Proceedings of Machine Learning Research vol 107:476–511, 2020

Gating creates slow modes and controls phase-space complexity in GRUs and LSTMs

Tankut Can [*†]
Initiative for the Theoretical Sciences, CUNY Graduate Center

Kamesh Krishnamurthy [*†]
Joseph Henry Laboratories of Physics and PNI, Princeton University

David J. Schwab
Initiative for the Theoretical Sciences, CUNY Graduate Center

Abstract

Recurrent neural networks (RNNs) are powerful dynamical models for data with complex temporal structure. However, training RNNs has traditionally proved challenging due to exploding or vanishing of gradients. RNN models such as LSTMs and GRUs (and their variants) significantly mitigate these issues associated with training by introducing various types of *gating* units into the architecture. While these gates empirically improve performance, how the addition of gates influences the dynamics and trainability of GRUs and LSTMs is not well understood. Here, we take the perspective of studying randomly initialized LSTMs and GRUs as dynamical systems, and ask how the salient dynamical properties are shaped by the gates. We leverage tools from random matrix theory and mean-field theory to study the state-to-state Jacobians of GRUs and LSTMs. We show that the update gate in the GRU and the forget gate in the LSTM can lead to an accumulation of slow modes in the dynamics. Moreover, the GRU update gate can poise the system at a marginally stable point. The reset gate in the GRU and the output and input gates in the LSTM control the spectral radius of the Jacobian, and the GRU reset gate also modulates the complexity of the landscape of fixed-points. Furthermore, for the GRU we obtain a phase diagram describing the statistical properties of fixed-points. We also provide a preliminary comparison of training performance to the various dynamical regimes realized by varying hyperparameters. Looking to the future, we have introduced a powerful set of techniques which can be adapted to a broad class of RNNs, to study the influence of various architectural choices on dynamics, and potentially motivate the principled discovery of novel architectures.

Keywords: RNN, GRU, LSTM, RMT, MFT

1. Introduction

Recurrent neural networks (RNNs) are a powerful class of dynamical systems that can implement a complex array of transformations between time-varying inputs and target outputs. These networks have proved to be highly effective tools in learning tasks involving data with complex temporal structure. However, RNNs in the form they were initially proposed are challenging to train due to the so-called exploding and vanishing gradients problem Bengio et al. (1994); Hochreiter et al. (2001); Pascanu et al. (2013).

More sophisticated RNN models such as long short-term memory networks (LSTMs) Hochreiter and Schmidhuber (1997) and gated recurrent units (GRUs) Cho et al. (2014) that feature some

[*] TC & KK contributed equally and listed alphabetically

[†] Corresponding authors: tankut.can@gmail.com and kameshk@princeton.edu

form of gating exhibit significantly improved trainability. Empirically, these network models also achieve significant improvements over traditional RNNs in areas such as language modeling Kiros et al. (2015), speech recognition Graves (2013), and neural machine translation Cho et al. (2014); Sutskever et al. (2014); Bahdanau et al. (2014). Thus gating seems to robustly improve the performance of RNNs.

In addition to improved trainability, gating likely has a significant influence on the *dynamics* of LSTMs and GRUs. However, the various architectural choices for the gates are somewhat *ad hoc*, and the precise nature of how the gates influence the dynamics and trainability of the network is not well understood. In this work, we take the approach of theoretically studying LSTMs and GRUs as dynamical systems and characterize how the gates shape the most salient aspects of the dynamics. We use techniques from random matrix theory and mean-field theory to analytically characterize the spectrum of the state-to-state Jacobian, and we study how each gate shapes the features of the spectra. Our formalism properly accounts for the highly recurrent nature of these systems which can lead to very different behavior from the feed-forward case. In the GRU, the two gate types produce very distinct effects: the update gate can lead to a clumping of Jacobian eigenvalues near unity, facilitating a proliferation of long timescales and *marginal stability*, whereas the reset gate influences the spectral radius and controls the topological complexity of phase space. We use the theory to develop a phase diagram for the GRU that characterizes the transition between different dynamical behaviors and the statistical properties of fixed points. For the LSTM, our spectral analysis of the Jacobian reveals the distinct roles of each gate: the forget gate primarily influences the accumulation of eigenvalues near unity, whereas all gates influence the spectral radius, though in differing degree. Finally, we provide a preliminary comparison of training on a sequential task in the various dynamical regimes.

2. Prior work

Recent work Chen et al. (2018) has focused on characterizing the role of gating in conditioning the Jacobian of a minimally gated model under the assumption that the weights are independent from one time step to the other; this untied weight assumption is akin to studying a feed-forward gated network. The focus in that work was more on selecting good parameter initialization values for constraining the second moment of the Jacobian singular values, which improves trainability. This approach of using independent weights at each time step was later extended to LSTMs and GRUs Gilboa et al. (2019). More recently Jordan et al. (2019) have looked at characterising GRUs as dynamical systems, but the focus is on small (two-dimensional) systems. Tallec and Ollivier (2018) have argued that the update gates in GRUs can help with dealing with time warping.

Lee et al. (2018); Schoenholz et al. (2017); Pennington et al. (2017) have used random matrix theory (RMT) to characterize signal propagation in deep feedforward networks at initialization with independent weights in each layer. Seminal work by Derrida and Pomeau (1986) and Sompolinsky et al. (1988) used mean-field theory to study dynamics of randomly connected recurrent networks without gating – i.e. purely additive interactions. Subsequent work Stern et al. (2014); Aljadeff et al. (2015b,a); Marti et al. (2018); Ahmadian et al. (2015) has extended this analysis using RMT to the case when the weight matrix has a correlational structure.

In the remainder of the paper, we first describe the gated networks we consider. Next we formulate the problem of studying the spectrum of the state-to-state Jacobian as a random matrix problem.

We then describe how the gates shape the short-term dynamics for GRUs and LSTMs. In Appendix H, we provide preliminary results on training for a sequential task in the various dynamical regimes.

3. Problem setup

The 'vanilla' RNN has no form of gating, and is described by a discrete time dynamical equation

$$\mathbf{h}_t = \phi\left(U_h \mathbf{h}_{t-1} + W_x \mathbf{x}_t + \mathbf{b}\right), \tag{1}$$

where $U_h \in \mathbb{R}^{N\times N}, W_x \in \mathbb{R}^{N\times N_{in}}, \mathbf{b} \in \mathbb{R}^N$, $\mathbf{x}_t \in \mathbb{R}^{N_{in}}$ is the input and $\mathbf{h}_t \in \mathbb{R}^N$ is the internal (hidden) state of the RNN at the (integer) time step t. The nonlinear activation function ϕ is commonly taken to be $\tanh$, which will be our choice unless otherwise stated. These RNNs have been successful at learning sequential tasks but are notoriously hard to train due to the problem of exploding/vanishing gradients. Recently, LSTMs and GRUs were introduced to mitigate this issue by augmenting the vanilla RNN with gates that control information flow.

For the GRU, in addition to the hidden state variables $\mathbf{h}_t$, there are two additional dynamical gating variables: an update gate $\mathbf{z}_t$ and a reset gate $\mathbf{r}_t$ which both take values $\mathbf{z}_t, \mathbf{r}_t \in (0,1)^N$. The dynamics of the GRU is given by

$$\mathbf{z}_t = \sigma\left(U_z \mathbf{h}_{t-1} + W_z \mathbf{x}_t + \mathbf{b}_z\right), \quad \text{update} \tag{2}$$
$$\mathbf{r}_t = \sigma\left(U_r \mathbf{h}_{t-1} + W_r \mathbf{x}_t + \mathbf{b}_r\right), \quad \text{reset} \tag{3}$$
$$\mathbf{y}_t = U_h(\mathbf{r}_t \odot \mathbf{h}_{t-1}) + W_h \mathbf{x}_t + \mathbf{b}_h, \tag{4}$$
$$\mathbf{h}_t = \mathbf{z}_t \odot \mathbf{h}_{t-1} + (1 - \mathbf{z}_t) \odot \phi(\mathbf{y}_t). \tag{5}$$

where $\odot$ denotes element-wise product.

Next, we consider LSTM networks introduced by Hochreiter and Schmidhuber (1997); Gers et al. (1999). LSTMs have three gates (input $\mathbf{i}_t$, forget $\mathbf{f}_t$ and output $\mathbf{o}_t$), and their state is characterized by a cell state $\mathbf{c}_t \in \mathbb{R}^N$ and the hidden state $\mathbf{h}_t \in \mathbb{R}^N$. The LSTM dynamics is given by

$$\mathbf{f}_t = \sigma(U_f \mathbf{h}_{t-1} + W_f \mathbf{x}_t + \mathbf{b}_f) \quad \text{forget} \qquad \mathbf{i}_t = \sigma(U_i \mathbf{h}_{t-1} + W_i \mathbf{x}_t + \mathbf{b}_i) \quad \text{input} \tag{6}$$
$$\mathbf{o}_t = \sigma(U_o \mathbf{h}_{t-1} + W_o \mathbf{x}_t + \mathbf{b}_o) \quad \text{output} \qquad \mathbf{y}_t = U_h \mathbf{h}_{t-1} + W_h \mathbf{x}_t + \mathbf{b}_h \tag{7}$$
$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \phi(\mathbf{y}_t) \qquad \mathbf{h}_t = \mathbf{o}_t \odot \phi(\mathbf{c}_t). \tag{8}$$

We use the common choice of the sigmoid $\sigma(x) = (1 + e^{-x})^{-1}$ for the gating nonlinearity in both cases. We will study the dynamics of these models upon random parameter initialization – i.e. the elements of the connectivity matrices and the bias vectors are assumed to be gaussian random variables, appropriately scaled by system size $(U_k)_{ij} \sim \mathcal{N}(0, a_k^2/N)$, $(\mathbf{b}_k)_i \sim \mathcal{N}(b_k, v_k)$, for $k \in \{z, r, h\}$ (GRU) and $k \in \{f, o, i, h\}$ (LSTM). In most of what follows, we assume the bias variance $v_k = 0$, unless otherwise explicitly stated.

Before we proceed with the analysis, we note that just from these dynamical equations one can understand simple intuitive features of each gate's effect. For example, in the GRU the update gate (2) controls the amount of *leak* and can thus potentially slow down the mixing of the inputs (as $\mathbf{z}_t \to 1$) due to self-coupling, thereby modulating memory in the network. The forget and input gate (6) would appear to have analogous roles for the LSTM. The GRU reset gate (3) modulates

(column-wise) the strength of the connectivity matrix U_h. One important consequence of this is that the reset gate has the power to change the landscape of dynamical fixed points, which is determined by U_h. The output gate (7) appears to have an analogous role in modulating U_h in the LSTM architecture. We will elaborate on these points below and provide quantitative understanding to these qualitative features.

Finally, we do not directly consider the effects of the input. However, nonzero bias is equivalent to the case in which the input is *constant* in time. We comment on the effects of the bias throughout the text.

Notation We use $\hat{\mathbf{v}}$ to denote the diagonal matrix whose entries are given by the elements of the vector $\mathbf{v}$. The gate vectors are denoted $\mathbf{k}_t = \sigma(\mathbf{x}_k)$ for $\mathrm{k} \in \{z, r\}$ (GRU) and $\mathrm{k} \in \{f, o, i\}$ (LSTM), and $\mathbf{x}_k$ is the appropriate argument of the sigmoid defined for GRU in (2- 3) and for LSTM in (6-7). We use prime to denote differentiation, with $\mathbf{k}'_t = \sigma'(\mathbf{x}_k)$, where $\sigma'(x) = \sigma(x)(1 - \sigma(x))$, and similarly for $\phi'(\mathbf{x})$. We denote the autocorrelation functions as $C_x(t, t') = \mathbb{E}\left[x_t x_{t'}\right]$, with the subscript indicating the variable and arguments (t, t') indicating time. For quenched U_k, the expectation is understood as an average over neurons. In the mean-field limit, this is equivalent to an average over the effective stochastic variables. Finally, the asymptotic notation $O(\cdot)$ and $\Theta(\cdot)$ have the standard definitions.

4. Spectral theory for the GRU

In this section, we focus on the GRU. First, we develop a mean-field theory (MFT) for the GRU and calculate the state-to-state Jacobian with parameters drawn at initialization. The Jacobians are an architecture-dependent combination of structured and random matrices, and thus we can use tools from random matrix theory, combined with the MFT, to elucidate how the spectrum of the Jacobian is shaped by the gates. The eigenvalues of the Jacobian will in principle depend on the particular realization of the connectivity matrices. However, for large random networks, due to self-averaging, the spectrum in fact only depends on the statistics of the state variables, and we develop a self-consistent mean-field theory taking into account recurrent dynamics to calculate these statistics.

4.1. Mean-field theory for the GRU

We first develop a mean-field theory (MFT) for the GRU which gives valuable insight into the structure of the dynamical phase space as the model parameters are varied. First, the theory provides insight into the structure of correlation functions, which is required in the calculation of the instantaneous Jacobian spectrum. Second, it is useful in analyzing the structure of fixed-points (FPs) of the dynamics. The MFT replaces the N dimensional update equation for the GRU with a single stochastic difference equation. Specifically, we can approximate the following terms as Gaussian processes:

$$\left(U_z \mathbf{h}_t + \mathbf{b}_z\right)_i \sim \zeta_t, \quad \left(U_r \mathbf{h}_t + \mathbf{b}_r\right)_i \sim \xi_t, \quad \left(U_h(\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b}_h\right)_i \sim y_t, \tag{9}$$

where ζ_t, ξ_t and y_t are independent Gaussian processes specified by their correlation functions which must be solved for self-consistently. From this, one can obtain a deterministic evolution equation for the correlation functions of various dynamical variables, e.g. $C_h(t, t') := \mathbb{E}[h_t h_{t'}]$. The details of the derivation and the equations for the correlation functions in the general case are

provided in Appendix A, and here, as an example, we provide a summary of the MFT results for FPs (i.e. time-independent solution of the dynamics).

At a FP, the MFT solution to the correlation functions are given by the following implicit equations:

$$C_h = \int Dx \left(\phi(x\sqrt{a_h^2 C_y + v_h} + b_h)\right)^2, \quad C_y = C_h \int Dx \left(\sigma(x\sqrt{a_r^2 C_h + v_r} + b_r)\right)^2, \quad (10)$$

where $Dx = dx \exp(-x^2/2)/\sqrt{2\pi}$ is the Gaussian measure. We use perturbation theory to find solutions in Sec.(4.4), and map out a phase diagram in Fig.(2) indicating topologicaly distinct regions of phase space.

Note that the update gate does not influence the fixed point solutions. This fact is apparent even at the level of Eq. (5). However, as we shall see, the update gate can strongly affect the response to perturbations around the fixed points. Unlike the fixed point solutions above, the MFT solution for the correlation function in a general time-dependent state are obtained by solving a difference equation (see Appendix A). We will use the solutions for the correlation functions from the MFT in our RMT analysis of the state-to-state Jacobian.

4.2. Spectral support for the GRU Jacobian

The spectrum of the Jacobian provides valuable insight into the instantaneous dynamics of the network. We wish to delineate the role of gates by determining how the choice of gates and parameters shapes the instantaneous Jacobian spectrum. To this end, we study the eigenvalues of the Jacobian of randomly initialized GRUs. Our first result is a formula expressing the boundary of the spectral support, which we refer to as the *spectral curve*, in terms of expected values of functions of the dynamical state variables. From a theoretical perspective, this allows us to ascertain precisely how various architectural choices affect the spectrum. We use the method of hermitian reduction Feinberg and Zee (1997) combined with the linearization trick from free probability theory Belinschi et al. (2018), which are necessary ingredients to study the highly structured random non-Hermitian Jacobian. Our implementation of these techniques to the case of GRU and LSTM Jacobian spectra is described in detail in Appendix (B).

The state-to-state Jacobian for the GRU is given by

$$\mathbf{J}_t = \frac{\partial \mathbf{h}_t}{\partial \mathbf{h}_{t-1}} = \hat{\mathbf{z}}_t + (\mathbb{1}_N - \hat{\mathbf{z}}_t)\hat{\phi}'(\mathbf{y}_t)U_h \left(\hat{\mathbf{r}}_t + \hat{\mathbf{h}}_{t-1}\hat{\mathbf{r}}_t' U_r\right) + \left(\hat{\mathbf{h}}_{t-1} - \hat{\phi}(\mathbf{y}_t)\right)\hat{\mathbf{z}}_t' U_z. \quad (11)$$

To study the spectrum of the Jacobian using random matrix techniques, the starting point is the resolvent $(\lambda \mathbb{1}_N - \mathbf{J}_t)^{-1}$. From this, the spectral density is given by

$$\mu(\lambda) = \frac{1}{\pi}\frac{\partial}{\partial \bar{\lambda}} \mathbb{E}\left[\frac{1}{N}\,\mathrm{tr}\left[(\lambda \mathbb{1}_N - \mathbf{J}_t)^{-1}\right]\right], \quad (12)$$

where the expectation is over the random weight matrices. The analysis proceeds by considering the Jacobian for an enlarged space $(\mathbf{h}_t, \mathbf{r}_t, \mathbf{z}_t)$,

$$\mathbf{M}_t = \begin{pmatrix} \hat{\mathbf{z}}_t + (\mathbb{1}_N - \hat{\mathbf{z}}_t)\hat{\phi}'(\mathbf{y}_t)U_h\hat{\mathbf{r}}_t & (\mathbb{1} - \hat{\mathbf{z}}_t)\hat{\phi}'(\mathbf{y}_t)U_h\hat{\mathbf{h}}_{t-1} & \hat{\mathbf{h}}_{t-1} - \hat{\phi}(\mathbf{y}_t) \\ \hat{\mathbf{r}}_t' U_r & 0 & 0 \\ \hat{\mathbf{z}}_t' U_z & 0 & 0 \end{pmatrix}. \quad (13)$$

More precisely, $\mathbf{M}_t$ describes the linear dynamics: $(\delta\mathbf{h}_{t+1}, \delta\mathbf{r}_t, \delta\mathbf{z}_t)^T = \mathbf{M}_t(\delta\mathbf{h}_t, \delta\mathbf{r}_t, \delta\mathbf{z}_t)^T$. The benefit of working with this representation is that each block is now linear in the Gaussian random weight matrices.[1] The only price we pay is that the eigenvalues of $\mathbf{J}_t$ must be found from a *generalized* eigenvalue problem $\mathbf{M}_t\mathbf{v} = \mathbf{I}_\lambda\mathbf{v}$ with the block diagonal matrix $\mathbf{I}_\lambda = \mathrm{bdiag}(\lambda\mathbb{1}_N, \mathbb{1}_N, \mathbb{1}_N)$. The resolvent for this expanded Jacobian is given by

$$\mathbf{G}(\lambda) = (\mathbf{I}_\lambda - \mathbf{M}_t)^{-1}, \tag{14}$$

which will be a $3N \times 3N$ matrix, whose first $N \times N$ block $(\lambda\mathbb{1}_N - \mathbf{J}_t)^{-1} = \mathbf{G}_{11}$ is the resolvent of the Jacobian, the object of interest. Since $\mathbf{J}_t$ is non-hermitian, we use the method of hermitian reduction to study the spectrum (details are in Appendix (B)).

A major result of this analysis is an expression for the spectral support of $\mathbf{J}_t$ in terms of the parameters a_z, a_r, a_h and correlation functions of the dynamical variables in the mean field theory $(z_t = \sigma(\zeta_{t-1}), r_t = \sigma(\xi_{t-1}), h_{t-1}, y_t)$, which we state below:

Theorem 1 (Spectral Support for GRU) *The support of the eigenvalue distribution of $\mathbf{J}_t$ in the limit of large N is given by $\Sigma(\mathbf{J}_t) := \{\lambda \in \mathbb{C} : \quad \mathcal{S}(\lambda) \geq 0\}$ where*

$$\mathcal{S}(\lambda) = \rho_t^2\, \mathbb{E}\left[\frac{(1-z_t)^2}{|\lambda - z_t|^2}\right] + a_z^2\, \mathbb{E}\left[\frac{(z_t')^2(h_{t-1} - \phi(y_t))^2}{|\lambda - z_t|^2}\right] - 1, \tag{15}$$

$$\text{and} \quad \rho_t^2 = a_h^2 C_{\phi'}(t,t)\left[C_r(t,t) + a_r^2 C_{r'}(t,t) C_h(t-1,t-1)\right], \tag{16}$$

and the boundary of the support defines the spectral curve $\partial\Sigma(\mathbf{J}_t) := \{\lambda \in \mathbb{C} : \quad \mathcal{S}(\lambda) = 0\}$.

The equation for the spectral curve Eq.(15) involves equal-time correlation functions of various dynamical variables, and to obtain these values accurately we resort to our self-consistent MFT for the correlation functions. Note that assuming independent weights at each time step will give erroneous results. Fig. (1 a - d) shows the empirical spectrum along with the spectral curve (in red) for various values of a_z and a_r.

One simple case that is nonetheless insightful is the zero fixed-point. The Jacobian around the zero fixed-point can be analysed to yield the density of eigenvalues:

Proposition 2 (Eigenvalue density for zero fixed point) *Around the zero fixed point ($h_t = 0$), $z_t = z = \sigma(b_z)$, $r_t = r = \sigma(b_r)$, and the eigenvalue density is*

$$\mu(\lambda) = \frac{1}{\pi(1-z)^2 r^2 a_h^2}, \quad \text{for } |\lambda - z| \leq (1-z) r a_h, \tag{17}$$

and zero otherwise.

[1] One might worry that, given the recurrent nature of the dynamics of the system of equations 5, all the state variables will depend in some nonlinear fashion on the connectivity matrices, thus making $\mathbf{M}_t$ a complicated nonlinear function of the random variables. Fortunately, we can assume that $\mathbf{h}_t$ and U_k are independent. This assumption was initially made by Amari (1972) and has been referred to as the "local chaos hypothesis" Cessac (1995) (see also Geman and Hwang (1982); Geman (1982)), where it was shown to hold for vanilla RNNs with asymmetric connectivity matrices. Formally, this behavior may be established by the central limit theorem Geman and Hwang (1982); Geman (1982). Our numerical experiments confirm that this assumption is borne out in the systems we study.

Therefore, for zero activity networks, the Jacobian spectrum is uniform and occupies a circular region in the complex plane. In this setting, the consequences of gating are minimal. To observe nontrivial shaping of the spectrum by the gates, one must tune parameters outside the region in which the zero fixed point is stable. In fact, the zero fixed point becomes unstable precisely when the spectral radius exceeds unity. This condition follows directly from Prop. (2)

Corollary 3 *The zero fixed point is* **stable** *for*

$$\sigma(b_z) + (1 - \sigma(b_z))\sigma(b_r)a_h < 1. \tag{18}$$

With zero bias, this condition reduces to that found in Kanai et al. (2017) for stability of the zero fixed point.

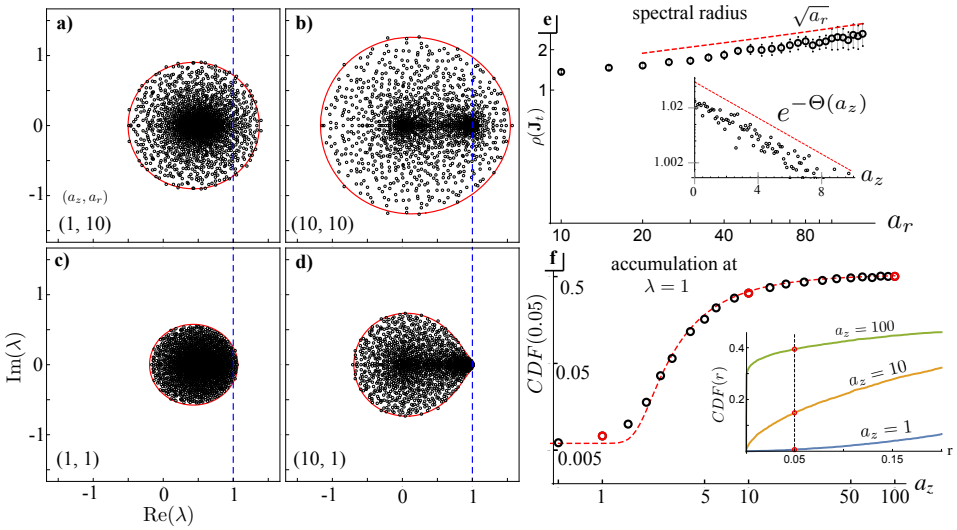

Figure 1: Empirical GRU Jacobian spectrum (black circles) with the spectral curve (15) predicted by RMT (red) in the steady state for $a_h = 3$ at various combinations of (a_z, a_r) given by **a)** $(1, 10)$; **b)** $(10, 10)$; **c)** $(1, 1)$; **d)** $(10, 1)$. In **e)** we show the $\sqrt{a_r}$ growth of the spectral radius with increasing a_r (for $a_z = 0$), as well as the exponential decrease (to unity) with increasing a_z (for $a_r = 0$) (inset). In **f)**, we show the cumulative distribution function $\mathrm{CDF}(r) = P(|\lambda - 1| < r)$ (inset) for various a_z. Increasing a_z causes more eigenvalues to accumulate at $\lambda = 1$ as illustrated in the main figure which shows the CDF at $r = 0.05$. The red dashed curve is the scaling function (94).

4.3. Shaping of dynamics by the *update* and *reset* gates

Having provided a method to calculate the spectral curve, we now look at how the two gates in the GRU – the update and reset gates – each shape the spectrum. For simplifying the discourse, we consider either gate by itself to isolate its contribution, but this is not requisite.

4.3.1. THE *update* GATE FACILITATES CREATION OF SLOW MODES

Here we consider a GRU with only the update gate, and the reset gate variance set to $a_r = 0$. Eigenvalues of the Jacobian close to 1.0 correspond to modes which will evolve slowly, and consequently a clumping of eigenvalues near 1.0 will give rise to a broad spectrum of timescales. In the limit $a_z = 0$, the eigenvalue density becomes circularly symmetric and centered at $\sigma(b_z)$, similar to that of a vanilla RNN. In the opposite limit of $a_z = \infty$, the eigenvalues accumulate near unity and a characterisation of the full density is easily illustrated by considering the nonlinear activation function ϕ to be piecewise-linear (i.e. the "hard-tanh" defined in Eq.(75)). The behavior does not qualitatively change with other saturating non-linearities but the expressions can be complicated. Under this approximation, we can evaluate the density of eigenvalues $\mu(\lambda)$ for $a_z \to \infty$ and $b_z/a_z = \beta$ constant

$$\mu(\lambda) = (1-\alpha)\delta(\lambda - 1) + \alpha(1-\eta)\delta(\lambda) + \frac{1}{\pi a_h^2 \sigma(b_r)^2}\mathbf{1}_{|\lambda| \leq R}, \tag{19}$$

where $R = \sqrt{\alpha\eta}a_h\sigma(b_r)$, η is the fraction of unsaturated activations, and α is the fraction of update gates which are zero (for a derivation of Eq. (19), see Appendix C). This fraction will change depending on $b_z = a_z\beta$, and in general will lead to an extensive number of eigenvalues at $\lambda = 1$. We show this accumulation at $\lambda = 1$ for the Jacobian of a generic time-dependent steady state in Fig.(1 f), which shows good agreement with a scaling function : $c_1 \mathrm{erfc}(c_2/a_z)$ (dashed red); the motivation for this scaling function appears in Appendix C. Interestingly, large a_z also leads to *pinching* of the spectral curve (for a_h in a certain range), thus further accentuating the accumulation of eigenvalues near 1 (Fig. 1 f) and 2b. We will discuss pinching in greater detail in the context of marginal stability (Sec. 4.5). The emergence of slow modes with a spectrum of timescales is likely useful for processing inputs which have dependencies over a wide range of timescales. The accumulation of these slow modes is a generic feature of the gates that control the rate of integration, and we will show later that the forget gate in the LSTM has similar characteristics.

4.3.2. THE *reset* GATE CONTROLS THE SPECTRAL RADIUS

The *reset* gate in the GRU has the effect of controlling the spectral radius $\rho(\mathbf{J}_t)$ of the Jacobian, defined $\rho(\mathbf{J}_t) = \max\{|\lambda_i|\}$. The spectral radius among other things informs us about stability of fixed points. Furthermore, we argue that the reset gate controls a transition from a topologically trivial dynamical phase space with only a single zero fixed point, to a "complex" landscape with a large number of non-trivial fixed points Wainrib and Touboul (2013). To isolate the effect of the reset gate, we set the variance of the update gate $a_z = 0$. A consequence of Theorem 1 is the following corollary regarding the spectral radius:

Corollary 4 *For $a_z = 0$, the spectral radius of the instantaneous Jacobian $\rho(\mathbf{J}_t)$ is given by*

$$\rho(\mathbf{J}_t) = \sigma(b_z) + (1 - \sigma(b_z))\rho_t, \tag{20}$$

with ρ_t defined in (16).

In the Appendix D, we give another proof of this using the Gelfand formula Gelfand (1941), which indeed aligns with our RMT prediction. We clearly observe that higher values of a_r lead to larger spectral radii, e.g. going from Figs. (1c $\to$ a) or (1d $\to$ b). Indeed, we can make this more precise with the following proposition:

Proposition 5 *The spectral radius of the Jacobian* $\mathbf{J}_t$ *for* $a_z = 0$ *grows with* a_r *asymptotically as* $\rho(\mathbf{J}_t) = \Theta(\sqrt{a_r})$.

The red dashed line in Fig (1e) demonstrates that this scaling takes over relatively soon.

4.4. Phase diagram for the GRU

The solutions of the MFT allow us to map out dynamical regimes of qualitatively different behavior. This allows us to construct a phase diagram which, among other things, describes the statistical structure of fixed-points for the GRU. To illustrate our main point, we assume for simplicity $v_h = v_r = b_h = b_r = 0$.

The zero solution $C_h = 0$ always exists, and is only stable for $a_h < 2$ per Corollary 3. A single (unstable) non-zero FP appears continuously from zero for $a_h > 2$ (Fig. 2). This motivates seeking a perturbative solution to Eq. (10) (details are in Appendix E). We study this solution assuming small C_h, and $a_h = 2 + \epsilon$ for $\epsilon > 0$ The perturbative FP solution can exhibit different behavior depending on the magnitude of a_r,

$$C_h \sim \begin{cases} \dfrac{4\epsilon}{8 - a_r^2}, & a_r < \sqrt{2}a_h, \\ \dfrac{\sqrt{\epsilon}}{4}, & a_r = \sqrt{2}a_h \quad , \quad a_h = 2 + \epsilon. \\ \dfrac{6(a_r^2 - 8)}{3a_r^4 + 24a_r^2 - 136} + \dfrac{f(a_r)\epsilon}{a_r^2 - 8}, & a_r > \sqrt{2}a_h \end{cases} \tag{21}$$

where $f(a_r)$ (given in (114)) is a rational polynomial which is positive and nonsingular for $a_r > \sqrt{8}$. For $a_h = 2 - \epsilon$ for positive ϵ, there is no nonzero perturbative solution for $a_r \leq \sqrt{2}a_h$, whereas for $a_r > \sqrt{2}a_h$ *two* nonzero fixed points arise (Fig. 2)

$$C_h \sim \begin{cases} \dfrac{4\epsilon}{a_r^2 - 8}, \\ \dfrac{6(a_r^2 - 8)}{3a_r^4 + 24a_r^2 - 136} - \dfrac{f(a_r)\epsilon}{a_r^2 - 8} \end{cases} \quad a_r > \sqrt{2}a_h, \quad a_h = 2 - \epsilon \tag{22}$$

Evidently, the first solution disappears at the critical value $a_h = 2$, while the second solution is continuous and nonzero across this threshold.

Close to $a_h = 2^-$, we see that above an a_r-threshold, nonzero fixed points appear. They do not exist for all $a_h < 2$; in fact from Eq. 10 (for large a_r) we see that for $a_h < \sqrt{2}$, there are no non-zero fixed points, and for $a_h = \sqrt{2} + \epsilon$, a non-zero fixed point appears perturbatively, scaling like $C_h \sim \epsilon/\sqrt{2}$. Therefore, we see that the reset gate modulates the topology of the dynamical phase space in the interval $a_h \in (\sqrt{2}, 2)$, and produces nonzero fixed points in this regime where the zero FP is stable. In this range for a_h, increasing a_r leads to a pitchfork-like bifurcation of the FP equations, from one (zero) FP to three FPs. At a fixed a_h, this bifurcation occurs at a critical $a_r^*(a_h)$ (with a correspoding variance C_h^*), which we refer to as the bifurcation curve, shown in Fig. (2, Left) .

Thus, the picture that emerges is the following (see Fig.2 Left): when $a_h < \sqrt{2}$, the zero fixed-point is the only stable solution (blue region). For $\sqrt{2} < a_h < 2$, the zero FP is stable, but for a_r above a critical value a_r^*, we also get time-dependent chaotic solutions; this coincides with the abrupt

appearance of the non-zero unstable FP solutions (in the pitchfork bifurcation) for C_h in the MFT which indicate the proliferation of unstable fixed-points in the phase space (green region). When $a_h > 2$ the zero FP is unstable, and the only solutions which are possible are time-dependent chaotic solutions (orange region). We emphasize that even though the appearance of chaotic dynamics coincides with the proliferation of unstable fixed-points, in this study we have only dealt with the analysis of fixed points. A detailed study of the chaotic solutions will be published elsewhere.

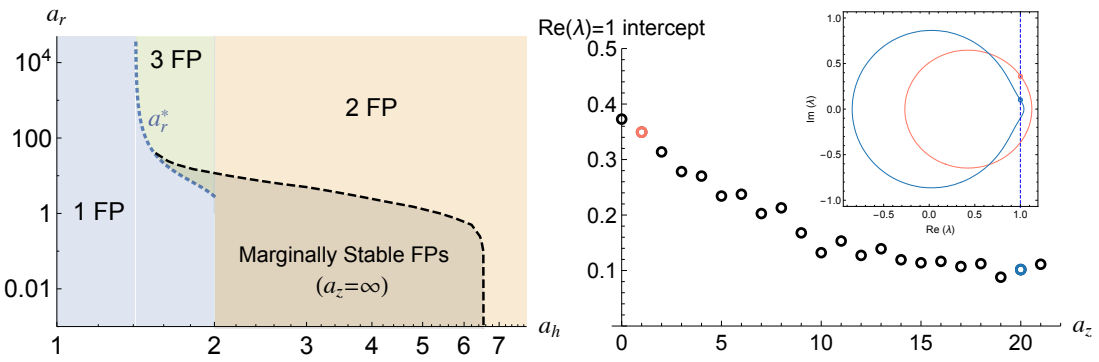

Figure 2: **(Left)** Phase diagram in the (a_h, a_r)-plane for $b_r = 0$ indicating regions with distinct topological structure: (Blue) a single trivial (zero) fixed point only; (Green) three fixed points, two of which are non-zero; (Orange) two fixed points with a single non-zero FP. The bifurcation curve a_r^* (dashed blue) separates the blue from the green region, and is computed numerically. Note that the zero FP always exists. (**Right**) The pinching of the spectral support at $\lambda = 1$ is illustrated by tracking the intercept of the spectral boundary with $\mathrm{Re}(\lambda) = 1$. As this intercept decreases, the spectrum is seen (inset) to become pinched around $\lambda = 1$. Parameters used for numerical simulations: $N = 1000$, $a_r = 2$, $a_h = 3$.

4.5. Stability of non-zero fixed points and emergence of marginal stability

Having looked at the structure of non-zero fixed points in the phase diagram above, we now comment on their stability and the ability of the update gate to make some of the fixed-points marginally stable. The stability of fixed points can be determined from the spectral shaping parameter, and from Thm. (1) we can get a condition for when a fixed-point will be stable. Denoting the spectral shaping parameter (16) at the fixed point ρ_0, we find the stability condition:

Corollary 6 (Fixed Point Stability) *A fixed point is stable iff the spectral shaping parameter (16) satisfies* $\rho_0 < 1$.

Proof For a fixed point, the second term in (15) vanishes. We make use of two observations: that $\mathcal{S}(\lambda)$ is a monotonically decreasing function of $|\lambda|$, and that $\mathcal{S}(|\lambda|e^{i\theta}) \leq \mathcal{S}(|\lambda|)$ for any nonzero θ. Together these imply that the spectral radius is determined by the largest positive real x such that $\mathcal{S}(x) = 0$. Thus, if $\mathcal{S}(1) < 0$, then this implies $x < 1$, and hence stability. Conversely, if the FP is stable, $x < 1$ which implies $\mathcal{S}(1) < 0$ since $\lambda = 1$ is outside the support. Finally, evaluating $\mathcal{S}(1) = \rho_0^2 - 1$, concludes the proof. ■

It turns out that for all non-zero fixed points in the green and orange region in Fig. (2), which satisfy the implicit mean-field equations (10), we find $\rho_t > 1$. Thus, all of the nonzero fixed points implied by the mean-field theory are *unstable*. However, the locally unstable manifold of these fixed-points can be reduced significantly by an appropriate scaling of the update gate variance a_z, and in the asymptotic limit, some of these fixed-points become marginally stable. We state this result below:

Theorem 7 (Emergence of Local Marginal Stability) *For large a_z, and $b_z = a_z\beta$, the asymptotic scaling of the spectral radius of the fixed point Jacobian $\mathbf{J}_0$ is given by*

$$\rho(\mathbf{J}_0) - 1 = \Theta(e^{-c\sqrt{C_h}a_z}), \tag{23}$$

where $c \in \mathbb{R}^+$ is determined implicitly by

$$\mathrm{erf}\left(\frac{c - \beta/\sqrt{C_h}}{\sqrt{2}}\right) = \frac{2}{\rho_0^2} - 1, \tag{24}$$

and ρ_0 is given by Eq.(16) evaluated at the fixed point.

Thus for fixed-points which admit a positive solution for c, we see that increasing a_z makes the spectral radius approach unity, and the fixed-points become marginally stable. Moreover, on increasing a_z not only does the spectral radius reduce to unity for these fixed-points, but the spectral density also gets *pinched* near unity leading to a concentration of eigenvalues near unity (Fig. 2 right). In appendix (C.1), we show that the leading edge of the spectral boundary curve is contained in a wedge near $\lambda = 1$ which scales like $\lambda_0 \exp\left(-c\sqrt{C_h}a_z\right)$, for some order one constant $\lambda_0 \in \mathbb{C}$.

We now comment on the condition for a fixed-point to become marginally stable: a positive solution for c in (24) only exists for

$$\rho_0^2 < \frac{1}{\alpha}, \quad \alpha = \frac{1}{2}\mathrm{erfc}(\beta/\sqrt{2C_h}). \tag{25}$$

In the $a_z \to \infty$ limit, $\alpha = 1 - \mathbb{E}[z_t]$ is simply the fraction of gates which are completely closed ($z_t = 0$). When the bias is zero, $\alpha = 1/2$, and the condition for the spectral shaping parameter becomes $\rho_0 < \sqrt{2}$.

Thus, while $\rho_0 > 1$ for all nonzero fixed points, there is a range of parameters for which $\rho_0 < \sqrt{2}$ and marginal stability and spectral pinching can occur for nonzero fixed points. This region in parameter space is shaded black in Fig. (2 Left). This pinching effect can be observed even for finite a_z (see Fig. (1 d) for $(a_z, a_r) = (10, 1)$, which falls inside the marginal stability region). A practical consequence of the pinching is that it will reduce the number of unstable modes, leading to a non-negligible chance of observing nonzero fixed points in finite random networks.

5. Spectral theory of LSTM

The analysis of the Jacobian for the LSTM proceeds in a similar way as the GRU. For an LSTM with N hidden and cell state variables, the Jacobian is the $2N \times 2N$ matrix

$$\mathbf{J}_t = \begin{pmatrix} \hat{\mathbf{f}}_t & \hat{\mathbf{g}}_t \\ \hat{\mathbf{m}}_t \hat{\mathbf{f}}_t & \hat{\mathbf{o}}'_t \hat{\phi}(\mathbf{c}_t) U_o + \hat{\mathbf{m}}_t \hat{\mathbf{g}}_t \end{pmatrix}, \quad \mathbf{g}_t = \hat{\mathbf{f}}'_t \hat{\mathbf{c}}_{t-1} U_f + \hat{\mathbf{i}}_t \hat{\phi}'(\mathbf{y}_t) U_h + \hat{\mathbf{i}}'_t \hat{\phi}(\mathbf{y}_t) U_i, \tag{26}$$

where $\mathbf{m}_t = \mathbf{o}_t \odot \phi'(\mathbf{c}_t)$. As with the GRU, we wish to evaluate the LSTM resolvent, and thus characterize the spectrum, in the limit of large N. We accomplish this by again combining linearization with hermitian reduction, then performing the ensemble average over U_k (for details see Appendix (G)). Anything we do not average over will enter into the final expression. However, as we demonstrate in the appendix, these expressions involve averages of state variables over the network. To state our results below, we assume the mean-field limit to simplify the resulting correlation functions. In this limit, $(U_k \mathbf{h}_{t-1} + \mathbf{b}_k)_i \sim \eta^k_{t-1}$ and $(U_h \mathbf{h}_{t-1} + \mathbf{b}_h)_i \sim y_t$ become independent Gaussian processes, and c_t, h_t are random variables (for details see App. (A)). For the gating variables we use the notation $k_t = \sigma(\eta^k_{t-1})$, and $k'_t = \sigma'(\eta^k_{t-1})$. Our first result concerns the trace of the resolvent of $\mathbf{J}_t$ in the large N limit in terms of the parameters a_f, a_i, a_o, a_h, as well as the spectral support. Note that we normalize the resolvent by N as in (12) instead of $2N$.

Theorem 8 (LSTM Resolvent) *The trace of the resolvent for the LSTM Jacobian in the large N limit in the mean-field theory is*

$$G(\lambda) = \begin{cases} \mathbb{E}\left[\dfrac{|\lambda|^2(\bar{\lambda} - f_t) + \bar{\lambda}|\lambda - f_t|^2 + F(\lambda)\left((\bar{\lambda} - f_t) q_t + \bar{\lambda} p_t\right)}{|\lambda|^2 |\lambda - f_t|^2 + F(\lambda)\left(|\lambda - f_t|^2 q_t + |\lambda|^2 p_t\right)} \right], & \lambda \in \Sigma, \\ \dfrac{1}{\lambda} + \mathbb{E}\left[\dfrac{1}{\lambda - f_t} \right], & \lambda \in \Sigma^c, \end{cases} \tag{27}$$

where

$$q_t = a_o^2 o_t'^2 \phi(c_t)^2, \qquad p_t = o_t^2 \phi'(c_t)^2 \left(a_f^2 c_{t-1}^2 f_t'^2 + a_i^2 i_t'^2 \phi(y_t)^2 + a_h^2 i_t^2 \phi'(y_t)^2 \right), \tag{28}$$

and $F(\lambda) \geq 0$ for $\lambda \in \Sigma$ solves the implicit equation

$$1 = \mathbb{E}\left[\frac{|\lambda - f_t|^2 q_t + |\lambda|^2 p_t}{|\lambda|^2 |\lambda - f_t|^2 + F(\lambda)\left(|\lambda - f_t|^2 q_t + |\lambda|^2 p_t\right)} \right], \tag{29}$$

whereas $F = 0$ outside the support $\lambda \in \Sigma^c$.

Outside the support, the resolvent becomes a holomorphic function of λ, whereas inside it depends on both λ and $\bar{\lambda}$. The trace of the resolvent $G(\lambda)$ is continuous on the complex plane, which means the holomorphic solution must match the non-analytic solution precisely at the boundary of the support. This allows us to deduce the spectral boundary curve by setting $F(\lambda) = 0$ in (29).

Corollary 9 (LSTM Boundary Curve) *The boundary of the spectral support is given by the curve*

$$\partial\Sigma(\mathbf{J}_t) = \{\lambda \in \mathbb{C} : \mathcal{S}(\lambda) = 0\} \tag{30}$$

where

$$\mathcal{S}(\lambda) = \frac{\mathbb{E}\left[q_t\right]}{|\lambda|^2} + \mathbb{E}\left[\frac{p_t}{|\lambda - f_t|^2} \right] - 1. \tag{31}$$

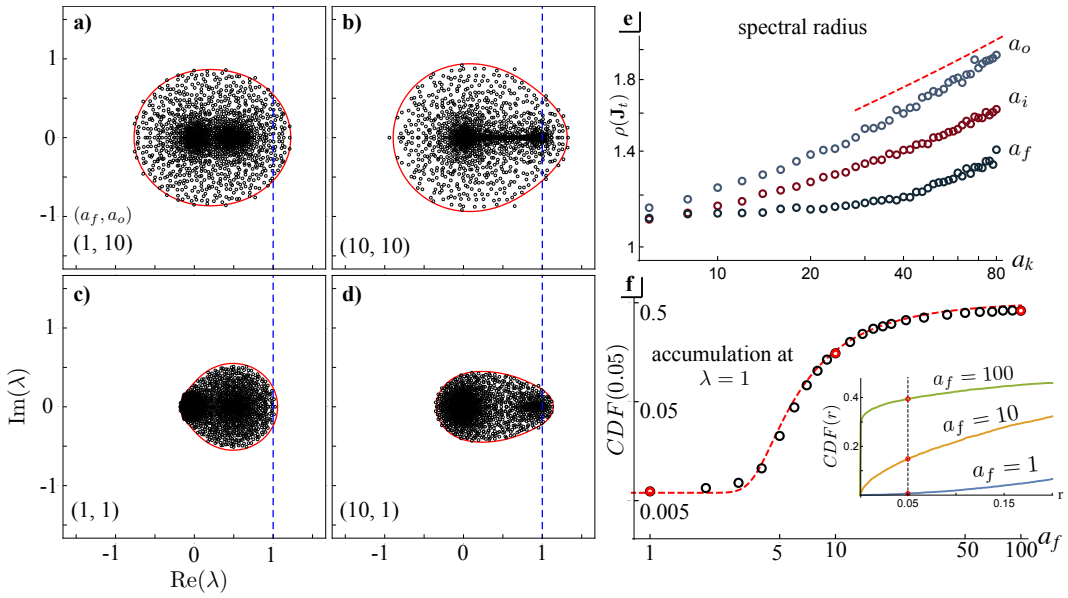

Figure 3: (**a - d**) Empirical spectrum of LSTM Jacobian in steady state (black circles) compared to RMT spectral boundary prediction (30) (red) using steady state correlation functions: $N = 1000$, $a_h = 3$, $a_i = 1$, and the values of (a_f, a_o) are indicated for each plot. (**e**) Numerically computed spectral radius as a function of a_k when all other parameters ($a_{k'}$, $k' \neq k$) are set to zero, with $\sqrt{a_k}$ scaling (dashed red) to guide the eye. (**f**) Cumulative distribution of eigenvalues around a small ball of radius $r = 0.05$ as a function of a_f, with $a_i = a_o = 0$ and zero biases, compared to scaling function (94) (dashed red). Inset shows CDF as a function for r for different a_f for $a_i = a_o = 0$

We present the general result for the density of eigenvalues in the appendix, which is not very illuminating. However, around the zero fixed point, the density of eigenvalues simplifies considerably:

Corollary 10 (Zero Fixed Point) *Around the zero fixed point $c_t = h_t = 0$, $p_t = p = \sigma(b_o)^2\sigma(b_i)^2 a_h^2$, $q_t = 0$, and the density of eigenvalues becomes*

$$\mu(\lambda) = \delta(\lambda) + \frac{1}{\pi p}, \quad \text{for } |\lambda - f| \leq \sqrt{p} \tag{32}$$

and zero otherwise.

Note with our normalization, $\int d\mu(\lambda) = 2$. From this, we can directly infer the stability condition for the zero fixed point

Corollary 11 *The zero fixed point is* **stable** *for*

$$\sigma(b_f) + \sigma(b_o)\sigma(b_i)a_h < 1 \tag{33}$$

It is worth emphasizing that the density of eigenvalues depends on most state variables (e.g. input and output gates, activation functions) only through the combinations which appear in (28). An exception is the forget gate, which enters separately and in a rather influential way. Consequently, we find that very naturally the effects of the input and output gates in shaping the spectrum are roughly comparable, while the forget gate plays a qualitatively distinct role.

5.1. The *forget* gate facilitates slow modes

The *forget* gate in the LSTM has a role very similar to that of the GRU *update* gate. In particular, large values of a_f facilitate the accumulation of eigenvalues near unity thus leading to slow modes (Fig. (3f)). In the limit $a_f = \infty$, using again a piecewise linear approximation of the activation functions ϕ, we find the density of the eigenvalues

$$\mu(\lambda) = \left(1 + \frac{(1-\eta)}{2}\right)\delta(\lambda) + \frac{1}{2}\delta(\lambda - 1) + \frac{1}{\pi R^2}\mathbf{1}_{\lambda\in\Sigma}, \quad \Sigma := \{\lambda \in \mathbb{C} : |\lambda| \leq \frac{\sqrt{\eta}}{\sqrt{2}}R\}, \quad (34)$$

where $R = \sigma(b_o)\sigma(b_i)a_h$, and η is again the fraction of activations which are unsaturated (see App. (G.2)). As before, we see that ramping up a_f leads to an accumulation of eigenvalues at $\lambda = 1$, becoming a delta function in the extreme limit.

When $a_o = 0$, the first N rows of $\mathbf{J}_t$ (26) become linearly dependent with the second N (due to U_0 vanishing), guaranteeing at least N exact zero eigenvalues. The accumulation at zero will generally grow in proportion to the fraction of saturated activations. Finally, in the limit $a_f \to \infty$, the mean density away from these accumulation points is flat, controlled by the input and output biases as well as the variance a_h.

An important assumption in arriving at this expression is that the forget gate is completely saturated, essentially becoming a binary variable. Practically speaking, this allows us to neglect contributions from f_t'. As a result, we find a spectral radius which is independent of a_f, fixed by other parameters in the problem. However, we observe numerically in Fig.(3) that the spectral radius in fact grows with a_f, albeit slower than with a_i and a_o. The scaling of the spectral radius in this regime is shown to be no greater than $\sqrt{a_f}$ in Appendix (G.2).

5.2. The *input* and *output* gates modulate the spectral radius

The *output* and *input* gates in the LSTM have similar effects on the shape of the Jacobian and are comparable in this regard to the *reset* gate in the GRU; the effect of the *forget* gate on the spectral radius is modest (see Fig.(3e). Setting $a_f = 0$, the spectral boundary curve simplifies considerably to read

$$1 = \frac{\mathbb{E}[q_t]}{|\lambda|^2} + \frac{\mathbb{E}[p_t]}{|\lambda - f|^2} \quad (35)$$

The spectral radius will consequently be controlled by the asymptotic behavior of the coefficients $\mathbb{E}[q_t]$ and $\mathbb{E}[p_t]$. First, we consider $a_o = 0$. In this case, it immediately follows that $q_t = 0$, and we are left with only p_t. The spectral support then becomes a circle centered on $\sigma(b_f)$ with radius $\sqrt{\mathbb{E}[p_t]}$. In appendix (G.3), we show that when $a_o = 0$, $\mathbb{E}[p_t] = O(a_i)$.

Isolating the effects of the output gate by setting $a_i = 0$, we find that $\mathbb{E}[p_t] = O(1)$, whereas $\mathbb{E}[q_t] = O(a_o)$.

Gathering these results, we find a universal scaling of the spectral radius with the gate variance:

Proposition 12 *The spectral radius of the LSTM Jacobian scales asymptotically as*

$$\rho(\mathbf{J}_t) = O(\sqrt{a_k}), \quad a_k \to \infty, \quad a_{k' \neq k} = \text{fixed} \tag{36}$$

We compare this scaling behavior with numerical experiments in Fig. (3), showing good agreement. We give additional arguments in Appendix (G.3) and (G.4) that in fact $\rho(\mathbf{J}_t) = \Theta(\sqrt{a_k})$ for the input and output gates $k = i, o$. This indicates the spectral radius indeed strictly grows with these gate variances. However, we lack a tight lower bound for the growth of the spectral radius with the forget gate, most likely because the lower bound is an order one constant, as suggested by the $a_f = \infty$ analysis. Nevertheless, we are able to observe the $\sqrt{a_f}$ scaling in Fig. (3) over a modest range of a_f.

6. Discussion: future directions and relation to training

We have undertaken a detailed analysis of the dynamics of randomly initialized GRUs and LSTMs. These dynamical properties are critical to how the networks process temporal input. We showed that the gates which affect the rate of integration – the update gate in the GRU and the forget gate in the LSTM – can faciliate long timescales by shaping the Jacobian spectrum to cluster near unity. Moreover, we showed how the update gate in the GRU, in certain regimes, can make the system *marginally stable*.

Marginal stability can propagate perturbations over a spectrum of long timescales and facilitate the existence of a continuous manifold of attractors. Recently Maheswaranathan et al. (2019) found that trained gated-RNNs naturally discover solutions for which the dynamics tends toward a line attractor. Our analysis shows that marginal stability is a generic feature in the presence of binary switch-like gates that control the rate of integration. We showed that the reset gate in the GRU and the output and input gates in the LSTM control the spectral radius of the Jacobian and thus shape the complexity of the dynamics. For the GRU, we obtained a full phase-diagram of dynamics and showed how the reset gate can lead to a proliferation of unstable fixed points coinciding with the abrupt appearance of a chaotic attractor.

How does our study of the dynamical properties of gated RNNs inform us about training performance? While the dynamical properties are crucial for performance, these properties will change during training; however, initialization is known to have a significant impact on trainability Sutskever et al. (2012). Furthermore, in feedforward networks, it was shown that parameters do not change significantly from initialization when training large networks with appropriately small learning rate Jacot et al. (2018); Lee et al. (2018). This regime is relatively unexplored for RNNs. A detailed study of how the dynamical properties change during training will be undertaken elsewhere, but in Appendix (H) we provide very preliminary results on training GRUs on a sequential task that requires processing long-time dependencies. Overall, we observe (see Fig 4) that both training time and training accuracy improve with larger values of a_z (more slow modes). Interestingly, we observe that values of a_h slightly above the critical point $a_h = 2.0$ (in the marginally stable region indicated in Fig. (2)) give the best performance (for high a_z). This bears some semblance to the benefits of "edge of chaos" initialization in vanilla RNNs Bertschinger and Natschl (2004), and feed forward networks Glorot and Bengio (2010); Schoenholz et al. (2017). The improved performance with increasing a_z furthermore suggests that proximity to marginal stability is also a boon for training. This might explain the relative ease of training GRUs and LSTMs, since, as we showed, gating in both architectures quite naturally leads to marginal stability. A more thorough characterization of

how dynamics evolve during training and the relation to training performance will be undertaken in future work. We have provided analytical tools to study the dynamics of RNNs that we expect can be fruitfully applied in contexts other than at initialization.

Acknowledgements We would like to thank Devon Wood-Thomas for comments on the draft. KK is supported by a C.V. Starr Fellowship and a CPBF Fellowship (through NSF PHY-1734030). DJS acknowledges support from NSF PHY-1734030, NIH 5R01EB026943 and a Simons Investigator grant in MMLS.

References

Yashar Ahmadian, Francesco Fumarola, and Kenneth D Miller. Properties of networks with partially structured and partially random connectivity. *Physical Review E*, 91:012820, 2015.

Johnatan Aljadeff, David Renfrew, and Merav Stern. Eigenvalues of block structured asymmetric random matrices. *Journal of Mathematical Physics*, 56:103502, 2015a.

Johnatan Aljadeff, Merav Stern, and Tatyana Sharpee. Transition to chaos in random networks with cell-type-specific connectivity. *Physical Review Letters*, 114:088101, 2015b.

Shun-Ichi Amari. Characteristics of random nets of analog neuron-like elements. *IEEE Transactions on Systems, Man, and Cybernetics*, (5):643–657, 1972.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*, 2014.

Serban T. Belinschi, Piotr Śniady, and Roland Speicher. Eigenvalues of non-Hermitian random matrices and Brown measure of non-normal operators: Hermitian reduction and linearization method. *Linear Algebra and Its Applications*, 537:48, 2018.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

Nils Bertschinger and Thomas Natschl. Real-Time Computation at the Edge of Chaos in Recurrent Neural Networks. *Neural Computation*, 1436:1413–1436, 2004.

B Cessac. Increase in Complexity in Random Neural Networks. *J Phys I France*, 5(3):409–432, 1995.

M. Chen, J. Pennington, and S. S. Schoenholz. Dynamical Isometry and a Mean Field Theory of RNNs: Gating Enables Signal Propagation in Recurrent Neural Networks. In *International Conference on Machine Learning (ICML)*, 2018.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014.

Bernard Derrida and Yves Pomeau. Random networks of automata: a simple annealed approximation. *EPL (Europhysics Letters)*, 1(2):45, 1986.

Joshua Feinberg and Anthony Zee. Non-hermitian random matrix theory: Method of hermitian reduction. *Nuclear Physics B*, 504(3):579–608, 1997.

I. Gelfand. Normierte ringe. *Rec. Math. [Mat. Sbornik] N.S.*, 9(51)(1):3–24, 1941.

Stuart Geman. Almost sure stable oscillations in a large system for randomly coupled equations. *SIAM J. Appl. Math.*, 42(4):695–703, 1982.

Stuart Geman and Chii-Ruey Hwang. A chaos hypothesis for some large systems of random equations. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 60(3):291–314, 1982.

Felix A. Gers, J. Schmidhuber, and Fred Cummins. Learning to Forget: Continual Prediction with LSTM. In *International Conference on Artificial Neural Networks*, volume 2, pages 850–855, 1999.

Dar Gilboa, Bo Chang, Minmin Chen, Greg Yang, Samuel S Schoenholz, Ed H Chi, and Jeffrey Pennington. Dynamical Isometry and a Mean Field Theory of LSTMs and GRUs. *arXiv preprint arXiv:1901.08987*, 2019.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, and Jürgen Schmidhuber. Gradient flow in recurrent nets: The difficulty of learning long-term dependencies. In *A Field Guide to Dynamical Recurrent Neural Networks*, pages 237–243. Wiley-IEEE Press, 2001.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8571–8580. 2018.

Li Jing, Caglar Gulcehre, John Peurifoy, Yichen Shen, Max Tegmark, Marin Soljacic, and Yoshua Bengio. Gated orthogonal recurrent units: On learning to forget. *Neural computation*, 31(4): 765–783, 2019.

Ian D Jordan, Piotr Aleksander Sokol, and Il Memming Park. Gated recurrent units viewed through the lens of continuous time dynamical systems. *arXiv preprint arXiv:1906.01005*, 2019.

Sekitoshi Kanai, Yasuhiro Fujiwara, and Sotetsu Iwamura. Preventing gradient explosions in gated recurrent units. In *Advances in Neural Information Processing Systems*, pages 435–444, 2017.

Giancarlo Kerg, Kyle Goyette, Maximilian Puelma Touzel, Gauthier Gidel, Eugene Vorontsov, Yoshua Bengio, and Guillaume Lajoie. Non-normal recurrent neural network (nnrnn): learning long time dependencies while improving expressivity with transient dynamics. In *Advances in Neural Information Processing Systems 32*, pages 13613–13623. 2019.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, pages 3294–3302, 2015.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-dickstein. Deep Neural Networks as Gaussian Processes. In *International Conference on Learning Representations (ICLR)*, 2018.

Niru Maheswaranathan, Alex H Williams, Matthew D Golub, Surya Ganguli, and David Sussillo. Line attractor dynamics in recurrent networks for sentiment classification. In *International Conference on Machine Learning (ICML)*, 2019.

Daniel Marti, Nicolas Brunel, and Srdjan Ostojic. Correlations between synapses in pairs of neurons slow down dynamics in randomly connected neural networks. *Physical Review E*, 97:062314, 2018.

Bernhard Mehlig and John T Chalker. Statistical properties of eigenvectors in non-hermitian gaussian random matrix ensembles. *Journal of Mathematical Physics*, 41(5):3233–3256, 2000.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning (ICML)*, pages 1310–1318, 2013.

Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In *Advances in Neural Information Processing Systems*, pages 4785–4795. 2017.

Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. In *International Conference on Learning Representations (ICLR)*, 2017.

H. Sompolinsky, A. Crisanti, and H. J. Sommers. Chaos in random neural networks. *Physical Review Letters*, 61(3):259, 1988.

M Stern, H Sompolinsky, and L F Abbott. Dynamics of random neural networks with bistable units. *Physical Review E*, 90:062710, 2014.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning (ICML)*, 2012.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112. 2014.

Corentin Tallec and Yann Ollivier. Can recurrent neural networks warp time? In *International Conference on Learning Representations (ICLR)*, 2018.

Gilles Wainrib and Jonathan Touboul. Topological and dynamical complexity of random neural networks. *Physical review letters*, 110(11):118101, 2013.

Appendix A. Mean Field Theory for GRUs and LSTMs

In this appendix, we describe the dynamical mean field equations for GRUs, and elaborate on aspects relevant for the analysis in this paper. A more thorough study of the dynamical equations which follow from the MFT will be undertaken in future work.

The mean field theory for the GRU uses the central limit theorem and the chaos hypothesis (Amari (1972); Geman and Hwang (1982); Cessac (1995)) to replace, for a given neuron, its N interactions with an effective stochastic variable. Concretely, the central limit theorem is invoked to write

$$(U_z \mathbf{h}_t + \mathbf{b}_z)_i \to \zeta_t \tag{37}$$
$$(U_r \mathbf{h}_t + \mathbf{b}_r)_i \sim \xi_t \tag{38}$$
$$(U_h(\mathbf{h}_t \odot \mathbf{r}_t) + \mathbf{b}_h)_i \to y_t \tag{39}$$

where (ζ_t, ξ_t, y_t) are independent Gaussian processes characterized by their moments

$$\mathbb{E}[y_t] = b_h, \quad \text{cov}[y_t y_{t'}] = a_h^2 C_r(t,t') C_h(t,t') + v_h, \tag{40}$$
$$\mathbb{E}[\zeta_t] = b_z, \quad \text{cov}[\zeta_t \zeta_t] = a_z^2 C_h(t,t') + v_z, \tag{41}$$
$$\mathbb{E}[\xi_t] = b_r, \quad \text{cov}[\xi_t \xi_{t'}] = a_r^2 C_h(t,t') + v_r, \tag{42}$$

with the kernels related to the correlation functions

$$C_h(t,t') = \mathbb{E}[h_t h_{t'}], \quad C_r(t,t') = \mathbb{E}[\sigma(\xi_t)\sigma(\xi_{t'})]. \tag{43}$$

A crucial aspect of the dynamical mean-field theory (DMFT) is that the Gaussian kernels are non-Markovian (i.e. noise at different time steps are correlated), and must be determined self-consistently from the correlation function of the hidden state variable h_t. Thus, as the network size $N \to \infty$, each neuron becomes essentially independent of the other neurons. The N dimensional dynamical system is then reduced to a one-dimensional stochastic difference equation for the hidden state variable, where the noise term embodies the effective interaction with the rest of the network

$$h_t = \sigma(\zeta_{t-1}) h_{t-1} + (1 - \sigma(\zeta_{t-1}))\phi(y_{t-1}). \tag{44}$$

In what follows, we will also make use of the additional correlation functions, which may be viewed as transforms of a Gaussian process

$$C_z(t,t') = \mathbb{E}[\sigma(\zeta_t)\sigma(\zeta_{t'})], \quad C_\phi(t,t') = \mathbb{E}[\phi(y_t)\phi(y_{t'})], \quad \kappa_t = \mathbb{E}[\sigma(\zeta_t)]. \tag{45}$$

For long times, the autonomous dynamics will approach a steady state described either by a fixed point or chaotic attractor. Limit cycles require fine tuned weight matrices, and are expected to have vanishing probability in the limit of large N. Furthermore, upon reaching this limiting distribution, the correlation function should only depend on the absolute difference in times, e.g. $C_h(t, t+k) \to$

$C_h(k)$ and averages become time-independent e.g. $\kappa_t = \kappa$. The DMFT in the steady state reduces to the second order difference equation

$$C_h(k) - \kappa C_h(k-1) - \kappa C_h(k+1) + C_z(k)C_h(k) = (1 - 2\kappa + C_z(k))\, C_\phi(k), \tag{46}$$
$$C_h(0) + C_z(0)C_h(0) - 2\kappa C_h(1) = (1 - 2\kappa + C_z(0))C_\phi(0). \tag{47}$$

Fixed points are described by time-independent distributions. Assuming time-independent in (46) and (47) gives the implicit equation (10) stated in the main text.

Another useful relation follows from the fact that the autocorrelation function is bounded at all times $C_h(|t - t'|) \leq C_h(0)$. Using this in (47) leads to the bound for the equal-time correlation function for time-dependent solutions

$$C_h(0) \leq C_\phi(0). \tag{48}$$

If we let C_h^{FP} denote the fixed point solution $C_h^{FP} = C_\phi(0)$, then this inequality implies $C_h(0) \leq C_h^{FP}$. Therefore, the fixed point (time-independent) solution is an upper bound on the equal-time autocorrelation function in a time-dependent steady state.

LSTM Applying a similar reasoning as above, we arrive at the following effective stochastic difference equation for the LSTM (8)

$$c_t = \sigma(\eta^f_{t-1})c_{t-1} + \sigma(\eta^i_{t-1})\phi(y_{t-1}), \tag{49}$$
$$h_t = \sigma(\eta^o_{t-1})\phi(c_t), \tag{50}$$

where y_t is a Gaussian process with mean b_h and covariance

$$\mathrm{cov}\left[y_t y_{t'}\right] = a_h^2 C_h(t,t') + v_h, \tag{51}$$

and η_t^k for $k \in \{f, i, o\}$ are Gaussian processes with mean b_k and covariance

$$\mathrm{cov}\left[\eta_t^k \eta_{t'}^{k'}\right] = \delta_{kk'}\left(a_k^2 C_h(t,t') + v_k\right), \quad C_h(t,t') = \mathbb{E}\left[h_t h_{t'}\right]. \tag{52}$$

c_t and h_t governed by such a stochastic difference equation will not be gaussian, and this makes the fixed point analysis difficult. However, we see at the very least that the η_t^k are independent. Henceforth, for simplicity we denote $k_t = \sigma(\eta^k_{t-1})$ and $k'_t = \sigma'(\eta^k_{t-1})$. Similarly for the GRU $z_t = \sigma(\zeta_{t-1})$, $z'_t = \sigma'(\zeta_{t-1})$, $r_t = \sigma(\xi_{t-1})$, and $r'_t = \sigma'(\xi_{t-1})$

Appendix B. GRUs: proof of Theorem 1

Here we outline the proof of our main result concerning the spectral curve. Our approach uses the ideas about hermitization method and free probability theory to truncate the cumulants. We are interested in determining the spectral properties of the state-to-state Jacobian (11). In the main text, we suggested that it is advantageous to look at an enlarged Jacobian $\mathbf{M}_t$ defined in eq.(13). This extension is properly thought of as an implementation of the linearization method (see e.g. Belinschi et al. (2018) and references therein) to deal with products of random matrices. The linearization in this problem happens to be naturally suggested by the structure of the GRU network dynamics.

Consider the generalized resolvent

$$\mathbf{G}(\lambda) = (\mathbf{I}_\lambda - \mathbf{M}_t)^{-1} = \begin{pmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} & \mathbf{G}_{13} \\ \mathbf{G}_{21} & \mathbf{G}_{22} & \mathbf{G}_{23} \\ \mathbf{G}_{31} & \mathbf{G}_{32} & \mathbf{G}_{33} \end{pmatrix}. \tag{53}$$

It is simple to check that $\mathbf{G}_{11} = (\lambda - \mathbf{J}_t)^{-1}$ is the resolvent of Jacobian. It is convenient to decompose the extended Jacobian (13) in the form (dropping the time index for simplicity) $\mathbf{M} = \hat{A} + \hat{B}\hat{U}\hat{C}$ where

$$\hat{A} = \begin{pmatrix} \hat{\mathbf{z}} & 0 & \hat{\mathbf{h}} - \hat{\phi}(\mathbf{y}) \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \hat{B} = \begin{pmatrix} (\mathbb{1} - \hat{\mathbf{z}})\hat{\phi}'(\mathbf{y}) & 0 & 0 \\ 0 & \mathbf{r}' & 0 \\ 0 & 0 & \mathbf{z}' \end{pmatrix}, \quad \hat{C} = \begin{pmatrix} \mathbf{r} & \mathbf{h} & 0 \\ \mathbb{1} & 0 & 0 \\ \mathbb{1} & 0 & 0 \end{pmatrix},$$

and $\hat{U} = \mathrm{bdiag}(U_h, U_r, U_z)$. Following the well-known method of hermitian reduction, we introduce an auxiliary complex variable η and consider the resolvent $\mathcal{G}$, whose inverse is

$$\mathcal{G}^{-1} = \begin{pmatrix} \eta\mathbb{1}_{3N} & \mathbf{I}_\lambda - \mathbf{M} \\ \mathbf{I}_{\bar{\lambda}} - \mathbf{M}^T & \eta\mathbb{1}_{3N} \end{pmatrix} = \begin{pmatrix} \eta\mathbb{1}_{3N} & \mathbf{I}_\lambda - \hat{A} \\ \mathbf{I}_{\bar{\lambda}} - \hat{A}^T & \eta\mathbb{1}_{3N} \end{pmatrix} - \begin{pmatrix} 0 & \hat{B}\hat{U}\hat{C} \\ \hat{C}^T\hat{U}^T\hat{B}^T & 0 \end{pmatrix}. \tag{54}$$

Denote the four $3N \times 3N$ blocks of $\mathcal{G}$ by $\mathcal{G}_{ab}$ for $a, b \in \{1, 2\}$. We identify $\mathcal{G}_{21} = \mathbf{G}(\lambda)$ as the generalized resolvent of interest. The expression (54) is of the form $\mathcal{G}^{-1} = \mathcal{G}_0^{-1} - \mathcal{H}$, where $\mathcal{H}$ contains the Gaussian random variables $\hat{U}$. Since the elements of the matrix $\mathcal{H}$ are gaussian, we utilize the so-called self-consistent Born approximation to compute the self-energy (see e.g. Mehlig and Chalker (2000)). In the language of Free Probability theory, this is equivalent to calculating the R transform by keeping only the second cumulant. This approximation is expected to be exact in the $N \to \infty$ limit. In this limit, the self-energy $\mathbf{\Sigma}$ will be block diagonal, with the $N \times N$ blocks given by

$$\mathbf{\Sigma}_{11}[\mathcal{G}] = \hat{B}\mathcal{Q}\left[\hat{C}\mathcal{G}_{22}\hat{C}^T\right]\hat{B}^T, \quad \mathbf{\Sigma}_{22}[\mathcal{G}] = \hat{C}^T\mathcal{Q}\left[\hat{B}^T\mathcal{G}_{11}\hat{B}\right]\hat{C}, \tag{55}$$

where the superoperator, defined by the action

$$\mathcal{Q}[R] = \begin{pmatrix} \frac{a_h^2}{N}\mathrm{tr}R_{11} & 0 & 0 \\ 0 & \frac{a_r^2}{N}\mathrm{tr}R_{22} & 0 \\ 0 & 0 & \frac{a_z^2}{N}\mathrm{tr}R_{33} \end{pmatrix}, \tag{56}$$

arises after ensemble averaging over the random connectivity matrices in the block diagonal $\hat{U}$. We have therefore replaced the explicit random variable $\mathcal{H}$ with an effective "self-energy" and rewritten the expression for the resolvent $\mathcal{G}$ in the form $\mathcal{G}^{-1} = \mathcal{G}_0^{-1} - \mathbf{\Sigma}[\mathcal{G}]$, known as the Dyson equation. Explicitly, to compute the resolvent we must solve the non-linear system of equations implied by the expression

$$\mathcal{G}^{-1} = \begin{pmatrix} -\mathbf{\Sigma}_{11}[\mathcal{G}] & \mathbf{I}_\lambda - \hat{A} \\ \mathbf{I}_{\bar{\lambda}} - \hat{A}^T & -\mathbf{\Sigma}_{22}[\mathcal{G}] \end{pmatrix}. \tag{57}$$

Since the self-energy involves traces of transformed blocks of the resolvent, it is useful to work directly with these transformed matrices. So we define

$$\mathcal{F}_{11} = \hat{B}^T \mathcal{G}_{11} \hat{B}, \quad \mathcal{F}_{22} = \hat{C} \mathcal{G}_{22} \hat{C}^T. \tag{58}$$

We will end up needing only the diagonal blocks of these matrices, since this is all that enters the self-energy. For $\mathcal{F}_{11}$, we denote the diagonal block matrices F_{ii}, for $i \in \{1, 2, 3\}$, and similarly for $\mathcal{F}_{22}$ with indices $i \in \{4, 5, 6\}$. Finally, we denote the normalized trace $f_{ii} = \frac{1}{N} \mathrm{tr} F_{ii}$. Furthermore, for simplicity, define the functions $Q = \frac{f_{22}+f_{33}}{f_{11}}$, $X = f_{11} f_{44}$ and $Y = f_{11} f_{55}$. The Dyson equation then leads to a system of equations for various Green's functions which we must solve self-consistently.

To present our results, we remove the bold-face and carat (e.g. $\hat{\mathbf{h}} \to h$), and denote

$$\frac{1}{N} \mathrm{tr} \left(\mathbf{Q}(\mathbf{h}, \mathbf{z}, \mathbf{y}, \mathbf{r}) \right) = \mathbb{E} \left[Q(h, z, y, r) \right], \tag{59}$$

where Q is a matrix-valued function of the state variables, e.g. $\mathbf{Q} = \hat{\mathbf{h}}$. In the large N limit where the mean field theory is expected to hold, we may take this to mean that the sum over all neurons is literally replaced by an expectation value over a single neuron with fluctuating fields.

An important object, which is a non-Hermitian generalization of the Stieltjes transform, is the trace of the resolvent $G(\lambda) = \mathbb{E} \left[\frac{1}{N} \mathrm{tr} (\lambda - \mathbf{J}_t)^{-1} \right]$. From the Dyson equation, we find

$$G(\lambda) = \mathbb{E} \left[\frac{(\bar{\lambda} - z)(1 - Y h^2 r'^2)}{D} \right], \tag{60}$$

where the denominator is given by

$$\begin{aligned} D = {} & |\lambda - z|^2 (1 - Y h^2 (a_r r')^2) - (1 - z)^2 (a_h \phi'(y))^2 (QX + X r^2) + XYQ (1 - z)^2 (a_h \phi'(y))^2 h^2 (a_r r')^2 \\ & - QY (h - \phi(y))^2 (a_z z')^2 - Y (r^2 - QY h^2 (a_r r')^2)(h - \phi(y))^2 (a_z z')^2, \end{aligned} \tag{61}$$

and the auxiliary functions must be computed self-consistently to satisfy the system of equations

$$1 = \mathbb{E} \left[\frac{(1 - z)^2 (a_h \phi'(y))^2 r^2}{D} \right] + Q \, \mathbb{E} \left[\frac{(1 - z)^2 (a_h \phi'(y))^2 \left(1 - Y h^2 (a_r r')^2 \right)}{D} \right] \tag{62}$$

$$\begin{aligned} X = {} & X \mathbb{E} \left[\frac{(1 - z)^2 (a_h \phi'(y))^2 r^2}{D} \right] - QXY \mathbb{E} \left[\frac{(1 - z)^2 (a_h \phi'(y))^2 h^2 (a_r r')^2}{D} \right] + Y \mathbb{E} \left[\frac{r^2 (h - \phi(y))^2 (a_z z')^2}{D} \right] \\ & + Y \mathbb{E} \left[\frac{h^2 (a_r r')^2 |\lambda - z|^2}{D} \right] - QY^2 \mathbb{E} \left[\frac{h^2 (a_r r')^2 (h - \phi(y))^2 (a_z z')^2}{D} \right] \end{aligned} \tag{63}$$

$$\begin{aligned} Y = {} & X \mathbb{E} \left[\frac{(1 - z)^2 (a_h \phi'(y))^2}{D} \right] - XY \mathbb{E} \left[\frac{(1 - z)^2 (a_h \phi'(y))^2 h^2 (a_r r')^2}{D} \right] \\ & + Y \mathbb{E} \left[\frac{(h - \phi(y))^2 (a_z z')^2}{D} \right] - Y^2 \mathbb{E} \left[\frac{h^2 (a_r r')^2 (h - \phi(y))^2 (a_z z')^2}{D} \right] \end{aligned} \tag{64}$$

$$\begin{aligned} Q = {} & Q \Big\{ \mathbb{E} \left[\frac{(h - \phi(y))^2 (a_z z')^2}{D} \right] - 2Y \mathbb{E} \left[\frac{h^2 (a_r r')^2 (h - \phi(y))^2 (a_z z')^2}{D} \right] \\ & - X \mathbb{E} \left[\frac{(1 - z)^2 (a_h \phi'(y))^2 h^2 (a_r r')^2}{D} \right] \Big\} + \left[\mathbb{E} \left[\frac{|\lambda - z|^2 h^2 (a_r r')^2}{D} \right] + \mathbb{E} \left[\frac{r^2 (h - \phi(y))^2 (a_z z')^2}{D} \right] \right] \end{aligned} \tag{65}$$

Near the boundary of the eigenvalue support, X and Y tend to zero, becoming exactly zero on the boundary. However, Q remains finite. We can use (62) to eliminate Q from (65), and after setting $X = Y = 0$, we obtain the equation for the spectral boundary curve

$$1 = \mathbb{E}\left[h^2(a_r r')^2\right] \mathbb{E}\left[\frac{(a_h\phi'(y))^2\,(1-z)^2}{|\lambda - z|^2}\right] + \mathbb{E}\left[\frac{(a_h\phi'(y))^2 r^2\,(1-z)^2}{|\lambda - z|^2}\right] \tag{66}$$

$$+ \mathbb{E}\left[\frac{(h-\phi(y))^2(a_z z')^2}{|\lambda - z|^2}\right]. \tag{67}$$

After invoking the MFT to factorize certain terms, this simplifies to read

$$1 = \mathbb{E}[(r^2 + a_r^2 h^2 r'^2)]\mathbb{E}\left[a_h^2(\phi'(y))^2\right] \mathbb{E}\left[\frac{(1-z)^2}{|\lambda - z|^2}\right] + \mathbb{E}\left[\frac{a_z^2(h-\phi(y))^2 z'^2}{|\lambda - z|^2}\right]. \tag{68}$$

The spectral boundary will consist of all $\lambda \in \mathbb{C}$ which satisfy this equation.

In the main text, we introduced a boundary curve function

$$\mathcal{S}(\lambda) = \rho^2 \mathbb{E}\left[\frac{(1-z)^2}{|\lambda - z|^2}\right] + a_z^2 \mathbb{E}\left[\frac{(h-\phi(y))^2(z')^2}{|\lambda - z|^2}\right] - 1, \tag{69}$$

which vanishes on the boundary. The difficulty here lies in making sense of the second term. First of all, if the steady state is a fixed point, one must have $h = \phi(y)$, and the second term simply vanishes. However, in general the attractor is not a simple fixed point (since we have shown that the fixed points are unstable), and we cannot expect this term to vanish. The main difficulty with this correlation function is that in general h and z are *not* independent, even in the mean field theory, and they may have non-trivial correlations.

Let us define

$$\mathcal{S}_{\rho,\nu}(\lambda) = \rho^2 \mathbb{E}\left[\frac{(1-z)^2}{|\lambda - z|^2}\right] + \nu^2\, \mathbb{E}\left[\frac{z'^2}{|\lambda - z|^2}\right], \quad \nu^2 = \mathbb{E}\left[(h-\phi(y))^2\right], \tag{70}$$

along with the set

$$\Sigma(\rho,\nu) := \{\lambda \in \mathbb{C} \,|\, \mathcal{S}_{\rho,\nu}(\lambda) \geq 0\}. \tag{71}$$

By Cauchy-Schwarz, we have that $\mathcal{S}(\lambda) \leq \mathcal{S}_{\rho,\nu}(\lambda)$, and by the positivity of the second term on the RHS of (15), it follows that $\mathcal{S}_{\rho,0}(\lambda) \leq \mathcal{S}(\lambda)$. Therefore, we conclude that

$$\Sigma(\rho,0) \subseteq \Sigma \subseteq \Sigma(\rho,\nu), \tag{72}$$

where Σ is the spectral support defined in Thm.(1). Thus we obtain bounding curves for the spectral density described parametrically by $\mathcal{S}(\rho,\nu) = 0$ and $\mathcal{S}(\rho,0) = 0$, both of which are expressible in terms of simple correlation functions.

Obtaining the trace of the resolvent is tractable in limited cases. The Dyson equation for the Green's function proves to be intractable when both reset and update gates have nonzero variance. However, when the reset gate variance $a_r = 0$, the equations simplify considerably and we can determine the trace of the resolvent:

Proposition 13 (Trace of the Resolvent) *For $a_r = 0$, the trace of the resolvent for the GRU instantaneous Jacobian is*

$$G(\lambda) = \begin{cases} \mathbb{E}\left[\dfrac{\bar{\lambda} - z_t}{|\lambda - z_t|^2 + F(\lambda)s_t}\right], & \lambda \in \Sigma \\ \mathbb{E}\left[\dfrac{1}{\lambda - z_t}\right], & \lambda \in \Sigma^c \end{cases} \tag{73}$$

where $s_t = r^2 a_h^2 (1 - z_t)^2 \phi'(y_t)^2 + a_z^2 z_t'^2 (h_{t-1} - \phi(y_t))^2$. For $\lambda \in \Sigma$, $F(\lambda) \geq 0$ is given implicitly by the equation

$$1 = \mathbb{E}\left[\frac{s_t}{|\lambda - z_t|^2 + F(\lambda)s_t}\right], \tag{74}$$

while for $\lambda \in \Sigma^c$, $F = 0$.

The proof follows easily after specializing Eqs.(62 - 65) for $a_r = 0$ in which the reset gate is fixed $r_t = r = \sigma(b_r)$. It is apparent by inspection of Eqs. (63) and (64) that $X = r^2 Y$. Consequently, by introducing the function $F(\lambda) = -(QY + Yr^2)$, we arrive at the statement of the proposition.

Appendix C. GRU: accumulation of eigenvalues at $\lambda = 1$

We provide the details of analytical demonstration of accumulation shown visually in Fig.(1 f) as a function of update gate variance a_z (with reset gate inactive: $a_r = 0$), and provide a derivation of the $a_z \to \infty$ limiting eigenvalue density Eq. 19 presented in the main text. Our starting point is the trace of the resolvent given by Prop. (13).

For $F(\lambda) \geq 0$, we must solve (74) first, then insert this solution back into (73) to determine the trace of the resolvent. Now we make two assumptions: first, we use a piecewise linear approximation for $\phi(y)$ (the "hard-tanh")

$$\phi(x) = \begin{cases} 1, & x > 1 \\ x, & -1 < x < 1 \\ -1, & x < -1 \end{cases}, \quad \phi'(x) = \begin{cases} 0, & x > 1 \\ 1, & -1 < x < 1 \\ 0, & x < -1 \end{cases}, \tag{75}$$

and secondly we assume $P(\zeta, y) = P(\zeta)P(y)$ which is justified in the mean-field limit. Evaluate the Gaussian integral over y using (40) gives

$$G(\lambda) = \eta \mathbb{E}\left[\frac{\bar{\lambda} - z}{|\lambda - z|^2 + F(1-z)^2 r^2 a_h^2}\right]_\zeta + (1 - \eta)\mathbb{E}\left[\frac{1}{\lambda - z}\right] \tag{76}$$

$$1 = \eta \mathbb{E}\left[\frac{(1-z)^2 r^2 a_h^2}{|\lambda - z|^2 + F(1-z)^2 r^2 a_h^2}\right]_\zeta \tag{77}$$

where

$$\eta = \int_{-1}^{1} \frac{1}{\sqrt{2\pi a_h^2 r^2 C_h}} \exp\left(-\frac{y^2}{2a_h^2 r^2 C_h}\right) \tag{78}$$

can be interpretted as the fraction of unsaturated activations $\eta \approx \mathbb{E}\left[(\phi')^2\right]$ (which is exact for the hard-tanh). We are left with a Gaussian integral over ζ, which has the distribution (41)

$$P(\zeta) = \frac{1}{\sqrt{2a_z^2 C_h}} \exp\left(-\frac{(\zeta - b_z)^2}{2a_z^2 C_h}\right). \tag{79}$$

We consider first the limit $a_z = \infty$ to derive Eq.(19), then consider the large a_z limit to obtain a scaling function describing the cumulative distribution function near $\lambda = 1$.

Resolvent for Binary Update Gate For $a_z, b_z \to \infty$ with $b_z/a_z = \beta$ kept fixed, we approximate $z = \sigma(\zeta)$ as a binary random variable with distribution $P(z) = \alpha\delta(z) + (1-\alpha)\delta(z-1)$, where

$$\alpha = 1 - \mathbb{E}[\sigma(\zeta)] = \int Dx \left(1 + \exp\left(-a_z(\sqrt{C_h}x + \beta)\right)\right)^{-1} \approx \frac{1}{2}\mathrm{erfc}\left(\frac{\beta}{\sqrt{2C_h}}\right), \tag{80}$$

which is the expression quoted in (25). The expectation over ζ in (73) and (74) is then trivial, and solving for F we find the trace of the resolvent

$$G(\lambda) = \begin{cases} \dfrac{(1-\alpha)}{\lambda - 1} + \dfrac{\alpha(1-\eta)}{\lambda} + \dfrac{\bar{\lambda}}{r^2 a_h^2}, & |\lambda| \leq \sqrt{\eta\alpha} r a_h \\ \dfrac{(1-\alpha)}{\lambda - 1} + \dfrac{\alpha}{\lambda}, & |\lambda| > \sqrt{\eta\alpha} r a_h. \end{cases} \tag{81}$$

Using $\mu(\lambda) = (1/\pi)\partial_{\bar{\lambda}} G(\lambda)$ then gives Eq.(19). It is especially interesting to note the delta functional concentration of eigenvalues at $\lambda = 1$. Now, we seek to understand how the accumulation grows with increasing a_z, in particular with the goal of obtaining a scaling function to understand the eigenvalue cumulative distribution function (CDF) centered at $\lambda = 1$ displayed in Fig.(1).

Scaling Function for CDF In order to obtain a scaling function to describe the accumulation of eigenvalues at $\lambda = 1$, we make a series of approximations. The good agreement we find with numerical experiments appears to justify these assumptions. First, we assume that we are at or near a fixed point, and thus can neglect terms involving $h - \phi(y)$ in the expression for the resolvent (this is also justified when a_z is large and the update gate is mostly saturated). To get an explicit expression, we use the hard-tanh activation function, so our starting point is again Eqs.(76) and (77).

We would like to obtain the scaling behavior of the cumulative distribution function $CDF(r) = P(|\lambda - 1| < r)$ close to the $\lambda = 1$. With a little foresight, we parameterize the coordinate λ as

$$\lambda = 1 + \epsilon, \quad \epsilon = e^{-\tilde{a}_z c + i\theta}, \quad \tilde{a}_z = a_z\sqrt{C_h}. \tag{82}$$

The CDF follows from the resolvent by the contour integral

$$CDF(|\epsilon|) = \frac{1}{2\pi i}\oint_{\mathcal{C}} d\epsilon\, G(1+\epsilon) = \frac{1}{2\pi}\int d\theta\, \epsilon G(1+\epsilon). \tag{83}$$

We must therefore find an approximation of $G(1+\epsilon)$ for small ϵ. The first observation is that F is not singular at $\lambda = 1$ (though we observe that it is not necessarily smooth), since

$$1 = \eta\mathbb{E}\left[\frac{(1-z)^2 r^2 a_h^2}{(1-z)^2 + F(1)(1-z)^2 r^2 a_h^2}\right] = \frac{\eta r^2 a_h^2}{1 + F(1) r^2 a_h^2}, \tag{84}$$

so we assume for simplicity $F(1+\epsilon) \approx F(1)$. Using this, we evaluate $G(\lambda)$. We may rewrite the resulting Gaussian integral using the fact that when $b_z = 0$, $\sigma(\zeta) = 1 - \sigma(-\zeta)$, and defining $G(\lambda) = G^a(\lambda) + G^b(\lambda)$, we get

$$G^a(\lambda) = \eta \int Dx \frac{e^{-\tilde{a}_z c - i\theta} + 2e^{-\tilde{a}_z(c+x)-i\theta} + e^{-\tilde{a}_z(2x+c)-i\theta} + 1 + e^{-\tilde{a}_z x}}{|e^{-\tilde{a}_z c + i\theta} + e^{-\tilde{a}_z(c+x)+i\theta} + 1|^2 + Fr^2 a_h^2}, \tag{85}$$

$$G^b(\lambda) = (1-\eta) \int Dx \frac{1 + e^{-\tilde{a}_z x}}{\epsilon + 1 + e^{-\tilde{a}_z(x+c)+i\theta}}. \tag{86}$$

We approximate these integrals for large a_z in four regions: I) $x \in (-\infty, -c)$, II) $x \in (-c, -c/2)$, III) $x \in (-c/2, 0)$ and IV) $x \in (0, \infty)$. We begin with G^a. Approximating the integrals by keeping only the dominant terms for large a_z, we get

$$G^a_I(\lambda) \approx \eta \int_{-\infty}^{-c} Dx \frac{e^{-\tilde{a}_z c - i\theta} e^{-2\tilde{a}_z x}}{e^{-2\tilde{a}_z c - 2a_z x}} = \frac{1}{\epsilon}\frac{\eta}{2}\text{erfc}\left(\frac{c}{\sqrt{2}}\right), \tag{87}$$

$$G^a_{II}(\lambda) \approx \eta \frac{\bar{\epsilon}}{|\epsilon+1|^2 + Fr^2 a_h^2} \int_{-c}^{-c/2} Dx e^{-2\tilde{a}_z x}, \tag{88}$$

$$G^a_{III}(\lambda) \approx \eta \frac{1}{|\epsilon+1|^2 + Fr^2 a_h^2} \left(\int_{-c/2}^{0} Dx e^{-\tilde{a}_z x} \right), \tag{89}$$

$$G^a_{IV}(\lambda) \approx \frac{\eta}{2} \frac{1}{1 + Fr^2 a_h^2}. \tag{90}$$

Repeating a similar analysis for $G^b(\lambda)$, we get the approximate expression

$$G^b_I(\lambda) \approx \frac{1}{\epsilon}\frac{(1-\eta)}{2}\text{erfc}\left(\frac{c}{\sqrt{2}}\right), \quad G^b_{II} + G^b_{III} + G^b_{IV} \approx (1-\eta)\frac{1}{\epsilon+1}\int_{-c}^{\infty} Dx e^{-\tilde{a}_z x} \tag{91}$$

Each of these contributions to the trace of the resolvent will produce a different contribution to the CDF: G_{II},G_{III} and G_{IV} will to leading order give a contribution that vanishes as $a_z \to \infty$. In contrast, G_I gives a contribution which does not vanish in this limit - in fact, it grows as $a_z \to \infty$! Therefore, focusing on this contribution we arrive at the CDF

$$P(|\lambda - 1| < e^{-\tilde{a}_z c}) = \frac{1}{2}\text{erfc}\left(\frac{c}{\sqrt{2}}\right). \tag{92}$$

Writing $r = e^{-a_z\sqrt{C_h}c}$, and $c = -\log(r)/a_z\sqrt{C_h}$, we get the cumulative distribution function

$$P(|\lambda - 1| < r) \approx \frac{1}{2}\text{erfc}\left(-\frac{\log(r)}{a_z\sqrt{2C_h}}\right). \tag{93}$$

This motivates the scaling function we plot in Figs.(1 f) and (3 f),

$$\text{CDF}(r) = c_1 \text{erfc}\left(c_2/a_z\right), \tag{94}$$

treating c_1 and c_2 as fitting parameters. This function appears to capture well the scaling behavior of the CDF with a_z in a small ball around $\lambda = 1$, even at small a_z where the derivation does not strictly make sense. This derivation shows that the dominant contribution to the CDF comes, quite naturally, from the domain of ζ over which the update gate $\sigma(\zeta) \approx 1$.

C.1. Proof of Theorem (7) on Spectral Pinching

Here we present a proof of the spectral pinching announced in Thm.(7). We start with the equation for the spectral boundary curve for fixed points, i.e. Eq.(15) with the second term set to zero, or equivalently $\mathcal{S}_{\rho,0}$ defined in (70). We propose an ansatz for the curve $\lambda(s)$ close to the leading edge

$$\lambda = 1 + \lambda_0 e^{-c\tilde{a}_z}, \quad \tilde{a}_z = \sqrt{C_h} a_z, \tag{95}$$

where $\lambda_0 \in \mathbb{C}$ is a complex constant, and $c \in \mathbb{R}^+$ is an order one real constant. We consider the large a_z, b_z limit while keeping C_h fixed, and the ratio $b_z/a_z = \beta$ also fixed. For convenience, let $\tilde{\beta} = \beta/\sqrt{C_h}$

We proceed now to constrain c via the support equation $\mathcal{S}_{\rho,0}(\lambda) = 1$. For large a_z we may approximate this boundary curve equation

$$1 = \rho^2 \int Dx \frac{(1-\sigma(\tilde{a}_z(x+\tilde{\beta})))^2}{|\lambda - \sigma(\tilde{a}_z(x+\tilde{\beta}))|^2} = \rho^2 \int Dx \frac{1}{|1+\lambda_0 e^{-c\tilde{a}_z}(1+e^{\tilde{a}_z(x+\tilde{\beta})})|^2} \tag{96}$$

$$\approx \rho^2 \left\{ \int_{-\infty}^{c-\tilde{\beta}} Dx \frac{1}{|1+\lambda_0 e^{-c\tilde{a}_z}|^2} + \frac{1}{|\lambda_0|^2} \int_{c-\tilde{\beta}}^{\infty} Dx\, e^{2\tilde{a}_z(c-\tilde{\beta}-x)} \right\} \tag{97}$$

$$\approx \rho^2 \frac{1}{2|1+\lambda_0 e^{-c\tilde{a}_z}|^2} \mathrm{erfc}\left(-\frac{(c-\tilde{\beta})}{\sqrt{2}} \right) + \frac{\rho^2}{2|\lambda_0|^2} e^{2\tilde{a}_z(\tilde{a}_z - \tilde{\beta} + c)} \mathrm{erfc}\left(\frac{2\tilde{a}_z - \tilde{\beta} + c}{\sqrt{2}} \right) \tag{98}$$

Now sending $a_z \to \infty$, we recover the implicit equation for c quoted in (24).

Appendix D. Spectral Radius via Gelfand's Formula and Proof of Proposition (5)

To understand the role of the reset gate, we set $a_z = 0$ and fix b_z. We want to show that the spectral radius will grow with increasing a_r. This could be calculated using our formula for the spectral curve, but here we provide an alternate approach using Gelfand's formula and take an average over the random weights. The Gelfand formula states that the spectral radius is given by the limit

$$\rho(\mathbf{J}_t) = \lim_{n\to\infty} ||\mathbf{J}_t^n||^{1/n} \tag{99}$$

for any matrix norm $||\cdot||$. We make a self-averaging assumption by taking

$$\left\langle \lim_{n\to\infty} ||\mathbf{J}_t^n||^{1/n} \right\rangle_U \approx \lim_{n\to\infty} \left(\langle ||\mathbf{J}_t^n|| \rangle_U \right)^{1/n}, \tag{100}$$

where $\langle ... \rangle_U$ indicates averaging over random connectivity matrices. In words, this assumption states that the spectral radius is stable between random realizations of the weights. Since the Gelfand formula is valid for any matrix norm, we resort to the sum of the singular values of J_t^n : μ_n

$$\mu_n = \left\langle \frac{1}{N} \mathrm{tr}(\mathbf{J}_n^T \mathbf{J}_n) \right\rangle. \tag{101}$$

With this, we may then use the results of Appendix F to find that in the limit of large N, the spectral radius of the Jacobian is given by

$$\langle \rho(\mathbf{J}_t) \rangle = \lim_{n \to \infty} (\mu_n)^{\frac{1}{2n}} = \sigma(b_z) + (1 - \sigma(b_z))\rho_t. \tag{102}$$

This expression also follows immediately from the equation for the spectral support when $a_z = 0$

$$|\lambda - \sigma(b_z)|^2 = \rho_t^2 (1 - \sigma(b_z))^2. \tag{103}$$

Through its effect on ρ_t, the reset gate will directly influence the spectral radius. To see this, recall again the expression for ρ_t

$$\rho_t^2 = a_h^2 C_{\phi'} \left(C_r + a_r^2 C_{r'} C_h \right), \tag{104}$$

where

$$C_{\phi'} = \int \left(\phi' \left(\sqrt{C_y} x \right) \right)^2 Dx, \quad C_y = C_r C_h \tag{105}$$

$$C_r = \int Dx \left(\sigma(a_r \sqrt{C_h} x) \right)^2, \quad C_{r'} = \int Dx \left(\sigma'(a_r \sqrt{C_h} x) \right)^2. \tag{106}$$

We have taken $b_r = 0$ for simplicity. We would like to examine the growth of ρ_t as a function of a_r at fixed C_h. The mean-field theory shows that since C_h only depends on a_r through the reset gate, it is not possible that the steady state distribution C_h will grow without bound. Thus, C_h must be bounded above and below by constant which do not grow with a_r. Next, C_r is a bounded function of a_r, and for fixed C_h $\frac{1}{4} \leq C_r \leq \frac{1}{2}$. Therefore, we may bound the spectral shaping parameter from below

$$a_h^2 C_{\phi'} \left(\frac{1}{4} + a_r^2 C_{r'} C_h \right) \leq \rho_t^2 \leq a_h^2 C_{\phi'} \left(\frac{1}{2} + a_r^2 C_{r'} C_h \right). \tag{107}$$

Furthermore, since $C_y = C_r C_h$, we have that C_y is similarly bounded for fixed C_h. An upper-bound on C_y implies a non-zero lower bound on $C_{\phi'}$ which we call c: $c \leq C_{\phi'} \leq 1$. Finally, we come to $C_{r'}$, which tends to zero with increasing a_r. Thus, it remains to show that this tendency does not overwhelm the a_r^2 prefactor. To see this, we may develop an asymptotic expansion of the integral

$$C_{r'} = \int Dx \left(\sigma'(a_r \sqrt{C_h} x) \right)^2 = \frac{1}{6 a_r \sqrt{2\pi C_h}} + O(a_r^{-2}). \tag{108}$$

Combining this with the bounds we have argued for, we get

$$a_h^2 c \left(\frac{1}{4} + \frac{a_r}{6\sqrt{2\pi}} \sqrt{C_h} \right) \leq \rho_t \leq a_h^2 \left(\frac{1}{4} + \frac{a_r}{6\sqrt{2\pi}} \sqrt{C_h} \right), \tag{109}$$

and thus $\rho_t = \Theta(\sqrt{a_r})$.

Appendix E. GRU Fixed Point Phase diagram

Here we describe the fixed point distributions implied by the implicit equations (10). We assume $v_h = v_r = b_h = 0$, but keep nonzero b_r. For fixed a_h and a_r, and $a_r^2 C_y << 1$, we expand the correlation functions

$$C_\phi = a_h^2 C_y - 2a_h^4 C_y^2 + \frac{17}{3} a_h^6 C_y^3 + O(C_y^4), \quad C_r = c_1 + c_2 a_r^2 C_h + c_3 a_r^4 C_h^2 + O(C_h^3), \tag{110}$$

$$c_1 = \sigma(b_r)^2, \quad c_2 = -\frac{(e^{-b_r} - 2e^{-2b_r})}{(1+e^{-b_r})^4}, \quad c_3 = -\frac{e^{2b_r}(e^{3b_r} - 18e^{2b_r} + 33e^{b_r} - 8)}{4(1+e^{b_r})^6}. \tag{111}$$

Then using $C_y = C_h C_r$, the implicit equation (10) becomes

$$C_h = a_h^2 c_1 C_h + \left(a_h^2 a_r^2 c_2 - 2a_h^4 c_1^2\right) C_h^2 + \left(c_3 a_h^2 a_r^4 - 4c_1 c_2 a_h^4 a_r^2 + \frac{17}{3} a_h^6 c_1^3\right) C_h^3 + O(C_h^4). \tag{112}$$

Setting $b_r = 0$ gives

$$1 = \frac{a_h^2}{4} + \frac{a_h^2}{16}\left(a_r^2 - 2a_h^2\right) C_h + \frac{a_h^2}{192}(17a_h^4 - 12a_h^2 a_r^2 - 6a_r^4) C_h^2 + O(C_h^3). \tag{113}$$

A straightforward analysis of this expression leads to the perturbative solutions quoted in the main text, with the function

$$f(a_r) = 4(31552 - 11424a_r^2 + 744a_r^4 + 72a_r^6 + 9a_r^8)/(3a_r^4 + 24a_r^2 - 136)^2, \tag{114}$$

which is positive for $a_r > \sqrt{8}$.

We comment briefly on the effects of finite bias b_r. The critical line past which only a single nonzero fixed point exists moves to $a_h = 1 + e^{-b_r}$. For a finite bias, there is still only a single zero fixed point for $a_h < \sqrt{2}$.

Appendix F. GRU : Singular values of the long-term Jacobian

In this section, we provide results on the moments of the singular values of the long-term Jacobian for fixed points in two limiting cases.

Binary Update Gate We first consider the Jacobian for fixed points with update gates which are switch-like ($a_z \to \infty$). In this case, $\mathbf{z}_t$ is vectors of binary variables with distribution $P(z) = \alpha\delta(z) + (1-\alpha)\delta(z-1)$, where α depends on b_z (see Eq.(25)) and $\mathbf{z}'_t = 0$, so that the Jacobian is

$$\mathbf{J} = \hat{\mathbf{z}} + (\mathbb{1} - \hat{\mathbf{z}})\hat{\phi}'(\mathbf{y}) U_h (\hat{r} + \hat{r}' \hat{h} U_r). \tag{115}$$

Since $\hat{\mathbf{z}}(\mathbb{1} - \hat{\mathbf{z}}) = 0$, and $\mathbf{z}^k = \mathbf{z}$, we have for the late-time Jacobian

$$\mathbf{J}_n = \mathbf{J}^n = \sum_{q=0}^{n} \left((\mathbb{1} - \hat{\mathbf{z}})\hat{\phi}'(\mathbf{y}) U_h (\hat{r} + \hat{r}' \hat{h} U_r)\right)^q \hat{\mathbf{z}}. \tag{116}$$

This allows us to obtain the mean square of the singular values to leading order in N

$$\mu_n = \left\langle \frac{1}{N} \mathrm{tr}(\mathbf{J}_n^T \mathbf{J}_n) \right\rangle = (1-\alpha) \sum_{k=0}^{n} \left(\alpha \rho^2\right)^k + O(N^{-1}). \tag{117}$$

Here, we have also taken the expectation over z, using for instance,

$$\frac{1}{N} \mathrm{tr}\left((\mathbb{1} - \hat{\mathbf{z}})^2 (\hat{\phi}'(\mathbf{y}))^2 \right) \approx \mathbb{E}[(1-z)^2 (\phi'(\eta))^2] = \alpha \mathbb{E}[(\phi'(\eta))^2]. \tag{118}$$

For $\alpha \rho^2 < 1$, the moment converges to

$$\mu_n \to \frac{(1-\alpha)}{1-\alpha\rho^2}. \tag{119}$$

Constant Update Gate In the limit of zero a_z, the behavior is quite different. Then $\hat{\mathbf{z}}$ becomes a constant diagonal matrix with elements $\sigma(b_z)$. The late time Jacobian is

$$\mathbf{J}_n = \sum_{q=0}^{n} \binom{n}{q} \hat{\mathbf{z}}^{n-q} \left((1-\hat{\mathbf{z}}) \hat{\phi}'(\mathbf{y}) U_h \left(\hat{\mathbf{r}} + \hat{\mathbf{r}}' \hat{\mathbf{h}} U_r \right) \right)^q. \tag{120}$$

The mean squared singular values can be computed again in the limit of large N to yield

$$\mu_n = \sum_{q=0}^{n} \binom{n}{q}^2 \sigma(b_z)^{2(n-q)} (1-\sigma(b_z))^{2q} \rho^{2q}, \quad \rho^2 = a_h^2 \sigma(b_r) \mathbb{E}[(\phi(y)')^2] \left(\mathbb{E}[r^2] + a_r^2 \mathbb{E}[(r')^2 h^2] \right). \tag{121}$$

Formally, this can be expressed in terms of the Gauss hypergeometric function

$$\mu_n = \sigma(b_z)^{2n} {}_1F_2\left(-n, -n; -1; \frac{(1-\sigma(b_z))^2 \rho^2}{\sigma(b_z)^2} \right). \tag{122}$$

Another representation which is useful for asymptotic analysis is

$$\mu_n = \int_0^{2\pi} \frac{d\theta}{2\pi} \left| \sigma(b_z) + (1-\sigma(b_z)) \rho e^{i\theta} \right|^{2n}. \tag{123}$$

This can be evaluated by a saddle-point argument for large n to give the spectral radius in (102).

Appendix G. LSTM: Proof of Theorem (8)

Linearization of the LSTM Jacobian Eq.(26) is achieved by (dropping the carat $\hat{}$ with the understanding that all boldfaced variables are diagonal matrices)

$$\mathbf{M}_t = \begin{pmatrix} \mathbf{f}_t & 0 & \mathbf{i}_t\phi'(\mathbf{y}_t) & 0 & \phi(\mathbf{y}_t) & \mathbf{c}_{t-1} \\ \mathbf{m}_t\mathbf{f}_t & 0 & \mathbf{m}_t\mathbf{i}_t\phi'(\mathbf{y}_t) & \phi(\mathbf{c}_t) & \mathbf{m}_t\phi(\mathbf{y}_t) & \mathbf{m}_t\mathbf{c}_{t-1} \\ 0 & U_h & 0 & 0 & 0 & 0 \\ 0 & \mathbf{o}_t'U_o & 0 & 0 & 0 & 0 \\ 0 & \mathbf{i}_t'U_i & 0 & 0 & 0 & 0 \\ 0 & \mathbf{f}_t'U_f & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{124}$$

We express this as $\mathbf{M}_t = \hat{A} + \hat{B}\hat{U}\hat{C}$, where

$$\hat{A} = \begin{pmatrix} \mathbf{f}_t & 0 & \mathbf{i}_t\phi'(\mathbf{y}_t) & 0 & \phi(\mathbf{y}_t) & \mathbf{c}_{t-1} \\ \mathbf{m}_t\mathbf{f}_t & 0 & \mathbf{m}_t\mathbf{i}_t\phi'(\mathbf{y}_t) & \phi(\mathbf{c}_t) & \mathbf{m}_t\phi(\mathbf{y}_t) & \mathbf{m}_t\mathbf{c}_{t-1} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad \hat{C} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathbb{1} & 0 & 0 & 0 & 0 \\ 0 & \mathbb{1} & 0 & 0 & 0 & 0 \\ 0 & \mathbb{1} & 0 & 0 & 0 & 0 \\ 0 & \mathbb{1} & 0 & 0 & 0 & 0 \end{pmatrix},$$

$\hat{U} = \mathrm{bdiag}\,(0, 0, U_h, U_o, U_i, U_f)$, and $\hat{B} = \mathrm{bdiag}\,(0, 0, \mathbb{1}, \mathbf{o}_t', \mathbf{i}_t', \mathbf{f}_t')$. We denote by $\mathrm{bdiag}(A_1, ..., A_n)$ a block diagonal matrix whose blocks A_i are $N \times N$ matrices.

The generalized eigenvalue problem is $\mathbf{M}_t\mathbf{v} = I_{\lambda,2}\mathbf{v}$, where $I_{\lambda,2} = \mathrm{bdiag}(\lambda, \lambda, 1, 1, 1, 1)$ and λ are eigenvalues of the Jacobian $\mathbf{J}_t$. As before, we define the generalized resolvent

$$\mathbf{G} = (I_{\lambda,2} - \mathbf{M}_t)^{-1}, \tag{125}$$

which is a $6N \times 6N$ matrix. We describe this matrix by $N \times N$ subblocks $\mathbf{G}_{ab}$, for $a, b = 1, .., 6$. The resolvent of the Jacobian is then found in the upper left corner of the generalized resolvent

$$G = (\lambda - \mathbf{J}_t)^{-1} = \begin{pmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{pmatrix} \tag{126}$$

Following exactly the hermitization procedure outlined in Sec. (B), we define $F(\lambda, \bar{\lambda}) = f_{99}(f_{33} + f_{44} + f_{55} + f_{66})$, and obtain the corresponding set of equations from the Dyson equation. To express these in compact form, let

$$\mathbf{Q}_t(\lambda) = |\lambda - \mathbf{f}_t|^2\mathbf{q}_t + |\lambda|^2\mathbf{p}_t, \quad \text{where} \tag{127}$$

$$\mathbf{q}_t = a_o^2(\mathbf{o}_t')^2\phi(\mathbf{c}_t)^2, \quad \mathbf{p}_t = \mathbf{o}_t^2(\phi'(\mathbf{c}_t))^2\left(a_f^2\mathbf{c}_{t-1}^2(\mathbf{f}_t')^2 + a_i^2(\mathbf{i}_t')^2\phi(\mathbf{y}_t)^2 + a_h^2\mathbf{i}_t^2(\phi'(\mathbf{y}_t))^2\right) \tag{128}$$

Then we find the set of equations for the trace of the resolvent $G(\lambda)$ and the auxiliary function $F(\lambda)$

$$G(\lambda) = \frac{1}{N}\text{tr}\left[\frac{|\lambda|^2(\bar{\lambda}-\hat{\mathbf{f}}_t)+\bar{\lambda}|\lambda-\hat{\mathbf{f}}_t|^2+F(\lambda,\bar{\lambda})\partial_\lambda\hat{\mathbf{Q}}_t(\lambda,\bar{\lambda})}{|\lambda|^2|\lambda-\hat{\mathbf{f}}_t|^2+F(\lambda,\bar{\lambda})\hat{\mathbf{Q}}_t(\lambda,\bar{\lambda})}\right]F(\lambda) \quad = \frac{1}{N}\text{tr}\left[\frac{F(\lambda)\hat{\mathbf{Q}}_t(\lambda)}{|\lambda|^2|\lambda-\hat{\mathbf{f}}_t|^2+F(\lambda)\hat{\mathbf{Q}}_t(\lambda)}\right] \tag{129}$$

Outside the support of the eigenvalue density, the resolvent is holomorphic. We see that a consistent solution to these equations has $F(\lambda) = 0$, in which case $G(\lambda)$ simplies to the result quoted in (27), which is holomorphic. We conclude that $F = 0$ defines the exterior of the support. Inside the spectral support, we find the implicit equation can be solved for $F > 0$, which leads to (29). Continuity of $G(\lambda)$ implies that at the boundary of the spectral support, the holomorphic solution must match the non-analytic solution. The curve for the boundary support then follows by considering (129) λ approaching the boundary from the interior. This limit allows us to divide through by F first, and then set $F = 0$ to finally find the boundary curve in Corollary (9). So far we have not assumed the mean field theory, and keep an explicit sum over neurons. In the mean field theory, the neurons (hidden and cell states) describe independent stochastic processes, and allows us to replace the trace with an expectation value over the effective stochastic variables for each site.

We use the mean field theory now to describe the density of eigenvalues which is obtained from the resolvent. Using the definition (12), we find

$$\mu(\lambda) = \frac{1}{\pi}\mathbb{E}\left[\frac{F(\lambda)(|\lambda-f_t|^4q_t+|\lambda|^4p_t)-\partial_{\bar{\lambda}}F(\lambda)\left(\bar{\lambda}|\lambda-f_t|^4q_t+(\bar{\lambda}-f_t)|\lambda|^4p_t\right)+F(\lambda)^2f_t^2p_tq_t}{\left(|\lambda|^2|\lambda-f_t|^2+F(\lambda)Q_t\right)^2}\right].$$

To find $\partial_{\bar{\lambda}}F$, we differentiate the implicit equation for F,

$$\partial_{\bar{\lambda}}F = -\mathbb{E}\left[\frac{\lambda|\lambda-f_t|^4q_t+(\lambda-f_t)|\lambda|^4p_t}{\left(|\lambda|^2|\lambda-f_t|^2+FQ_t\right)^2}\right]\left(\mathbb{E}\left[\frac{Q_t^2}{\left(|\lambda|^2|\lambda-f_t|^2+FQ_t\right)^2}\right]\right)^{-1}$$

G.1. Proof of Corollaries (10) and (11): Biases Only

The equations all simplify considerably when $a_k = 0$ for $k \in \{f, i, o\}$. In this case, $k_t = k = \sigma(b_k)$. We get $p_t = a_h^2o^2i^2(\phi'(c_t))^2(\phi'(y_t))^2$, the spectral support

$$\Sigma := \{\lambda \in \mathbb{C} : |\lambda - f|^2 = \mathbb{E}\,[p_t]\}, \tag{130}$$

and the resolvent inside the support

$$G(\lambda) = \frac{1}{\lambda} + \mathbb{E}\left[\frac{(\bar{\lambda}-f)}{|\lambda-f|^2+F(\lambda)p_t}\right] \tag{131}$$

$$1 = \mathbb{E}\left[\frac{p_t}{|\lambda-f|^2+F(\lambda)p_t}\right]. \tag{132}$$

These become particular simple at the zero FP when $c = h = 0$. In this case, we have $p = \sigma(b_o)^2\sigma(b_i)^2a_h^2$, and straightforward substitution into the equations above gives Corollaries (10) and (11).

G.2. Only Forget

Here we examine the effects of the forget gate by setting $a_o = a_i = 0$. This gives $q_t = 0$, and $p_t = o_t^2(\phi'(c_t))^2\left(a_f^2 c_{t-1}^2 (f_t')^2 + a_h^2 i^2 (\phi'(y_t))^2\right)$, and the resolvent inside the spectral support

$$G(\lambda) = \frac{1}{\lambda} + \mathbb{E}\left[\frac{(\bar{\lambda} - f_t)}{|\lambda - f_t|^2 + F(\lambda)p_t}\right] \tag{133}$$

$$1 = \mathbb{E}\left[\frac{p_t}{|\lambda - f_t|^2 + Fp_t}\right]. \tag{134}$$

The spectral boundary curve is

$$1 = \mathbb{E}\left[\frac{o_t^2(\phi'(c_t))^2\left(a_f^2 c_{t-1}^2 (f_t')^2 + a_h^2 i^2 (\phi'(y_t))^2\right)}{|\lambda - f_t|^2}\right] \tag{135}$$

In the limit when $a_f = \infty$, we must take care with taking expectation values since in the mean-field limit, the distributions of c and f are not separable. We proceed with reasonable arguments. To arrive at explicit formulas, we assume a hard-tanh activation function ϕ. When $a_f = \infty$, f becomes a binary variable. We may then consider two cases

$$c_t \sim \begin{cases} c_{t-1} + i_t\phi(y_t), & f_t = 1 \\ i_t\phi(y_t), & f_t = 0 \end{cases} \tag{136}$$

In the first instance, c_t exhibits fast growth that overwhelms the finite noise term $\phi(y_t)$, and in this case we approximate the distributions of c_t and y_t as independent. In the latter case, c_t is controlled exclusively by the fluctuations of $\phi(y_t)$, and far from being independent, they are slaved. Defining for ease of notation $\tilde{F} = a_h^2\sigma(b_0)^2\sigma(b_i)^2 F$, we get the implicit equation for F:

$$1 = \frac{1}{2}\mathbb{E}\left[\frac{a_h^2 o^2 i^2 \phi'(c)^2\phi'(y)^2}{|\lambda|^2 + \tilde{F}\phi'(c_t)^2\phi'(y_t)^2}\right]_{P(c,y|f=0)} + \frac{1}{2}\mathbb{E}\left[\frac{a_h^2 o^2 i^2 \phi'(c_t)^2\phi'(y_t)^2}{|\lambda - 1|^2 + \tilde{F}\phi'(c_t)^2\phi'(y_t)^2}\right]_{P(c,y|f=1)} \tag{137}$$

When $f = 1$, we assume $\phi'(c) = 0$, since the activation saturates for large c. Therefore, we are left with the first case, when $c_t \sim \sigma(b_i)\phi(y)$ and use the Gaussianity of y_t to evaluate

$$1 = \frac{1}{2}\mathbb{E}\left[\frac{a_h^2 o^2 i^2 [\phi'(\sigma(b_i)\phi(y))]^2\phi'(y)^2}{|\lambda|^2 + \tilde{F}[\phi'(\sigma(b_i)\phi(y))]^2\phi'(y)^2}\right], \tag{138}$$

$$= \frac{o^2 i^2 a_h^2 \eta}{2}\frac{1}{|\lambda|^2 + i^2 o^2 a_h^2 F}, \quad \eta = \int_{-1/a_h}^{1/a_h} \frac{e^{-y^2/2C_h}}{\sqrt{2\pi C_h}}. \tag{139}$$

Using this we may evaluate the resolvent

$$G(\lambda) = \frac{1}{\lambda} + \frac{1}{2}\mathbb{E}\left[\frac{\bar{\lambda}}{|\lambda|^2 + \tilde{F}\phi'(c_t)^2\phi'(y_t)^2}\right]_{P(c,y|f=0)} + \frac{1}{2}\mathbb{E}\left[\frac{\bar{\lambda}-1}{|\lambda-1|^2 + \tilde{F}\phi'(c_t)^2\phi'(y_t)^2}\right]_{P(c,y|f=1)} \tag{140}$$

$$\approx \frac{1}{\lambda} + \frac{(1-\eta)}{2}\frac{1}{\lambda} + \frac{\bar{\lambda}}{o^2 i^2 a_h^2} + \frac{1}{2}\frac{1}{\lambda - 1}. \tag{141}$$

Then using Eq.(12) results in Eq. (34).

Effects on Spectral Radius The previous section assumes switch-like forget gate. Here we relax this assumption and consider the growth of the spectral radius with a_f. We focus on the term involving gradients of f_t in the equation for the spectral boundary curve (135). Since f_t' will be sharply peak around $\eta_t^f = 0$ (for zero bias), we treat it like a delta functional and find this contribution behaves approximately like

$$\frac{a_f}{6\sqrt{2\pi C_h}}\frac{1}{|\lambda - 1/2|^2}\mathbb{E}\left[(\phi'(c_t)^2)c_{t-1}^2\right]_{P(c,y|f=1/2)} \tag{142}$$

When $f = 1/2$, the update equation for the cell state is $c_t = (1/2)c_{t-1} + i\phi(y_t)$, and due to the fast decay of c_t we argue that $\phi'(c_t) \approx 1$. Combining this with the results in the previous section gives for the spectral boundary curve

$$1 = \frac{a_f \mathbb{E}\left[c_{t-1}^2\right]_{P(c|f=1/2)}}{6\sqrt{2\pi C_h}|\lambda - 1/2|^2} + \frac{o^2 i^2 a_h^2 \eta}{2|\lambda|^2} \tag{143}$$

Assuming of course that $C_h > 0$, which is true for networks with nonzero activity. We also need to make some reasonable assumption about the behavior of the correlation function which appears in this expression. For large a_f, the cell state will generally grow without bound for $f \approx 1$. However, conditioned on $f = 1/2$, we may argue that the autocorrelation remains bounded with increasing a_f. Together, these assumptions show that for increasing a_f, the first term takes over and the spectral radius grows like $\sqrt{a_f}$.

This poses an apparent contradiction: the spectral radius would appear to grow with a_f, yet in the limit $a_f = \infty$ it is independent of a_f according to (34). This likely occurs in the following way: the conditional probability used to evaluate the autocorrelation function must approach measure zero as $a_f \to \infty$. Therefore, we conclude that the spectral radius does not grow monotonically with a_f. In Fig.(3), we manage to capture this initial growth, which we conjecture is followed by a suppression for sufficiently large a_f.

G.3. Only Input

Now we isolate the input gate's effect on the Jacobian by setting $a_f = a_o = 0$, which makes $q_t = 0$, and $p_t = o^2\phi'(c_t)^2\left(a_i^2 i_t'^2\phi(y_t)^2 + a_h^2 i_t^2\phi'(y_t)^2\right)$. using Cor. (9) to write spectral boundary curve, we may immediately extract the spectral radius

$$\rho(\mathbf{J}_t) = \sigma(b_f) + \sqrt{\mathbb{E}\left[p_t\right]}, \quad p_t = o^2(\phi'(c_t))^2\left(a_i^2(i_t')^2\phi(y_t)^2 + a_h^2 i_t^2(\phi'(y_t))^2\right). \tag{144}$$

To obtain the asymptotic scaling of the spectral radius with increasing a_i, we may use the Cauchy-Schwartz inequality to establish the bound

$$\mathbb{E}\left[p_t\right] \leq a_i^2 o^2 \mathbb{E}\left[(i_t')^2\right] \mathbb{E}\left[(\phi'(c_t))^2 \phi(y_t)^2\right] + a_h^2 o^2 \mathbb{E}\left[i_t^2\right] \mathbb{E}\left[(\phi'(c_t))^2 \phi'(y_t)^2\right]. \tag{145}$$

First of all, $\phi'(c_t)$, $\phi(y_t)$ and $\phi'(y_t)$ are all bounded functions, which implies that their correlations cannot grow with a_i. Furthermore, $\mathbb{E}[i_t^2] \leq 1/2$, so we are left with $\mathbb{E}[(i_t')^2]$. Using the same arguments as in Sec. D, we find

$$\mathbb{E}\left[(i_t')^2\right] = \frac{1}{6a_i\sqrt{2\pi C_h}} + O(a_i^{-2}). \tag{146}$$

Combining all of this, we arrive at our main result

$$\rho(\mathbf{J}_t) = O(\sqrt{a_i}). \tag{147}$$

A further assumption that $\mathbb{E}\left[(i_t')^2 \phi'(c_t)^2\right] = \Theta\left(\mathbb{E}\left[(i_t')^2\right] \mathbb{E}\left[\phi'(c_t)^2\right]\right)$ would allow us to also obtain a lower bound that would imply $\rho(\mathbf{J}_t) = \Theta(\sqrt{a_i})$. Our numerical experiments indicate that such a lower bound should hold.

G.4. Only Output

Here we study the effects of the output gate on the spectral radius. Fixing the other gates by setting $a_f = a_i = 0$, using Cor. (9) the spectral boundary curve becomes

$$1 = \frac{\mathbb{E}[q_t]}{|\lambda|^2} + \frac{\mathbb{E}[p_t]}{|\lambda - \sigma(b_f)|^2}, \quad q_t = a_o^2 (o_t')^2 \phi(c_t)^2, \quad p_t = a_h^2 o_t^2 i^2 (\phi'(c_t))^2 (\phi'(y_t))^2 \tag{148}$$

Since o_t^2 is bounded, p_t will remain bounded as a_o is increased. We can thus make the obvious bounds

$$0 \leq \mathbb{E}\left[p_t\right] \leq \frac{1}{2}\sigma(b_i)^2 a_h^2, \tag{149}$$

which is just to show that this factor will not grow with a_o. Next, note that in the mean field limit for the LSTM network, the cell state c_t is *independent* of o_t, and therefore we may write

$$\mathbb{E}\left[q_t\right] = a_o^2 \mathbb{E}[(o_t')^2] \mathbb{E}\left[(\phi(c_t)^2\right] \tag{150}$$

Now following the same arguments as in Sec. D, we evaluate the correlation function

$$\mathbb{E}[(o_t')^2] = \frac{1}{6a_o\sqrt{2\pi C_h}} + O(a_o^{-2}). \tag{151}$$

Therefore, we find

$$\mathbb{E}[q_t] \sim a_o \frac{\mathbb{E}\left[\phi(c_t)^2\right]}{6\sqrt{2\pi C_h}} \tag{152}$$

Which means that for large a_o, the spectral radius grows like $\rho(\mathbf{J}_t) = \Theta(\sqrt{a_o})$.

Appendix H. GRU : Training on sequential MNIST

Here we provide details of training a GRU on the sequential MNIST task Jing et al. (2019); Kerg et al. (2019). Each 28x28 MNIST image is fed into a 256 dimensional GRU one pixel per timestep. The output of the GRU is used to decide the digit class at the end of the 784 timesteps. We also use zero biases for the GRU to make the training closer to the theory we have worked out – although incorporating biases in the theory is a trivial extension. The batch size chosen was 100 (for 60000 images in the training set), and the maximum number of epochs was taken to be 15. For ease of discussion, we show results for a low value of $a_r = 0.1$.

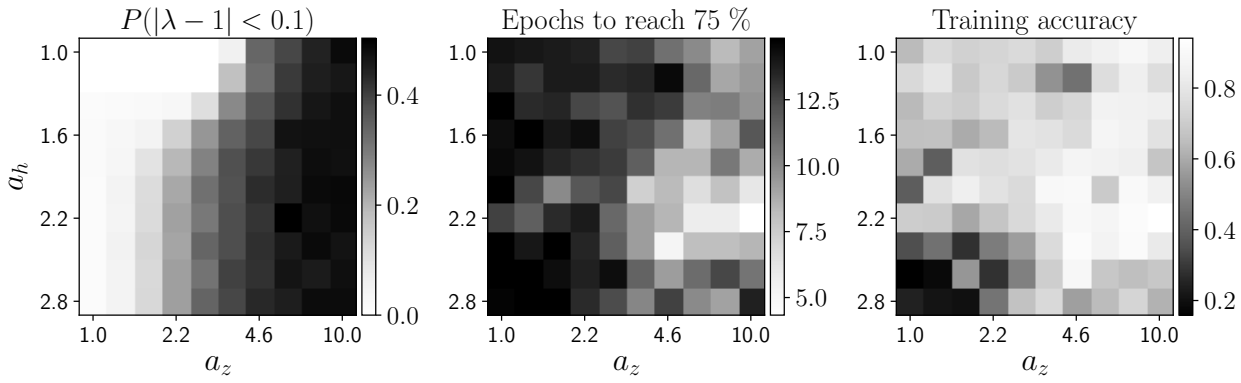

Figure 4: Training on sequential MNIST: (**left**) The density of eigenvalues of the Jacobian near 1.0 as a function of a_h and a_z. (**middle**) The number of epochs required to reach a training accuracy of 75 % as a function of a_h and a_z. (**right**) The training accuracy after 15 epochs as a function of a_h and a_z.

Fig. 4 shows the i) density of Jacobian eigenvalues near 1.0 (left); ii) the average number of epochs required to reach a training accuracy of 75 % (middle) and iii) the average training accuracy (right) as a function of a_h and a_z. As we can see in Fig 4 (left) increasing a_z leads to an increase in the density of eigenvalues near 1.0. The training accuracy and the number of epochs needed to reach 75% accuracy both improve with increasing a_z, however, the best values – especially for the training time – seem to be near $a_h = 2.0$ which is the critical value at which the zero fixed-point becomes unstable; benefits of training at the "edge-of-chaos" has been noted in previous work Bertschinger and Natschl (2004); Glorot and Bengio (2010). Training appears to be harder once we get further into the chaotic regime. This is likely related to the rate of growth of gradients in this regime, and will be investigated elsewhere.