly, to make φ large it is best to make $\omega_\perp = 0$. Also, it is clear from inspection that $\cos\varphi$ starts to initially decrease as ω'_u increases from zero, being dominated by the linear term.

However, as the quadratic term in ω'_u in the denominator becomes more and more important, $\cos\varphi$ reaches a minimum and starts to increase. Imposing $d\cos\varphi/d\omega'_u = 0$, we can find that this minimum is reached at

$$\omega'_u = \frac{\sin\theta\sin\phi\ (W^2-y^2)}{y\cos\theta|\cos\alpha|} = \overline{W}\sqrt{\frac{(W/y)^2-1}{(W/y_{\mathtt{cr,lin}})^2-1}} \leq \overline{W}\ , \tag{A.44}$$

where we substituted $y_{\mathtt{cr,lin}}$ from (A.42) and used the fact that $W \geq y \geq y_{\mathtt{cr,lin}}$ to obtain the inequality. This proves that the minimum $\cos\varphi$ indeed occurs at an allowed positive value of $\omega'_u \leq \overline{W}$. Substituting the above ω'_u in (A.43), after some algebra we find that this minimum value of $\cos\varphi$, or equivalently the maximum φ, is given by

$$\cos\varphi = \sqrt{\frac{y^2\cos^2\theta\cos^2\alpha - (W^2-y^2)\sin^2\theta\sin^2\phi}{(W^2-y^2)\cos^2\phi\sin^2\theta}}\ . \tag{A.45}$$

The certainty condition then requires

$$\cos^2\varphi = \frac{y^2\cos^2\theta\cos^2\alpha - (W^2-y^2)\sin^2\theta\sin^2\phi}{(W^2-y^2)\cos^2\phi\sin^2\theta} > \cos^2\gamma\ , \tag{A.46}$$

which can be recast as

$$y > y_{\mathtt{cr}} \equiv W\sqrt{\frac{\cos^2\gamma\sin^2\theta\cos^2\phi + \sin^2\theta\sin^2\phi}{\cos^2\alpha\cos^2\theta + \cos^2\gamma\sin^2\theta\cos^2\phi + \sin^2\theta\sin^2\phi}}\ , \tag{A.47}$$

It is illuminating to express y-critical in terms of the projections, e_y, e_u, e_{s*}, of the synaptic direction, $\widehat{e}$, respectively along the data vector, $\widehat{y}$, the unconstrained unit vector, $\widehat{u}$, and the semi-constrained boundary vector, $\vec{s}_*$:

$$y_{\mathtt{cr}} = W\sqrt{\frac{e_{s*}^2+e_u^2}{e_y^2+e_{s*}^2+e_u^2}}\ , \tag{A.48}$$

where

$$e_y \equiv \widehat{e}\cdot\widehat{y} = \frac{\sum_{\mu=1}^{\mathcal{C}} y_\mu e_\mu}{\sqrt{\sum_{\mu=1}^{\mathcal{C}} y_\mu^2}} = \cos\theta\cos\alpha\ ,\ e_{s*} \equiv \widehat{e}\cdot\widehat{s}_* = \sqrt{\sum_{\mu\in A_-} e_\mu^2} = -\mathrm{Sgn}(\cos\alpha)\sin\theta\cos\phi\cos\gamma\ , \tag{A.49}$$

and e_u is given by (A.22). We note that setting $\phi = 0$ precisely reproduces the correct limit with no unconstrained directions (A.37).

Regarding orthogonal input patterns in recurrent networks

While our analysis of the solution space and the certainty condition (A.48) translate directly to recurrent networks, the requirement of orthogonality for the derivation of our certainty condition imposes certain technical restrictions on its scope when it comes to recurrent neural networks (RNNs).

The certainty condition we derived for feedforward networks can be applied to two different RNN set ups. First, let us consider RNNs where neurons have self-couplings. A consequence of having orthogonal response patterns in this case is that the certainty condition can only be satisfied for self-couplings w_{ii}, as long as $W \geq 1$. This is because the imposition of orthogonality in response patterns also restricts the correlation between the target neuron and the other neurons:

$$\sum_{m=1}^{\mathcal{N}} Z_{\mu m}Z_{\nu m} = \delta_{\mu\nu} \Rightarrow \sum_{\mu=1}^{\mathcal{N}} Z_{\mu m}Z_{\mu n} = \delta_{mn}\ . \tag{A.50}$$

However, for the synapse-sign to be certain, the responses of pre and postsynaptic neurons need to be correlated. To see the problem more quantitatively, suppose we are interested in constraining the synapse from the m^{th} neuron onto the i^{th} neuron, as before. Now, the first $\mathcal{P}$ elements of the unit vectors, $\widehat{e}_i$ and $\widehat{e}_m$ contain the responses of the i^{th} and the m^{th} neuron in the $\mathcal{P}$ patterns. We have already derived a decomposition of $\widehat{e}_m$ in terms of its projections onto the constrained, semi-constrained and unconstrained subspaces (A.20). Similarly, $\widehat{e}_i$ can be decomposed as

$$\widehat{e}_i = y\widehat{y} + \sqrt{1-y^2}\widehat{y}_\perp \tag{A.51}$$

where $\widehat{y}$ lies entirely along the constrained directions, and $\widehat{y}_\perp$ is orthogonal to it and only has components along unconstrained directions. Then, orthogonality implies

$$\widehat{e}_i \cdot \widehat{e}_m = y\cos\theta\cos\alpha + \sqrt{1-y^2}\sin\theta\sin\phi\ (\widehat{y}_\perp \cdot \widehat{u}) = 0 \Rightarrow \sin\theta\sin\phi = -\frac{y\cos\theta\cos\alpha}{(\widehat{y}_\perp \cdot \widehat{u})\sqrt{1-y^2}}\ . \tag{A.52}$$

Starting from the certainty condition (A.48), we can now go through a sequence of (in)equalities:

$$\begin{aligned} y^2 \ &> \ W^2 \frac{\sin^2\theta\sin^2\phi + \sin^2\theta\cos^2\phi\cos^2\gamma}{\sin^2\theta\sin^2\phi + \sin^2\theta\cos^2\phi\cos^2\gamma + \cos^2\theta\cos^2\alpha} \geqslant \frac{W^2\sin^2\theta\sin^2\phi}{\sin^2\theta\sin^2\phi + \cos^2\theta\cos^2\alpha} \\ &= \ \frac{W^2y^2\cos^2\theta\cos^2\alpha}{y^2\cos^2\theta\cos^2\alpha + (1-y^2)(\widehat{y}_\perp\cdot\widehat{u})^2\cos^2\theta\cos^2\alpha} = \frac{W^2y^2}{y^2+(1-y^2)(\widehat{y}_\perp\cdot\widehat{u})^2} \ \geq W^2y^2, \end{aligned} \tag{A.53}$$

where we substituted $\sin\theta\sin\phi$ from (A.52). Note that the RHS is minimized when $\widehat{u}$ and $\widehat{y}_\perp$ are either aligned or anti-aligned. Even in this case, RHS $= W^2y^2$, and thus the certainty condition cannot be satisfied if $W \geq 1$. One can check that when $i = m$, because the RHS in the first equation of (A.52) is one and not zero, no similar constraints appear. Indeed, the certainty condition may be satisfied depending upon the specific response patterns.

As a second possibility, suppose that no self-couplings are present. Then to be able to apply our framework and determine the couplings w_{im} for a given i, we only need the truncated row vectors of z whose i^{th} column entry is absent, to be orthonormal. Therefore, the response of the i^{th} driven neuron, which consists of the entries of the i^{th} column, can now be chosen independently from the responses of its input neurons. In other words, $\widehat{e}_i$ and $\widehat{e}_m$, $m \neq i$, no longer need to satisfy orthogonality constraint of (A.52). Consequently, the w_{im} connectivities can indeed satisfy the certainty condition, just as in the feedforward case.

B. Estimating the Probability that a Synapse is Certain in Large Feedforward Networks:

For given values of $\mathcal{N} = \mathcal{I}, \mathcal{C}, \mathcal{P}$ in a feedforward setting we will here try to assess how likely is it that noiseless orthonormal neuronal responses require a given synapse to be nonzero. As we have seen (14), whether a synapse is certain to exist depends on six parameters, $\theta, \phi, \gamma, \alpha, W$, and y. The first four quantities depend on how $\widehat{e}$ is oriented with respect to various directions in the weight space. Since $\widehat{e}$ is a unit vector, typically we expect its component along any given direction to be $\sim \mathcal{O}(1/\sqrt{\mathcal{N}})$. Thus, we typically expect

$$e_y^2 = \cos^2\theta\cos^2\alpha \sim \frac{1}{\mathcal{N}}\ ;\ e_s^2 = \sin^2\theta\cos^2\phi \sim \frac{\mathcal{S}}{\mathcal{N}}\ ;\ e_u^2 = \sin^2\theta\sin^2\phi \sim \frac{\mathcal{U}}{\mathcal{N}} \ \text{ and } \ e_{s*}^2 = \cos^2\gamma\ e_s^2 \sim \frac{\mathcal{S}}{2\mathcal{N}}\ . \tag{B.1}$$

Hence we approximate the typical $y_{\mathtt{cr}}$ as

$$y_{\mathtt{cr}} = W\sqrt{\frac{\mathcal{S}/(2\mathcal{N}) + \mathcal{U}/\mathcal{N}}{1/\mathcal{N} + \mathcal{S}/(2\mathcal{N}) + \mathcal{U}/\mathcal{N}}} = W\sqrt{\frac{\mathcal{S}+2\mathcal{U}}{2+\mathcal{S}+2\mathcal{U}}}\ . \tag{B.2}$$

Let us now suppose that all the dimensions scale with the network size, such that

$$\mathcal{S} = \sigma\mathcal{N},\ \text{and}\ \mathcal{U} = \upsilon\mathcal{N}\ . \tag{B.3}$$

Then, we find that as the network size increases $y_{\mathtt{cr}}$ behaves as

$$\frac{y_{\mathtt{cr}}}{W} \approx 1 - \left(\frac{1}{\sigma+2\upsilon}\right)\frac{1}{\mathcal{N}}\ , \tag{B.4}$$

and $y_{\mathtt{cr}}$ is essentially pushed up towards W.

However, the typical scale of y behaves similarly as the dimensions increase. To see this concretely, let us define $\vec{y}_{cons} \equiv \{y_\mu | \mu = 1 \ldots \mathcal{C}\}$ as a $\mathcal{C}$-dimensional vector, and assume that every possible $\vec{y}_{cons}$ is equally likely within a sphere of radius W (larger activity levels of the target neuron admit no solutions). Then the average and median values of $y = |\vec{y}_{cons}|$ are given by

$$\begin{aligned} &< y >= \frac{\int_0^W dy y^{\mathcal{C}}}{\int_0^W dy y^{\mathcal{C}-1}} = W\left(\frac{\mathcal{C}}{\mathcal{C}+1}\right) \ \Rightarrow \ \frac{< y >}{W} = \frac{1}{1+1/(\eta\mathcal{N})} \approx 1 - \frac{1}{\eta\mathcal{N}}\ ,\ \text{where}\ \eta \equiv \frac{\mathcal{C}}{\mathcal{N}}\ , \\ \text{and} \ &\frac{\int_0^{y_M} dy y^{\mathcal{C}-1}}{\int_0^W dy y^{\mathcal{C}-1}} = \frac{1}{2} \Rightarrow \frac{y_M}{W} = \left(\frac{1}{2}\right)^{\frac{1}{\mathcal{C}}} \ \Rightarrow \ \frac{y_M}{W} \approx 1 - \frac{\ln 2}{\eta\mathcal{N}}\ , \end{aligned} \tag{B.5}$$

respectively. Since y and $y_{\mathtt{cr}}$ scale similarity as one increases the network size, the probability of a synapse being certain should not change as the network size increases. In Fig. 6C, we show that if we choose y as the median value in simulations with random input-output configurations (see Appendix E for details), then the number of certain synapses does indeed increase linearly with $\mathcal{N}$.

To quantitatively estimate the probability of finding a certain synapse, we can compute the fraction of volume of $\vec{y}_{cons}$'s for which the synapse is certain for the typical projections (B.2), as compared to the volume of $\vec{y}_{cons}$'s for which solutions to the steady-state equations exist. We know that $\vec{y}_{cons}$ has to lie within a $\mathcal{C}$-dimensional sphere of radius W in order for there to be any solutions to the problem [2]. On the other hand, for the synapse sign to be identifiable, we need $W \geq |\vec{y}_{cons}| = y > y_{\mathtt{cr}}$, where $y_{\mathtt{cr}}$ is given by (B.2). So, we need to compare the spherical shell volume, $V_{W\geq y>y_{\mathtt{cr}}}$, with the volume of the $\mathcal{C}$-dimensional sphere, $V_{y\leq W}$. In order to find $V_{W\geq y>y_{\mathtt{cr}}}$, we have to subtract the $\mathcal{C}$-dimensional spherical volume with radius $y_{\mathtt{cr}}$ from the spherical volume with radius W. Since n-dimensional spherical volumes scale as the nth power of the radius, the probability, P, of ascertaining the sign of the synapse is approximately given by

$$P \approx \frac{V_{W\geq y>y_{\mathtt{cr}}}}{V_{y\leq W}} = \frac{W^{\mathcal{C}} - y_{\mathtt{cr}}^{\mathcal{C}}}{W^{\mathcal{C}}} = 1 - \left(\frac{y_{\mathtt{cr}}}{W}\right)^{\mathcal{C}} . \tag{B.6}$$

Now, when $\mathcal{S},\mathcal{U} \gg 1$ we can approximately evaluate the RHS as follows:

$$\begin{aligned} \left(\frac{y_{\mathtt{cr}}}{W}\right)^{\mathcal{C}} &= \left(\frac{\mathcal{S}+2\mathcal{U}}{2+\mathcal{S}+2\mathcal{U}}\right)^{\mathcal{C}/2} = \left(1+\frac{2}{\mathcal{S}+2\mathcal{U}}\right)^{-\mathcal{C}/2} \\ \Rightarrow \ln\left(\left(\frac{y_{\mathtt{cr}}}{W}\right)^{\mathcal{C}}\right) &= -\frac{\mathcal{C}}{2}\ln\left(1+\frac{2}{\mathcal{S}+2\mathcal{U}}\right) = \frac{-\mathcal{C}}{\mathcal{S}+2\mathcal{U}}\left[1+\mathcal{O}\left(\frac{1}{\mathcal{S}+2\mathcal{U}}\right)\right] . \end{aligned} \tag{B.7}$$

Thus we get

$$P \approx 1 - e^{-\frac{\mathcal{C}}{\mathcal{S}+2\mathcal{U}}} . \tag{B.8}$$

The most prominent feature of (B.8) is that the probability only depends on the ratios of the various dimensions. Hence it doesn't change as we increase the size of the network as long as the ratios are kept constant.

For the purpose of illustration and numerically testing this feature we assessed how certainty predictions changed when the network size is increased while holding the ratios between $\mathcal{C}$, $\mathcal{S}$ and $\mathcal{U}$ fixed. In Fig. 6C we have plotted the number of certain synapses in simulations generated from random data as we scale up $\mathcal{N}$ maintaining the ratios between $\mathcal{C},\mathcal{S}$ and $\mathcal{U}$ (see Appendix E for more details). We illustrate two cases. In the first example, no unconstrained directions were present, and $\mathcal{S} = 3\mathcal{C}$. Then $P = 1 - e^{-1/3} \approx 0.28$, so one has a 28% chance of being able to determine the sign of the connections. This answer incidentally is the same as an example with $\mathcal{C} = \mathcal{S} = \mathcal{U}$. As another example, Fig. 6C considered the case when $\mathcal{U} = 2\mathcal{C} = 2\mathcal{S}$. According to (B.8), then $P = 1 - e^{-1/5} \approx 0.18$, so the chance of determining the sign drops to about 18%. We only expect these numbers to be approximate. For example, our arguments relied on the assumption that all target responses admitting solutions are equally likely, an assumption that definitely needs to be revisited for realistic networks. However, the scaling behavior should hold for other probabilistic distributions as long as the scale of $\vec{y}_{cons}$ behaves similarly to (B.5) with increasing $\mathcal{N}$.

C. Nonzero-error Certainty Conditions

There are various reasons why we may want to not only consider weights that exactly reproduce the specified neuronal responses, but also weights that do so approximately. For instance, we are always limited by the accuracy of the measurement apparatus. More importantly, there are various sources of biological noise that typically lead to uncertainties in observed values of neuronal responses. For the purpose of this paper we will consider any set of weights to be part of the ϵ-error solution space if it is able to reproduce the specified neuronal responses with an error $\mathcal{E} \leq \epsilon$ (see Eq. (20) for definition of $\mathcal{E}$). We will neglect uncertainties in the input responses to the target neuron, but we will comment on their possible effects towards the end of this appendix.

Errors in feedforward networks

Let us first focus on feedforward networks. Allowing for error increases the value of $y_{\mathtt{cr}}$ by expanding the solution space. One way to think about this is to realize that we have to now make sure that (14) is satisfied not only for

---

[2] The allowed $\vec{y}_{cons}$ must also lie in the all positive orthant, but as we will compute the ratio of two spherical volumes the reduction factor will cancel out.

$\vec{y} = \vec{y}_0$ but for any non-negative $\vec{y} = \vec{y}_0 + \vec{\delta}$, where $\vec{\delta}$ is a vector in the $\mathcal{P}$-dimensional activity-constrained subspace with $|\vec{\delta}| \leq \epsilon$. So our strategy will be to first find the minimum y_0 needed to have a certain synapse for a given $\vec{\delta}$. We then find the maximum among these y_0 values as we let $\vec{\delta}$ vary within the ϵ-ball. Since this procedure will guarantee that the $w = 0$ hyperplane doesn't intersect the entire solution space with $\mathcal{E} \leq \epsilon$, this means that the synapse must exist for the network to generate the specified responses patterns, and the synapse's sign will match the zero-error analysis. We will first estimate y-critical when the error is small enough to not induce topological transitions in the error surface. In the subsequent section, we will include the effects of topological transitions.

When no topological transitions occur: To understand how errors affect the certainty-conditions, let us consider the case when the allowed error, $\epsilon < \min\{y_\mu\}_{\mu=1,\ldots,\mathcal{C}}$, so that no topological transitions can occur. Without loss of generality, let us assume that $\vec{\delta}$ only has nonzero components along the first $\mathcal{Q}$ activity-constrained dimensions. If $\mathcal{Q} > \mathcal{C}$ (*i.e.* some responses which were previously zero are now nonzero), then both e_y and e_{s*} can change. e_y changes to

$$\overline{e}_y = \frac{\widehat{e} \cdot (\vec{y}_0 + \vec{\delta})}{|\vec{y}_0 + \vec{\delta}|} \ . \tag{C.1}$$

And, if some previously semi-constrained components that contributed to $\widehat{s}_*$ have now become constrained [3], then $\widehat{s}_*$ no longer has those components. This means that we have to subtract these components from e_{s*}:

$$\widehat{\overline{s}}_* = \widehat{s}_* - \sum_{\mu=\mathcal{C}+1}^{\mathcal{Q}} A_\mu e_\mu \vec{\epsilon}_\mu \ \Rightarrow \overline{e}^2_{s*} = e^2_{s*} - \sum_{\mu=\mathcal{C}+1}^{\mathcal{Q}} A_\mu e^2_\mu \ , \tag{C.2}$$

where A_μ is 1 if $(\widehat{y} \cdot \widehat{c})$ and e_μ have the same sign and 0 otherwise. This follows from the definition of the boundary projection vector (A.33) and e_{s*} (A.49). Thus, for a given $\vec{\delta}$ the certainty condition (A.48) yields

$$|\vec{y}|^2 = |\vec{y}_0 + \vec{\delta}|^2 > W^2 \left(\frac{\overline{e}^2_{s*} + e^2_u}{\overline{e}^2_y + \overline{e}^2_{s*} + e^2_u} \right) \Rightarrow \frac{|\widehat{e} \cdot (\vec{y}_0 + \vec{\delta})|^2}{\overline{e}^2_{s*} + e^2_u} + |\vec{y}_0 + \vec{\delta}|^2 > W^2. \tag{C.3}$$

As before, one can interpret the above inequality as equivalently specifying either y-critical or W-critical. For a fixed $\vec{y}_0$, one can obtain a minimum value of the left hand side (LHS) of the latter inequality by varying $\vec{\delta}$ within the ϵ-ball. The square root of this is W-critical. Then as long as W is less than W-critical, we will have a certain synapse. Inverting the relation, one finds y-critical as the minimum y_0 needed to make the synapse sign certain for all $\vec{\delta}$ and given $\widehat{y}_0$ and W. More explicitly, equating the two sides of the inequality for any given $\widehat{y}_0$, $\vec{\delta}$, and W, we get a minimal-y_0 that depends on $\vec{\delta}$. To find y-critical, we have to take the maximum of the minimal-y_0 as we vary over all possible $\vec{\delta}$ in the ϵ-ball.

Let us first obtain a lower bound on y-critical. By inspection of the LHS of the above inequality, it is clear that the more the $\vec{\delta}$-dependent terms can cancel the $\vec{y}_0$-dependent terms, the harder it is to satisfy the certainty condition. We observe that in (C.3), the second term is minimized when $\vec{\delta} = -\epsilon \widehat{y}_0$ [4]. Accordingly, one can obtain a lower bound on y-critical by substituting $\delta = -\epsilon \widehat{y}_0$ in (C.3):

$$(\widehat{e} \cdot \widehat{y}_0)^2 (y_0 - \epsilon)^2 + (e^2_{s*} + e^2_u)(y_0 - \epsilon)^2 > W^2 (e^2_{s*} + e^2_u) \ , \text{ or, } y_0 > y_{\mathtt{cr},\min} \equiv W \sqrt{\frac{e^2_{s*} + e^2_u}{e^2_y + e^2_{s*} + e^2_u}} + \epsilon \ , \tag{C.4}$$

where we have used $e_y = \widehat{e} \cdot \widehat{y}_0$ and $\overline{e}^2_{s*} = e^2_{s*}$, since $\vec{\delta}$ has no components along the semi-constrained directions. We will see later that this simple lower bound approximates the actual y-critical very well in many situations.

Next, we can find an upper bound for y-critical by noting

$$\frac{|\widehat{e} \cdot (\vec{y}_0 + \vec{\delta})|^2}{\overline{e}^2_{s*} + e^2_u} + |\vec{y}_0 + \vec{\delta}|^2 \geq \frac{|\widehat{e} \cdot (\vec{y}_0 + \vec{\delta})|^2}{e^2_{s*} + e^2_u} + |\vec{y}_0 + \vec{\delta}|^2 \geq \frac{(\widehat{e} \cdot \vec{y}_0)^2 + 2(\widehat{e} \cdot \vec{y}_0)(\widehat{e} \cdot \vec{\delta})}{e^2_{s*} + e^2_u} + y_0^2 + 2 \ \vec{y}_0 \cdot \vec{\delta} \ , \tag{C.5}$$

---

[3] This will happen if a component of $\vec{y}$ that was zero now has a nonzero component. The component of $\vec{\delta}$ along that direction is positive.

[4] This assumes that $y_0 > \epsilon$. Smaller values of y_0 permit $\vec{y} = \vec{0}$ and all weights can be set to zero

where the first inequality is true because $\overline{e}^2_{s*} \leq e^2_{s*}$, and the second inequality because we have dropped positive $\mathcal{O}(\delta^2)$ terms. Then we can obtain an upper bound on y-critical by finding a y_0 such that even the last expression on the RHS is greater than W^2. Specifically,

$$(\widehat{e}\cdot\vec{y}_0)^2 + 2(\widehat{e}\cdot\vec{y}_0)(\widehat{e}\cdot\vec{\delta}) + (e^2_{s*}+e^2_u)(y_0^2 + 2\ \vec{y}_0\cdot\vec{\delta}) > W^2(e^2_{s*}+e^2_u)\ , \tag{C.6}$$

So, let us try to find the $\vec{\delta}$ that minimizes the LHS:

$$\begin{aligned}\text{LHS} &= y_0^2 e_y^2 + 2y_0 e_y(\widehat{e}\cdot\vec{\delta}) + (e^2_{s*}+e^2_u)(y_0^2 + 2y_0\ \widehat{y}_0\cdot\vec{\delta}) \\ &= y_0^2(e_y^2+e^2_{s*}+e^2_u) + 2y_0(e_y\widehat{e} + (e^2_{s*}+e^2_u)\widehat{y}_0)\cdot\vec{\delta} \\ &= y_0^2(e_y^2+e^2_{s*}+e^2_u) + 2y_0\ \vec{\xi}\cdot\vec{\delta}\end{aligned} \tag{C.7}$$

where $\vec{\xi} \equiv e_y\sum_{\mu=1}^{\mathcal{P}} e_\mu\vec{\epsilon}_\mu + (e^2_{s*}+e^2_u)\widehat{y}_0$, and we have noted that $\widehat{u}\cdot\vec{\delta}=0$ because $\vec{\delta}$ must be in the activity-constrained subspace. It is now clear that LHS is minimized if $\vec{\delta}$ anti-aligns with $\vec{\xi}$. Then (C.6) yields

$$y_0^2(e_y^2+e^2_{s*}+e^2_u) - 2y_0|\vec{\xi}|\epsilon > W^2(e^2_{s*}+e^2_u)\ . \tag{C.8}$$

Equating the two sides of (C.8) and solving for y_0 [5], we now get an upper bound for y-critical:

$$y_{\mathtt{cr},\max} \equiv \sqrt{W^2\left(\frac{e^2_{s*}+e^2_u}{e_y^2+e^2_{s*}+e^2_u}\right) + \frac{\epsilon^2\xi^2}{(e_y^2+e^2_{s*}+e^2_u)^2}} + \frac{\epsilon\xi}{e^2_{s*}+e^2_u+e_y^2}, \tag{C.9}$$

where ξ is the norm of $\vec{\xi}$ and can be simplified as

$$\begin{aligned}\xi^2 &= \sum_{\mu=1}^{\mathcal{P}}[e_y e_\mu + (e^2_{s*}+e^2_u)\widehat{y}_{0,\mu}]^2 = e_y^2\sum_{\mu=1}^{\mathcal{P}} e_\mu^2 + (e^2_{s*}+e^2_u)^2\sum_{\mu=1}^{\mathcal{C}}\widehat{y}^2_{0,\mu} + 2(e^2_{s*}+e^2_u)e_y\sum_{\mu=1}^{\mathcal{C}}\widehat{y}_{0,\mu}e_\mu \\ &= e_y^2e_p^2 + (e^2_{s*}+e^2_u)^2 + 2(e^2_{s*}+e^2_u)e_y^2 = (e_y^2+e^2_{s*}+e^2_u)^2 + e_y^2(e_p^2-e_y^2)\ .\end{aligned} \tag{C.10}$$

Thus, we have

$$y_{\mathtt{cr},\max} \equiv \sqrt{W^2\left(\frac{e^2_{s*}+e^2_u}{e_y^2+e^2_{s*}+e^2_u}\right) + \epsilon^2\left(1+\frac{e_y^2(e_p^2-e_y^2)}{(e_y^2+e^2_{s*}+e^2_u)^2}\right)} + \epsilon\sqrt{1+\frac{e_y^2(e_p^2-e_y^2)}{(e_y^2+e^2_{s*}+e^2_u)^2}}\ . \tag{C.11}$$

Finally, we would like to point out that for small errors one can also obtain a leading order correction to y-critical that lies in between $y_{\mathtt{cr},\min}$ and $y_{\mathtt{cr},\max}$. To obtain this estimate, let us first write down the bound on y_0 that one would obtain from (C.3) as $\delta\to 0$:

$$y_0 > W\sqrt{\frac{\overline{e}^2_{s*}+e^2_u}{\overline{e}_y^2+\overline{e}^2_{s*}+e^2_u}}\ . \tag{C.12}$$

If $\vec{\delta}$ has components along any semi-constrained direction that contributes towards the original $\widehat{e}_{s*}$ vector, then $\overline{e}^2_{s*} < e^2_{s*}$, and comparing (A.48) and (C.12) we see that, as $\delta\to 0$, the bound on y_0 will be less than the zero-error $y_{\mathtt{cr}}$. In other words, for sufficiently small errors, if $\vec{\delta}$ explores directions that contribute to e_{s*}, then the corresponding bound on y_0 is going to be smaller than even the zero-error $y_{\mathtt{cr}}$. Thus, for these small errors the leading order corrections to (A.48) is obtained only if $\vec{\delta}$ do not have any components along these semi-constrained directions. This means $\overline{e}^2_{s*} = e^2_{s*}$, and we can reorder the indices such that the semi-constrained directions along which excursions of $\vec{\delta}$ will be considered range from $\mathcal{C}+1,\ldots,\mathcal{Q}$, *i.e.*, $\mathrm{Sgn}(\widehat{s}'_\mu)$ is negative for these, and only these, semi-constrained indices. To obtain the certainty condition, one can then follow steps (C.5) [6] through (C.9) except that $\vec{\delta}$ is restricted to only

---

[5] This quadratic equation obviously has two solutions. The correct one can easily be identified, for instance, by taking the $\epsilon\to 0$ limit.

[6] Since we are only interested in the leading order correction, we could also drop the $\mathcal{O}(\delta^2)$ terms needed to arrive at an expression such as the RHS of (C.5).

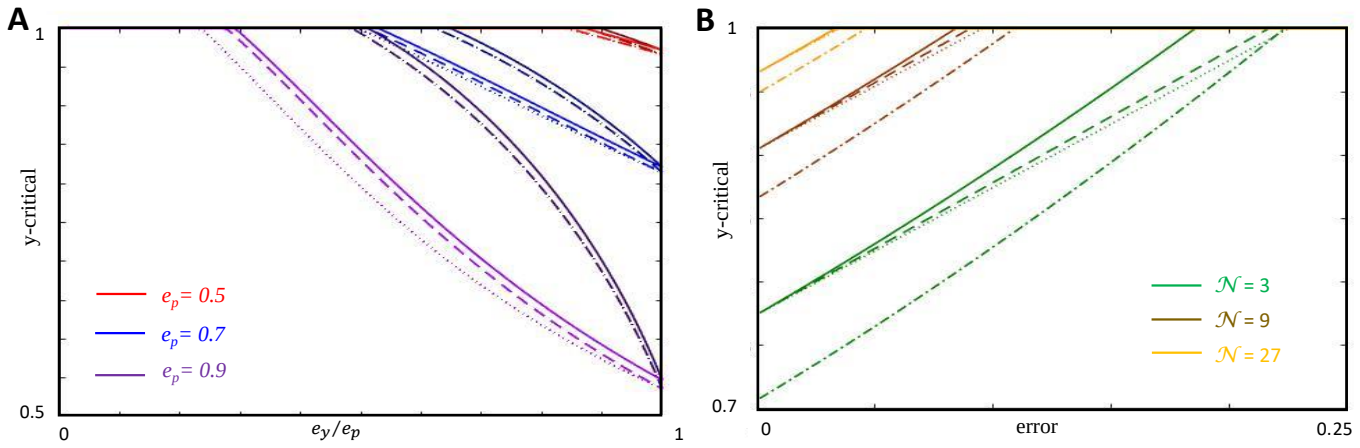

FIG. 8: **Dependence of y-critical on various parameters for nonzero errors.** (**A**) The red, blue and purple curves track y-critical as a function of e_y/e_p for $e_p = 0.5, 0.7$, and 0.9, respectively. The dotted, dashed and bold curves represent the lower bound, leading-order and upper bound y-critical curves for a fixed error, $\epsilon = 0.1W$. The darker shade correspond to the most nonlinear case when $e_{s*}/\sqrt{e_p^2 - e_y^2} = 1$, while the lighter shade correspond to $e_{s*} = 0$. These latter curves are also the ones that one obtains in a linear theory. Clearly, the difference between the linear and nonlinear theory increases as e_p increases. In all these cases y-critical decreases with increase of e_p, and for a given e_p, as e_y/e_p increases. Also, as e_{s*} increases and the semi-constrained dimensions become more important, it becomes harder to constrain the synapse sign, and therefore y-critical increases. (**B**) The green, brown and orange curves again track y-critical, but this time as a function of ϵ, for networks with $\mathcal{N} = 3, 9$, and 27 input neurons, respectively. The dotted, dashed and bold curves plot the lower bound, leading-order and upper bound on y-critical for typical values of e_p, e_y and e_{s*} that one expects in these networks (B.1). We see that these curves come closer together as the network size increases. The dot-dashed curves correspond to the linear theory ($e_{s*} = 0$), which remains clearly separated from the nonlinear curves. In each of these networks, $\mathcal{P}/\mathcal{N} = 2/3$ and $\mathcal{C}/\mathcal{P} = 1/2$.

have nonzero components along constrained directions and those semi-constrained directions that do not contribute to $\widehat{e}_{s*}$, *i.e.*, for $\mu = 1 \ldots \mathcal{Q}$. In other words, it can at the most anti-align with a truncated $\vec{\xi}$,

$$\vec{\xi}_{\texttt{trunc}} \equiv \sum_{\mu=1}^{\mathcal{Q}} \xi_\mu \vec{\epsilon}_\mu = \sum_{\mu=1}^{\mathcal{Q}} [e_y e_\mu + (e_{s*}^2 + e_u^2)\widehat{y}_{0,\mu}]\vec{\epsilon}_\mu \ , \ \text{and} \ (\vec{\xi}\cdot\vec{\delta})_{\min} = -\epsilon\xi_{\texttt{trunc}} \ , \tag{C.13}$$

where $\xi_{\texttt{trunc}}$ is the norm of $\vec{\xi}_{\texttt{trunc}}$ and can be simplified as

$$\begin{aligned} \xi_{\texttt{trunc}}^2 &= \sum_{\mu=1}^{\mathcal{Q}} [e_y e_\mu + (e_{s*}^2 + e_u^2)\widehat{y}_{0,\mu}]^2 = e_y^2 \sum_{\mu=1}^{\mathcal{Q}} e_\mu^2 + (e_{s*}^2 + e_u^2)^2 \sum_{\mu=1}^{\mathcal{C}} \widehat{y}_{0,\mu}^2 + 2(e_{s*}^2 + e_u^2) e_y \sum_{\mu=1}^{\mathcal{C}} \widehat{y}_{0,\mu} e_\mu \\ &= e_y^2(e_p^2 - e_{s*}^2) + (e_{s*}^2 + e_u^2)^2 + 2(e_{s*}^2 + e_u^2)e_y^2 = (e_y^2 + e_{s*}^2 + e_u^2)^2 + e_y^2(e_p^2 - e_{s*}^2 - e_y^2) \ . \end{aligned} \tag{C.14}$$

Substituting $\xi = \xi_{\texttt{trunc}}$ into the counterpart of (C.9), and keeping only the linear terms in ϵ, we thus get the leading order correction to (A.48):

$$y_{\texttt{cr,lead}} \approx W\sqrt{\frac{e_{s*}^2 + e_u^2}{e_y^2 + e_{s*}^2 + e_u^2}} + \frac{\epsilon\xi_{\texttt{trunc}}}{e_y^2 + e_{s*}^2 + e_u^2} = W\sqrt{\frac{e_{s*}^2 + e_u^2}{e_y^2 + e_{s*}^2 + e_u^2}} + \epsilon\sqrt{1 + \frac{e_y^2(e_p^2 - e_{s*}^2 - e_y^2)}{(e_y^2 + e_{s*}^2 + e_u^2)^2}} \ . \tag{C.15}$$

Reassuringly, we see that at $\epsilon = 0$, $y_{\texttt{cr,lead}}$, $y_{\texttt{cr,max}}$ and $y_{\texttt{cr,min}}$, all reduce to the zero-error $y_{\texttt{cr}}$ (A.48). Also it is obvious that the coefficient of ϵ in $y_{\texttt{cr,lead}}$ is greater than that of $y_{\texttt{cr,min}}$ but less than that of $y_{\texttt{cr,max}}$. Finally, note that $y_{\texttt{cr,lead}}$ coincides with $y_{\texttt{cr,min}}$ in the maximally nonlinear case where $e_{s*}^2 = e_p^2 - e_y^2$. In Fig. 8A, we have plotted how these different quantities depend on e_p, e_y and e_{s*}. In particular we note that as the network size increases, these curves typically come closer together (Fig. 8B), so that the lower bound already provides a good approximation for y-critical.

Effects of topological transitions: Let us next consider a situation where a topological transition in the error surface occurs as one moves from $\mathcal{E} = 0$ to $\mathcal{E} = \epsilon$. Suppose this happens for the $\mathcal{C}^{th}$ direction because $y_{\mathcal{C}} < \epsilon$. This effectively means that for errors more than $y_{\mathcal{C}}$, one needs to also consider the possibility that the $\vec{\epsilon}_{\mathcal{C}}$ direction behaves as a

semi-constrained direction rather than a constrained direction. To account for this case, one needs to revisit (C.4), (C.11), and (C.15), and calculate $y_{\texttt{cr,min}}$, $y_{\texttt{cr,max}}$, and $y_{\texttt{cr,lead}}$, respectively with $\vec{\epsilon}_{\mathcal{C}}$ included among the semi-constrained dimensions. Also, $\epsilon^2 \to \epsilon^2 - y_{\mathcal{C}}^2$, since we are committed to making at least an error of $y_{\mathcal{C}}$ to convert the $\mathcal{C}^{th}$ response to a semi-constrained dimension. Once the $\mathcal{C}$th dimension is considered to be semi-constrained, some of the projections that enter in the y-critical calculations will change:

$$\begin{aligned} e_y &\to \frac{\sum_{\mu=1}^{\mathcal{C}-1} y_\mu e_\mu}{\sqrt{\sum_{\mu=1}^{\mathcal{C}-1} y_\mu^2}} \\ e_{s*} &\to \sqrt{\sum_{\mu\in A_-} e_\mu^2 + \Theta(\mathrm{Sgn}(e_y)e_{\mathcal{C}})e_{\mathcal{C}}^2}. \end{aligned} \tag{C.16}$$

To check whether we have a certain prediction, the new bound on y-critical has to be satisfied along with the original bound that was obtained by treating the $\mathcal{C}^{th}$ dimension as constrained.

It is not hard to see how this process should be continued as one has more than one topological transition within the allowed error. Since we know the precise sequence of topological transitions, all the sequential certainty-conditions can in principle be obtained. A synapse is certain if all of its certainty-conditions are satisfied.

Comparing predictions from linear and nonlinear models: To assess the effects of nonlinearity it is useful to compare the predictions for certain-synapses between the linear and nonlinear theory. In a linear theory, there are no semi-constrained directions, and therefore, a lower bound, leading order and upper bound on y-critical can be obtained from (C.4), (C.15) and (C.11) respectively by setting $e_{s*} = 0$:

$$y_{\texttt{cr,lin,min}} = W\sqrt{\frac{e_u^2}{e_y^2+e_u^2}} + \epsilon \ , \tag{C.17}$$

$$y_{\texttt{cr,lin,lead}} = W\sqrt{\frac{e_u^2}{e_y^2+e_u^2}} + \epsilon\sqrt{1+\frac{e_y^2(e_p^2-e_y^2)}{(e_y^2+e_u^2)^2}} \ , \tag{C.18}$$

$$y_{\texttt{cr,lin,max}} = \sqrt{W^2\left(\frac{e_u^2}{e_y^2+e_u^2}\right) + \epsilon^2\left(1+\frac{e_y^2(e_p^2-e_y^2)}{(e_y^2+e_u^2)^2}\right)} + \epsilon\sqrt{1+\frac{e_y^2(e_p^2-e_y^2)}{(e_y^2+e_u^2)^2}} \ . \tag{C.19}$$

We note that since all these quantities are increasing function of e_{s*}, the linear values are always less than or equal to the nonlinear counterparts. In Fig. 8, we show a comparison between the upper bound on y-critical obtained in the linear and the nonlinear theories.

New sources of corrections in recurrent neural networks:

It is clear that RNNs inherit error corrections to y-critical that were already present in the feedforward case. There are two additional sources of error that one could consider as one moves from feedforward to recurrent networks. However, our numerical simulations of recurrent networks suggest that these are sometimes small effects, and we leave their systematic study for the future.

Firstly, we could account for the fact that the ϵ_μ-directions themselves can change. This is because the inputs driving any given driven neuron can no longer be assumed to be fixed at $z_{\mu m}$ if the other driven neurons suffer from noise. However, these activity patterns define the ϵ_μ-directions and ω_μ-coordinates. Allowing noise in input neurons would lead to similar corrections.

Secondly, the total error in (20) may be unevenly distributed across the driven neurons. If the total squared error summed over all responses and neurons is $\epsilon_{\texttt{tot}}^2$, then on average, the root mean square error associated with each driven neuron is $\epsilon_{\texttt{tot}}/\sqrt{\mathcal{D}}$. We can thus hope that a substitution of $\epsilon = \epsilon_{\texttt{tot}}/\sqrt{\mathcal{D}}$ in the various y-critical formulas will provide a good approximation. However, it's also possible that a few neurons will incur most of the error (up to $\epsilon_{\texttt{tot}}$), potentially leading to violation of the certainty conditions computed from the root mean square error over neurons.

D. Low Dimensional Numerical Tests

Because the recurrent network solution space separates into several feedforward solution spaces at zero error, we numerically treated the driven neurons in the RNN of Fig. 4A sequentially. Here, we first detail the analysis of Fig. 4B for the driven neuron receiving feedforward drive, and we then detail the analysis of Fig. 4C for the driven neuron with self-coupling.

Feedforward Analysis: To test (13, A.47), and especially the dependence of $y_{\mathtt{cr}}$ on the various parameters of the response data, viz. θ, ϕ, α and γ, we decided to perform low dimensional simulations where we could vary the data matrix in a controlled way and compare simulation results for $y_{\mathtt{cr}}$. Below, we consider an $\mathcal{N} = 3$, feedforward network with $\mathcal{C} = \mathcal{S} = \mathcal{U} = 1$. We assumed the response of the input neurons to be given by the following form [7]:

$$Z_{\mu m} = \begin{pmatrix} \cos\psi' & -\sin\psi' & 0 \\ \sin\psi'\cos\chi' & \cos\psi'\cos\chi' & -\sin\chi' \\ \sin\psi'\sin\chi' & \cos\psi'\sin\chi' & \cos\chi' \end{pmatrix} , \tag{D.1}$$

where $\mu = 1, 2$ correspond to the observed responses of three input neurons, and the $\mu = 3$ row is an orthonormal extension of the data matrix as defined previously. We supposed that the driven neuron response vector is given by

$$\vec{y} = \begin{pmatrix} y \\ 0 \\ 0 \end{pmatrix} , \tag{D.2}$$

and focused on the certainty condition for w_1, for which the ψ', χ' variables in the response data (D.1) can be easily related to θ and ϕ projection angles depicted in Fig. 3C. Now, (A.47) can be simplified for different cases as follows: If, $\psi', \chi' \in [0, \pi/2]$, then, $\theta = \psi',\ \phi = \chi',\ \alpha = \gamma = 0$, and the critical value of y for which we can be certain of the sign of w_1 is given by

$$y_{\mathtt{cr}} = W \sin\psi' = W \sin\theta \ . \tag{D.3}$$

The θ-dependence of $y_{\mathtt{cr}}$ is depicted in Fig. 4C, the pink curve, where we have set $W = 1$. There is however, no ϕ dependence as can be seen in the light green curve being flat in the same figure. When $\chi' \in [\pi/2, \pi]$, the component of $\widehat{e}_1$ along the semi-constrained direction is negative, and therefore $\theta = \psi',\ \phi = \pi - \chi',\ \alpha = 0$ and $\cos\gamma = 0$. In this case, $y_{\mathtt{cr}}$ depends both on θ (purple curve) and ϕ (dark green),

$$y_{\mathtt{cr}} = W \frac{\sin\theta \sin\phi}{\sqrt{\cos^2\theta + \sin^2\theta \sin^2\phi}} \ , \tag{D.4}$$

as can be seen in Fig. 4C.

To test the analytic dependence, we wanted to simulate solutions without biasing ourselves by the particular search algorithm used to find solutions. Accordingly, to find solutions to the fixed point equations (A.2) with very small error ($\mathcal{E} < \epsilon = 0.01$) we performed a random screen where each weight was chosen randomly from a uniform distribution between -1 and $+1$. For feedforward circuits, given the synaptic weights, one can obtain the fixed point responses of the target neuron by direct substitution of the known input responses in (A.2) and then compare these simulated target responses with the known target responses. We varied ψ', χ' in the response data (D.1) systematically in steps of $\Delta\psi' = \Delta\chi' = 6^\circ$ [8]. And, for a given choice of ψ', χ', we systematically varied y between 0 and 1, in intervals of, $\Delta y = 0.01$. For each value of ψ', χ' and y, we obtained $\sim \mathcal{O}(10^2 - 10^4)$ solutions [9] satisfying the error and the biological bound (A.11) from five to ten million different trial weight vectors. We then identified the maximal value of y for which the solutions had both positive and negative w_1's. This simulation point should lie beneath the theoretical $y_{\mathtt{cr}}$ if no error is allowed. However, since the error is small but nonzero, occasionally the y-criticals determined from simulations did slightly exceed the theoretical value. Also, since we vary y by small amounts $\Delta y = 0.01$, we expected the simulated y-criticals to be discrete but close to the theoretical predictions, which is exactly what we found in Fig. 4B.

Recurrent Analysis: To test our formalism and analytical results in general recurrent networks with self-couplings, we performed low dimensional simulations where the $\mathcal{N} \times \mathcal{N}$ matrix $Z_{\mu m}$ is orthogonal and we focused on the self-couplings. As discussed in Appendix A, when one moves from feedforward to RNNs, the inputs can no longer be

---

[7] The first two rows of the Z matrix here can be identified with the x matrix in (19) with the following identifications: the first and second rows are interchanged, $\psi' = \psi$, $\chi' = \chi - \pi/2$, the first, second and third columns of Z are identified with second, third and first columns of x respectively. Thus, w_{y_1,x_2} in the main text is w_1 here.

[8] Since here we were primarily interested in the zero-error result, we restricted ourselves to a range of ψ', χ' where no topological transitions can occur due to the small but finite error we had to allow for numerical simulations.

[9] The number of solutions varied between 200 and 40,000 depending primarily on the value of y, the higher the value, typically the more difficult it was to find solutions.

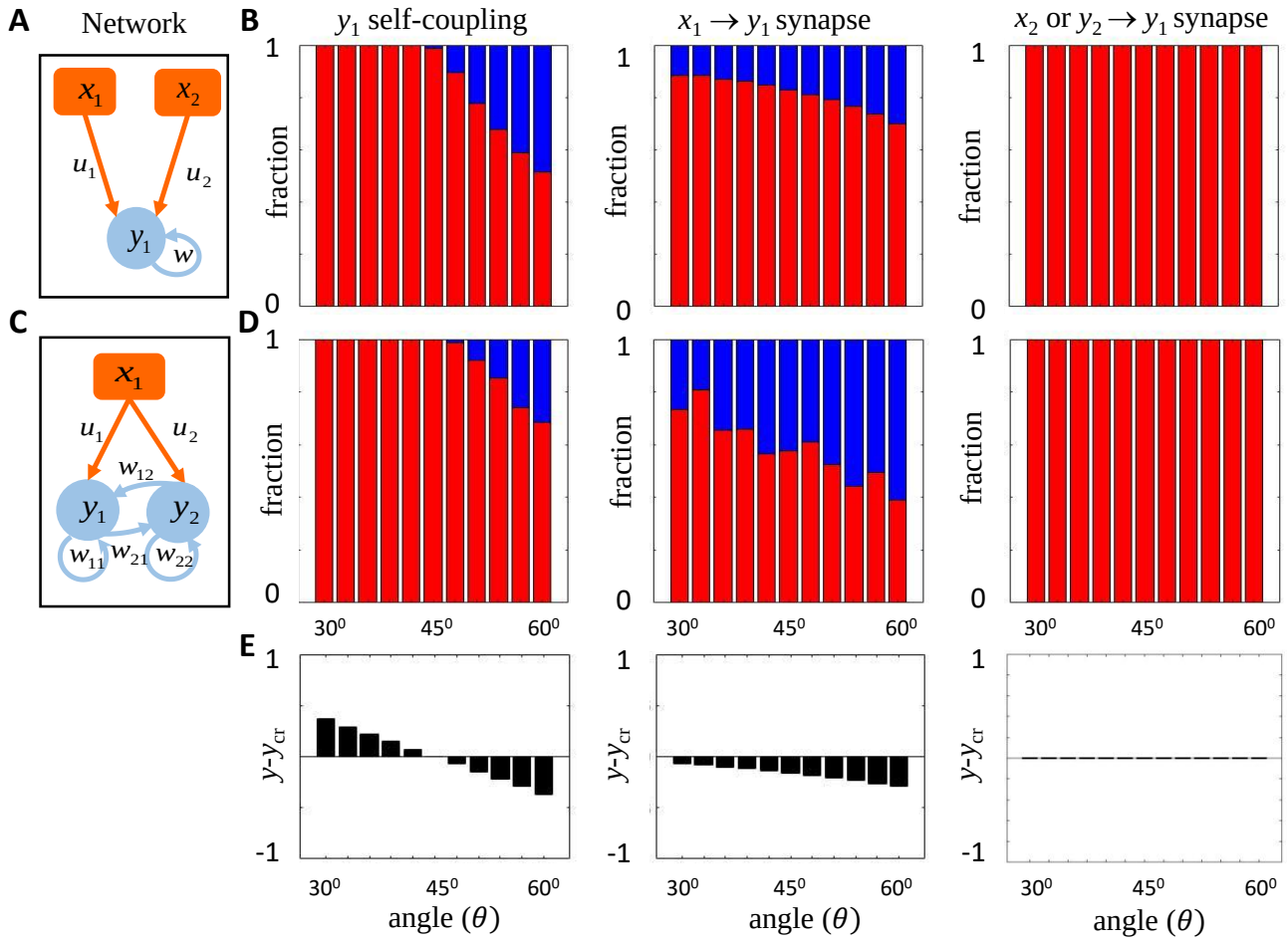

FIG. 9: **Comparing simulation and theoretical results in $\mathcal{N} = 3$ recurrent network.** (**A**) A simple $\mathcal{N} = 3$ RNN with one driven and two input neurons. Note that the y_1 neuron shown here maps onto the y_3 neuron in Fig. 4A by interpreting the x_1 and x_2 neurons shown here as the x_3 and y_2 neurons shown in Fig. 4A. (**B**) Bar graphs depicting the fraction of positive (red) and negative (blue) weights from the network depicted in (A). (**C**) Another $\mathcal{N} = 3$ RNN, this time with two driven and one input neuron. (**D**) Bar graphs depicting the fraction of positive (red) and negative (blue) weights from the network depicted in (**C**). (**E**) The black bars depict $y - y_{\rm cr}$ for the corresponding synapses.

considered independent from the driven responses. We considered an $\mathcal{N} = 3$ RNN with one driven and two input neurons, see Fig. 9A, where the response data was given by

$$Z_{\mu m} = \begin{pmatrix} \cos\chi & \sin\psi\sin\chi & \cos\psi\sin\chi \\ 0 & \cos\psi & -\sin\psi \\ -\sin\chi & \sin\psi\cos\chi & \cos\psi\cos\chi \end{pmatrix} , \tag{D.5}$$

which is essentially a reordering of (D.1). We chose the first column of (D.5) to correspond to the responses of a driven neuron, but it also provides a self-input. The two input neuronal responses are given by the second and third columns. For the purpose of numerical testing we assumed $\psi, \chi \in [0, \pi/2]$, which in particular ensured that the driven neuronal response was non-negative. There is one self-coupling, say w, and two feedforward couplings, say u_1, u_2, coming from the two input neurons, see Fig. 9A. (D.5) corresponds to a case where $\mathcal{P} = 2$, $\mathcal{C} = 1$, and $\mu = 1, 2$ correspond to the constrained and semi-constrained response patterns respectively. $\mu = 3$ row represents the unconstrained pattern. For this response structure, one can calculate $y_{\rm cr}$'s for all the three couplings from (A.48):

$$\begin{aligned} y_{{\rm cr},w} &= W\sin\chi \\ y_{{\rm cr},u_1} &= W\sqrt{\cos^2\psi + \cos^2\chi\sin^2\psi} \\ y_{{\rm cr},u_2} &= W\cos\chi \end{aligned} \tag{D.6}$$

To be certain of the synapse signs we have to make sure that $y_{\rm cr}$'s are smaller than the magnitude of $\vec{y}$,

$$y = \cos\chi \ . \tag{D.7}$$

The same exact response matrix (D.5) can also be used to consider an $\mathcal{N} = 3$ network with the first two columns encoding the responses of the two recurrent neurons, y_1 and y_2, respectively, that are being driven by a single input neuron (third column), Fig. 9C.

We set $W = 1$. Then we see that the condition for certainty is always *just* not satisfied for u_2. Here this implies that all solutions have $u_2 \geq 0$. This is because it is a very special case where $\cos\gamma = e_{s*} = 0$ and $\widehat{y}_\perp$ in (A.51) is aligned with $\widehat{u}$, so that both the inequalities in (A.53) turn into equalities. Next, basic trigonometric manipulations tell us that the certainty condition can never be satisfied for u_1, and the self coupling become certain only if

$$\cos\chi > \sin\chi \ , \ \text{or,} \ \ \chi < 45^\circ \ . \tag{D.8}$$

This is depicted in Figs. 9B, 4C, and 9D. In each of these plots we tracked the fraction of positive and negative signs in the synapses, as we varied $\chi = \theta$. In particular, we see that u_1 always had mixed signs, w had a unique sign as long as $\theta < 45^\circ$, and u_2 was non-negative [10].

To find solutions for these simple RNNs, we fixed χ, ψ and then performed screens with random weights, selected in the same manner as the feedforward simulations discussed earlier. For each set of weights, and for each $\mu = 1, 2$, we obtained the late time values of y by solving the time evolution equation (Eq. (1) with $\tau_i = 20$ms) using Euler's method starting with initial conditions $y_i(0) = y_{\mu i}$, for $\mu = 1, 2$. We used a time step of $\Delta t = 0.2$ ms. The $\widetilde{y}_\mu$'s obtained from the simulation at late times, $t \sim 600$ ms, were then compared with y_μ to obtain $\mathcal{E}$. If the weights satisfied, $\mathcal{E} < 0.05\sqrt{\mathcal{D}}$, and the biological bound (A.11), then we considered the weights as solutions and checked the sign of the couplings. For every value of ψ, χ, we had at least 50 solutions [11] to test the certainty predictions.

E. High Dimensional Numerics

Feedforward Simulations: For feedforward networks we performed two different types of simulations. In the first type of simulation, our goal was to show the effect of the different corrections that we encounter when we have non-zero error. For this purpose we needed to explore the parameter space comprehensively. Accordingly, we simulated moderate size feedforward networks, and Fig. 6B shows results with $\mathcal{N} = 6$, $\mathcal{P} = 5$ and $\mathcal{C} = 2$. We constructed 101 orthonormal input response patterns by first generating an $\mathcal{N} \times \mathcal{N}$ antisymmetric matrix, G, where each of its entries was randomly selected from a uniform distribution between -1 and $+1$, and then we antisymmetrized the matrix: $G \to (G - G^T)/2$. The orthogonal response matrix was then obtained via exponentiation: $Z = e^G$. We focused our attention on a single synaptic weight. For every input response pattern, we also chose a random direction of the driven response, $\hat{y}$, with $\mathcal{C}$ nonzero responses. This was done by first randomly selecting numbers between 0 and 1 for every nonzero response. To examine the effects of topological transitions, we next set one nonzero response of the target neuron to a small value, 0.1ϵ. $\widehat{y}$ was then obtained by dividing the response vector by its norm. Also, for simplicity we only kept those $\widehat{y}$'s whose other entries were large enough to not show a transition. This was done by ensuring that all the other entries of $\vec{y}$ were greater than the theoretical lower bound for y-critical [12]. This way, typically one and only one constrained direction could become semi-constrained when we allow solutions with errors $\lesssim \epsilon$.

We next wanted to find a large number of solutions that would widely sample the solution space with non-zero error. In Fig. 6B we show an example with $\mathcal{E} < \epsilon = 0.1$. Since scanning a 6-dimensional synaptic weight-space randomly is not numerically efficient, we applied gradient descent learning [13] to obtain about 10^5 solutions to the fixed point equations (A.2). The initial weights were chosen randomly from a uniform distribution between -1 and 1. The weight vector was then rescaled to have a norm between 0 and 1, chosen randomly. We varied y, the norm of the target response vector, $\vec{y} = y\widehat{y}$, systematically by $\Delta y = 0.01$ in a manner similar to the low dimensional simulations (Appendix D). This way we found a maximum value of y in our simulations that still admitted mixed signs for the

---

[10] As we explained before, $y_{{\rm cr},u_2} = y$ for all values of θ. Hence the certainty condition is not satisfied because u_2 may be zero. The fact that u_2 can vanish is not discernible from our simulations because the weight magnitudes were generated randomly, and weights where $u_2 = 0$ comprise a zero measure set.

[11] The number of solutions varied between 250 and 20,000 for the $\mathcal{D} = 1$ simulation shown in Fig. 4C, between 10^4 and 10^5 for the $\mathcal{D} = 1$ simulation in Fig. 9B and between 56 and 110 for the $\mathcal{D} = 2$ simulation in Fig. 9D. Note that for the $\mathcal{D} = 2$ simulation we have a six dimensional weight-space which makes it a lot harder to find solutions through random scanning. Also, for this latter case we only checked that the biological constraint is satisfied by the incoming weights to y_1.

[12] The true y-critical is larger and could thus still induce a transition. However, this procedures empirically seemed sufficient for our numerical testing of transition effects.

[13] Learning rate = 0.02.

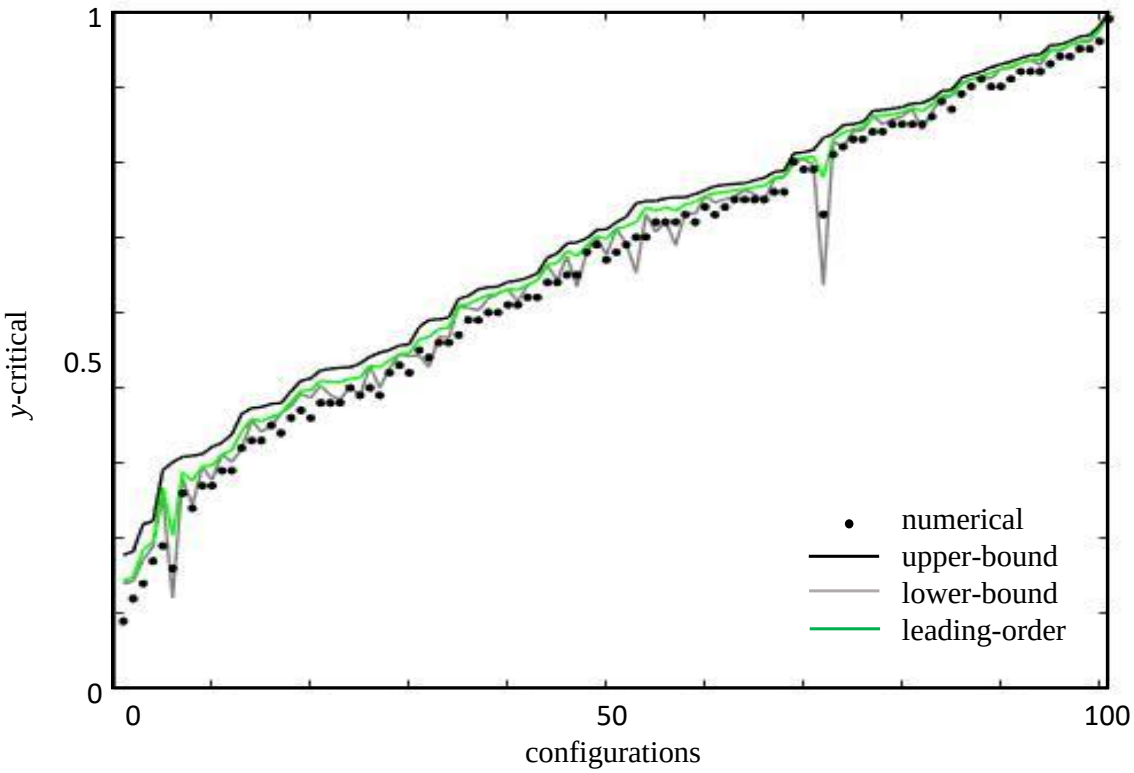

FIG. 10: **Testing bounds on y-critical for solutions with error.** We show the same 101 random configurations of input-output activity as Fig. 6B. The bold black, green, and gray curves represent the upper bound (C.11), leading-order (C.15), and lower bound y-critical values (C.4), respectively. The black dots correspond to the maximum value of y in our simulations that resulted in mixed signs for the synaptic weights under consideration.

synapse under consideration. In Figs. 6B and 10, these are depicted by black dots. As one can see in Fig. 10, these dots always lie below, and closely track, the y-critical curve that accounts for the leading order error correction and the topological transition (green). Most of these dots also lie below the lower bound y-critical curve but some certainly do lie above (gray). As expected, the upper bound curve is consistently above the dots and other curves (black).

In the second set of simulations, depicted in Fig. 6C, our goal was to see how our predictions for certain synapses fare as we scale the size of the network. Concurrently, we also wanted to assess whether the number of certain-synapses scale with the network size as our heuristic argument in Appendix B suggested. Accordingly, we considered feedforward networks with $\epsilon = 0.001\sqrt{\mathcal{P}}$, and $\mathcal{C}$ and $\mathcal{P}$ scaling with $\mathcal{N} = \mathcal{I}$: here the number of input neurons is also the number of synapses onto the target neuron. The orthonormal input response patterns were generated randomly as in the previous simulation. The target response vector was also chosen as before, except that its norm was fixed such that 50% of vectors in the unit sphere have shorter lengths (B.5).

For a given value of $\mathcal{N}$, we considered a single input-output configuration and found a single solution with gradient descent learning [14]. The bold curves in light brown and light purple plot the number of certain synapses, as predicted by our formalism, as we increase $\mathcal{N}$ in the network. The purple curve corresponds to the case when $\mathcal{N} = 2\mathcal{P} = 4\mathcal{C}$ and the brown curve when $\mathcal{N} = \mathcal{P} = 4\mathcal{C}$. The circled dots represent the number of synapses that had the correctly predicted sign in the simulations. As one can see, our predictions were always accurate. Also, the number of certainty dimensions did seem to scale with $\mathcal{N}$. However, the best-fit slopes were $0.16(\pm 0.01)$ and $0.07(\pm 0.01)$ at 95% confidence level, somewhat less than the zero error theoretical estimates of 0.28 and 0.18, respectively.

Recurrent Simulations: Finally, we considered an RNN to test the reliability of our formalism in ascertaining the solution space when we have nonzero error, as well as its ability in predicting the sign of synapses whenever the

[14] Learning rate = 0.005.

conservative certainty condition $y > y_{\mathtt{cr,max}}$ (C.11) is met. For this purpose, we considered networks without the self-coupling terms and obtained solutions using gradient descent, as we will describe in the next paragraph. For these simulations, we focused on couplings onto a given, say the $\mathcal{D}^{th}$, driven neuron, and ensured that the neuronal input patterns it receives are orthogonal. To achieve the latter, we started by choosing a $\mathcal{P}\times\mathcal{N}$ dimensional matrix containing the responses of $(\mathcal{D}-1)$ driven and $\mathcal{I}\geq\mathcal{P}-1$, input neurons. We made sure that the responses of the driven neurons were all non-negative, as the threshold nonlinearity dictates, by choosing them to lie randomly between 0 and 1. To mimic a sparse response pattern, we randomly set 50% of the driven responses to 0. The feedforward inputs, on the other hand, were randomly selected between -1 and 1. Next we orthogonalized the input responses to the target neuron as follows. For each row, $\nu=2\ldots\mathcal{P}$, in a sequential order we performed the following operations:

- We started by defining a $\nu-1$ dimensional square matrix, x':

$$x'_{\mu m}\equiv x_{\mu m}\text{ for }\mu,m=1\ldots(\nu-1). \tag{E.1}$$

- We next changed the first $\nu-1$ elements of the ν^{th} row of x:

$$x_{\nu m}\equiv-\sum_{\mu=1}^{\nu-1}x'^{-1}_{m\mu}\left(\sum_{i=1}^{\mathcal{D}-1}y_{\mu i}y_{\nu i}\right). \tag{E.2}$$

  The other elements of the ν^{th} row of x were left unchanged.

- Finally, we rescale all the elements of the ν^{th} row of x and y, except the $\mathcal{D}^{th}$ y-column that contains the responses of the target neuron:

$$z_{\nu m}\to\frac{z_{\nu m}}{\sqrt{\sum_{n\neq\mathcal{D}}z^2_{\nu n}}}\;\forall\;m\neq\mathcal{D}. \tag{E.3}$$

Essentially the algorithm uses the responses of the input neurons to the νth stimulus to ensure that the full νth response pattern involving both the driven and input neurons is orthogonal to all $\mu\leq\nu-1$ patterns. Separately we randomly chose responses of the $\mathcal{D}^{th}$ driven neuron, whose in-synapses were our main interest. These responses were chosen in the same manner as the target neuron response in the feedforward simulation investigating scaling behavior.

Having orthogonalized the input patterns to the $\mathcal{D}^{th}$ driven neuron, we proceeded to obtain solutions with errors, $0<\mathcal{E}<0.25\sqrt{\mathcal{D}}$. To find solutions for the RNNs in Figs. 6D-E, we first used the modified loss function,

$$\bar{\mathcal{E}}^2\equiv\sum_{i=1}^{\mathcal{D}}\sum_{\mu=1}^{\mathcal{P}}\left(\widetilde{y}_{\mu i}-y_{\mu i}\right)^2\equiv\sum_{i=1}^{\mathcal{D}}\bar{\mathcal{E}}_i^2\;,\;\text{where}\;\;\widetilde{y}_{\mu i}=\left(\sum_{m=1}^{\mathcal{N}}z_{\mu m}w_{im}\right)\Theta\left(\sum_{m=1}^{\mathcal{N}}z_{\mu m}w_{im}\right)\;, \tag{E.4}$$

to perform gradient descent. Since the responses of the driven neurons can vary for nonzero errors, the two loss functions, $\mathcal{E}$ and $\bar{\mathcal{E}}$, differ. However, it is numerically a lot quicker to obtain solutions via gradient descent with $\bar{\mathcal{E}}$ as compared to using back-propagation through time to consider the entire time evolution of the network. Thus, the strategy we adopted to find solutions with, say $\mathcal{E}\lesssim\epsilon$, was to first find weights satisfying $\bar{\mathcal{E}}\lesssim\bar{\epsilon}=\epsilon/10$. Also, the gradient descent [15] was done in two stages. In the first stage we minimized the error associated with each individual driven neuron, $\bar{\mathcal{E}}_i$. Once each of these errors were less than $\bar{\epsilon}$, we performed a second stage of gradient descent to minimize $\bar{\mathcal{E}}$ down to $\bar{\epsilon}$. Next, we obtained the late time values of y_i's by solving the time evolution equations (Eq. (1) with $\tau_i=20$ ms) using Euler's method with step time $\Delta t=0.2$ ms for the weights obtained via gradient descent and starting with initial conditions $y_i(0)=y_{\mu i},\forall\mu,i$. The $\widetilde{y}_{\mu i}$'s obtained at late times, $t\sim600$ ms, this way were compared with $y_{\mu i}$ to obtain $\mathcal{E}$. If the weights satisfied the error and the biological bounds [16] (A.11), then we considered the weights to be solutions and checked the sign of the couplings.

In Figs. 6D-E, we show results of a simulation for a $\mathcal{N}=10,\mathcal{D}=4,\mathcal{I}=7,\mathcal{P}=8$, and $\mathcal{C}=3$ network where the norm of $\vec{y}$ was fixed to 0.77, which is the median value of y for the given values of $\mathcal{N},\mathcal{P}$ and $\mathcal{C}$ (B.5). For these simulations, and for a given response pattern obtained as prescribed above, we found ~4500 weight matrices that

---

[15] For both stages a learning rate = 0.004 was used.

[16] We only checked that the target weights satisfied the weight bound, as that's what matters for the certainty conditions. Since we initialized weights amongst the non-target neurons to be between -7 and 7, it's likely that other components of the weight matrix were large.

satisfied the weight bound (A.11) for the input synapses of the $\mathcal{D}^{th}$ driven neuron, and whose overall error satisfied $0.017\sqrt{\mathcal{D}} \lesssim \mathcal{E} \lesssim 0.25\sqrt{\mathcal{D}}$. Fig. 6D shows the projection of the corresponding solution space along two ω-directions, one of which is constrained and the other semi-constrained. While the extension of the solution space along the negative semi-constrained direction is clearly discernible one does also see a variation in error (color). Firstly, this is because the error is a sum of squared differences between simulated and observed responses for all stimulus conditions, but we are only seeing a projection along two. The error for other stimulus conditions will add variation in the error, and this is going to be true for a feedforward network as well. Secondly, because we have an RNN with more than one driven neuron, and we are only considering the pooled error over all the neurons, this leads to an additional variation in error values. Finally, due to recurrence we expect the semi-constrained directions to no longer be perfectly cylindrical, as the inputs to the $\mathcal{D}^{th}$ target neuron coming from the other driven neurons can themselves vary. However, in spite of all the sources of variation, the geometric structure of the solution space seems to be conforming to our theoretical prediction rather well.

Finally, we wanted to test the reliability of our formalism in predicting the sign of synapses whenever the conservative certainty condition was met. In Fig. 6E, the cyan and orange curves respectively plot how the number of certain synapses predicted by the y-critical upper bounds in linear and nonlinear models change as one increases the root mean squared error over neurons, $\epsilon/\sqrt{\mathcal{D}}$. The circled dots represent the numbers of synapses whose sign were predicted correctly by the two different theoretical models, while the crosses represent the simulations which resulted in at least one incorrect prediction. While there are several blue crosses indicating incorrect linear predictions, there are no red crosses indicating contradictions with the nonlinear model.

[1] Watson JD, Crick FHC (1953) Molecular structure of nucleic acids. *Nature* 171:737-738.

[2] Milo R et al. (2002) Network motifs: simple building blocks of complex networks. *Science* 298:824-827.

[3] Hunter P, Nielsen P (2005) A strategy for integrative computational physiology. *Physiology* 20:316-325.

[4] Seung HS (2009) Reading the book of memory: sparse sampling versus dense mapping of connectomes. *Neuron* 62:17-29.

[5] Bargmann CI, Marder E (2013) From the connectome to brain function. *Nature Methods* 10:483.

[6] Bock DD et al. (2011) Network anatomy and in vivo physiology of visual cortical neurons. *Nature* 471:177.

[7] Varshney LR, Chen BL, Paniagua E, Hall DH, Chklovskii DB (2011) Structural properties of the *Caenhabditis elegans* neuronal network. *PLoS Comput Biol* 7:e1001066.

[8] Ahrens MB et al. (2012) Brain-wide neuronal dynamics during motor adaptation in zebrafish. *Nature* 485:471-477.

[9] Schrodel T, Prevedel R, Aumayr K, Zimmer M, Vaziri A (2013) Brain-wide 3D imaging of neuronal activity in *Caenorhabditis elegans* with sculpted light. *Nat Methods* 10:1013-1020.

[10] Ohyama T et al. (2015) A multilevel multimodal circuit enhances action selection in *Drosophila*. *Nature* 520:633-639.

[11] Lemon WC et al. (2015) Whole-central nervous system functional imaging in larval *Drosophila*. *Nat Commun* 6:7924.

[12] Naumann EA et al. (2016) From whole-brain data to functional circuit models: the zebrafish optomotor response. *Cell* 167:947-960.

[13] Hildebrand DGC et al. (2017) Whole-brain serial-section electron microscopy in larval zebrafish. *Nature* 545:345.

[14] Scheffer LK et al. (2020) A connectome and analysis of the adult *Drosophila* central brain. *BioRxiv* 10.1101/2020.01.21.911859.

[15] Ben-Yishai R, Bar-Or RL, Sompolinsky H (1995) Theory of orientation tuning in visual cortex. *Proc Natl Acad Sci USA* 92:3844-3848.

[16] Skaggs WE, Knierem JJ, Kudrimoti HS, McNaughton BL (1995) A model of the neural basis of the rat's sense of direction. *Adv Neural Inf Process Syst* 7:173-180.

[17] Kim SS, Rouault H, Druckmann S, Jayaraman V (2017). Ring attractor dynamics in the Drosophila central brain. *Science* 356:849-853.

[18] Turner-Evans DB et al. (2020) The neuroanatomical ultrastructure and function of a biological ring attractor. *Neuron* 108:1-19.

[19] Kim JS et al. (2014) Space-time wiring specificity supports direction selectivity in the retina. *Nature* 509:331-336.

[20] Kornfeld J et al. (2017) EM connectomics reveals axonal target variation in a sequence-generating network. *eLife* 6:e24364.

[21] Wanner AA, Friedrich RW (2020) Whitening of odor representations by the wiring diagram of the olfactory bulb. *Nat Neuro* 23:433-442.

[22] Marder E, Taylor AL (2011) Multiple models to capture the variability in biological neurons and networks. *Nat Neurosci* 14:133-138.

[23] Friston K, Harrison L, Penny W (2003) Dynamic causal modeling. *Neuroimage* 19:1273–1302.

[24] Schneidman E, Berry MJ, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440:1007-1012.

[25] Pillow JW et al. (2008) Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature* 454:995-999.

[26] Huang H, Ding M (2016) Linking functional connectivity and structural connectivity quantitatively: a comparison of methods. *Brain Connectivity* 6:99-108.

[27] Tschopp FD, Reiser MB, Turaga SC (2018) A connectome based hexagonal lattice convolutional network model of the Drosophila visual system. *arXiv* 1806.04793.

[28] Zarin AA, Mark B, Cardona A, Litwin-Kumar A, Doe CQ (2019) A multilayer circuit architecture for the gener-

ation of distinct locomotor behaviors in *Drosophila*. *eLife* 8:e51781.

[29] Litwin-Kumar A, Turaga SC (2019) Constraining computational models using electron microscopy wiring diagrams. *Curr Opin Neurobiol* 58:94-100.

[30] Prinz AA, Bucher D, Marder E (2004) Similar network activity from disparate circuit parameters. *Nat Neurosci* 7:1345.

[31] Fisher D, Olasagasti I, Tank DW, Aksay ER, Goldman MS (2013) A modeling framework for deriving the structural and functional architecture of a short-term memory microcircuit. *Neuron* 79:987-1000.

[32] Goaillard J-M, Taylor AL, Schulz DJ, Marder E (2009) Functional consequences of animal-to-animal variation in circuit parameters. *Nat Neuro* 12:1424-1430.

[33] Baldi P, Hornik K (1989) Neural networks and principal component analysis: learning from examples without local minima. *Neural Networks* 2:53-58.

[34] Dauphin YN et al. (2014) Identifying and attacking the saddle point problem in high-dimensional optimization. In *Advances in Neural Information Processing Systems 2014*.

[35] Kawaguchi K (2016) Deep learning without poor local minima. In *Advances in Neural Information Processing Systems 2016*.

[36] Machta BB, Chachra R, Transtrum MK, Sethna JP (2013). Parameter space compression underlies emergent theories and predictive models. *Science* 342:604-607.

[37] Transtrum MK et al. (2015) Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *J Chem Phys* 143(1):07B201.

[38] O'Leary T, Sutton AC, Marder E (2015) Computational models in the age of large datasets. *Curr Opin Neurobiol* 32:87-94.

[39] Abbott LF, Regehr WG (2004) Synaptic computation. *Nature* 43:796-803.

[40] Spruston N (2008) Pyramidal neurons: dendritic structure and synaptic integration. *Nat Rev Neurosci* 9:206-221.

[41] Zeng H, Sanes JR (2017) Neuronal cell-type classification: challenges, opportunities and the path forward. *Nat Rev Neurosci* 18:530-546.

[42] Grant SGN (2018) Synapse molecular complexity and the plasticity behaviour problem. *Brain and Neuroscience Advances* 2:1-7.

[43] Curto C, Morrison K (2019) Relating network connectivity to dynamics: opportunities and challenges for theoretical neuroscience. *Curr Opin Neurobiol* 58:11-20.

[44] Billeh YN et al. (2020) Systematic integration of structural and functional data into multi-scale models of mouse primary visual cortex. *Neuron* 106:388-403.

[45] Almog M, Korngreen A (2016) Is realistic neuronal modeling realistic? *J Neurophy* 116:2180-2209.

[46] Bittner SR et al. (2019) Interrogating theoretical models of neural computation with deep inference. *bioRxiv* 10.1101/837567.

[47] Goncalves et al. (2020) Training deep neural density estimators to identify mechanistic models of neural dynamics. *bioRxiv* 10.1101/838383.

[48] Treves A, Rolls ET (1991) What determines the capacity of autoassociative memories in the brain? *Network-Comp Neural* 2:371-397.

[49] Salinas E, Abbott LF (1996) A model of multiplicative neural responses in parietal cortex. *Proc Natl Acad Sci USA* 93:11956–11961.

[50] Hahnloser RLT (1998) On the piecewise analysis of networks of linear threshold neurons. *Neural Networks* 11:691–697.

[51] Hahnloser RH, Seung HS, Slotine JJ (2003) Permitted and forbidden sets in symmetric threshold-linear networks. *Neural Comput* 15:621-38.

[52] Morrison K, Degeratu A, Itskov V, Curto C (2016) Diversity of emergent dynamics in competitive threshold-linear networks: a preliminary report. *arXiv*:1605.04463.

[53] Curto C, Geneson J, Morrison K (2019) Fixed points of competitive threshold-linear networks. *Neural Comput* 31:94-155.

[54] Marder E (2012) Neuromodulation of neuronal circuits: back to the future. *Neuron* 76:1-11.

[55] Mu Y et al. (2019) Glia accumulate evidence that actins are futile and suppress unseccessful behavior. *Cell* 178:27-43.

[56] Aitchison et al. (2017) Model-based Bayesian inference of neural activity and connectivity from all-optical interrogation of a neural circuit. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA.

[57] Grienberger C, Konnerth A (2012) Imaging calcium in neurons. *Neuron* 73:862-885.

[58] Wilt BA, Fitzgerald JE, Schnitzer MJ (2013) Photon shot noise limits on optical detection of neuronal spikes and estimation of spike timing. *Biophys J* 104:51-62.

[59] Theis L et al. (2016) Benchmarking spike rate inference in population calcium imaging. *Neuron* 90:471–482.

[60] Logothetis NK, Pfeuffer J (2004) On the nature of the BOLD fMRI contrast mechanism. *Magn Reson Imaging* 22:1517-1531.

[61] Bartolo MJ et al. (2011) Stimulus-induced dissociation of neuronal firing rates and local field potential gamma power and its relationship to the resonance blood oxygen level-dependent signal in macaque primary visual cortex. *Eur J Neurosci* 34:1857–1870.

[62] Heinzle J, Koopmans PJ, den Ouden HEM, Raman S, Stephan KE (2016) A hemodynamic model for layer BOLD signals. *Neuroimage* 125:556-570.

[63] Nair V, Hinton GE (2010) Rectified linear units improve restricted Boltzmann machines. In *ICML 2010*.

[64] Krizhevsky A (2010) Convolutional deep belief networks on CIFAR-10.

[65] Xu B, Wang N, Chen T, Li M (2015) Empirical evaluation of rectified activations in convolutional network. *arXiv*:1505.00853.

[66] Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. *Nature* 323:533-536.

[67] Frankle J, Carbin M (2018) The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv*:1803.03635.

[68] Zhou H, Lan J, Liu R, Yosinski J (2019) Deconstructing lottery tickets: Zeros, signs, and the supermask. *arXiv*:1905.01067.

[69] Cover TM (1965) Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers* 14:326-334.

[70] Gardner E (1988) The space of interactions in neural network models. *J Phys A: Math Gen* 21:257-270.

[71] Marr D (1969) A theory of cerebellar cortex. *J Physiol*

202:437-470.1.

[72] Olshausen BA, Field DJ (1996) Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Res* 37:3311-3325.

[73] Glorot X, Bordes A, Bengio Y (2011) Deep sparse rectifier neural networks. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS) 2011*, Fort Lauderdale, FL, USA.

[74] Kubo F et al. (2014) Functional architecture of an optic flow-responsive area that drives horizontal eye movements in zebrafish. *Neuron* 81:1344-1359.

[75] Burnstock, G (2004) Cotransmission. *Current Opinion in Pharmacology* 4 (1): 47–52

[76] Chen B.L., D. H. Hall, and D. B. Chklovskii (2006) Wiring optimization can relate neuronal structure and function *PNAS* 103 (12) 4723-4728

[77] Brunel N, Hakim V, Isope P, Nadal J, Barbour B (2004) Optimal information storage and the distribution of synaptic weights: perceptron versus Purkinje cell. *Neuron* 43:745-757.