Dynamics of continuous-time gated recurrent neural networks

Tankut Can *
Initiative for the Theoretical Sciences
The Graduate Center, CUNY
New York, NY

Kamesh Krishnamurthy *
Joseph Henry Laboratories for Physics
Princeton University
Princeton, NJ

David J. Schwab
Initiative for the Theoretical Sciences
The Graduate Center, CUNY
New York, NY

Abstract

We study how gates shape the dynamics of a continuous-time gated recurrent network, closely related to the Gated Recurrent Unit (GRU). As a function of the initialization hyperparameters, we map out the phase diagram of the gated RNN showing it can exhibit a range of rich phenomena. In addition to highlighting the different dynamical phases, this phase diagram provides a principled map for initialization choices. We show that gating can robustly produce slow modes and line attractor dynamics – a mechanism useful for tasks involving long time dependencies; furthermore, gating can also lead to a first-order (discontinuous) transition to chaos, challenging the usual heuristic of initializing at the "edge of chaos". [2]

1 Introduction

Modern recurrent neural networks (RNNs) are able to learn complex sequence processing tasks in large part due to the use of *gating*. This architectural innovation, pioneered with the Long Short-Term Memory (LSTM) in [5] and its various variants, was intuitively argued to provide an efficient solution to the exploding and vanishing gradients problem, allowing RNNs to be trained by powerful gradient-based methods.

Despite their popularity in RNNs, a more systematic study of how gates shape the dynamics and gradients has been lacking. In particular, each of the gates has additional hyperparameters, making the space of hyperparameters larger and more challenging to navigate. One particularly important set of hyperparameters is the mean and variance of the distribution used for the initialization of weight matrices [9]. To date, there is no principled method to inform the choices of these hyperparameters for gated RNNs.

The goal of this work is to map out the phase diagram of dynamics in the space of the hyperparameters of gated RNNs, which we find to be remarkably rich. In particular, the phase diagram informs us about the autonomous dynamics of a gated RNN at initialization, and the adjoint sensitivity analysis shows that the behavior of the gradients at initialization is also intimately linked to this picture [6].

*corresponding authors, listed alphabetically: tankut.can@gmail.com; kameshk@princeton.edu
[2]This extended abstract is based on the preprint arXiv:2007.14823 [6].

Third Workshop on Machine Learning and the Physical Sciences (NeurIPS 2020), Vancouver, Canada.

2 Gated Recurrent Neural Networks

We study a continuous-time gated RNN with two gates: one which dynamically controls the time constant (z-gate), and another which modulates the network connectivity matrix (r-gate). The hidden units $\mathbf{h} \in \mathbb{R}^N$ are coupled to the dynamical gating variables $\mathbf{r}, \mathbf{z} \in \mathbb{R}^N$ which follow the dynamical equations:

$$\frac{d\mathbf{h}}{dt} = \sigma(\mathbf{z}) \odot \Big[-\mathbf{h} + g_h J^h \Big(\phi(\mathbf{h}) \odot \sigma(\mathbf{r}) \Big) \Big], \tag{1}$$

$$\tau_z \frac{d\mathbf{z}}{dt} = -\mathbf{z} + \alpha_z J^z \phi(\mathbf{h}), \quad \tau_r \frac{d\mathbf{r}}{dt} = -\mathbf{r} + \alpha_r J^r \phi(\mathbf{h}), \tag{2}$$

where $\odot$ indicates element-wise product, and the nonlinearities ϕ and σ are applied element-wise to their vector argument. Here our focus is on autonomous dynamics, and we leave the influence of inputs for future work. The weight matrices have elements drawn from Gaussian distributions $\mathcal{N}(0, 1/N)$. We use the activation $\phi(x) = \tanh(x)$, and a gating sigmoid function $\sigma(x) = 1/(1 + \exp(-x))$. We consider the effect on the dynamics of varying the hyperparameters $\Theta = \{g_h, \alpha_z, \alpha_r\}$, which set the scale of the different weight matrices. For most of what follows, we consider $\tau_z = \tau_r = 1$.

We comment briefly on the relation of our network equations to the popular discrete-time GRU [3]. The $\mathbf{z}$ variable corresponds to the update gate, whereas the $\mathbf{r}$ variable is naturally analogous to the reset gate. Conventionally, these variables do not have intrinsic dynamics themselves, a limit which can be recovered by setting $\tau_z = \tau_r = 0$. The other subtle difference is that in the GRU, the nonlinearity ϕ typically comes *after* the linear transformation J^h. We find that the static mean-field theory derived for our model actually matches that for the GRU found in [1] (see Appendix A for details), and thus we expect much of the phenomena described here to be present in the GRU as well.

2.1 Dynamical mean-field theory

For the gated RNN with random weight matrices, we develop a dynamical mean-field theory which reduces the description of the $3N$ deterministic differential equations to three stochastic differential equations driven by Gaussian noise processes whose statistics have to be determined self-consistently. Specifically,

$$\frac{dh}{dt} = \sigma(z) \left(-h + g_h \eta_h \right), \quad \frac{dz}{dt} = -z + \alpha_z \eta_z, \quad \frac{dr}{dt} = -r + \alpha_r \eta_r, \tag{3}$$

where the Gaussian noise processes η^a for $a \in \{h, r, z\}$ are non-Markovian and have self-consistently determined variances. Let $\mathbf{x}(t) = (h(t), r(t), z(t))$ stand for the triple of state variables, and $\varphi(\mathbf{x}(t))$ denote any functional of these variables. Then we define $C_{\varphi(\mathbf{x})}(t, t') = \mathbb{E}\left[\varphi(\mathbf{x}(t))\varphi(\mathbf{x}(t'))\right]$, where the expectation is taken over the stochastic processes η^a. The self-consistency condition for the noise covariances are then given by

$$\mathbb{E}\left[\eta_h(t)\eta_h(t')\right] = C_{\phi(h)}(t, t')C_{\sigma(r)}(t, t'), \quad \mathbb{E}\left[\eta_z(t)\eta_z(t')\right] = \mathbb{E}\left[\eta_r(t)\eta_r(t')\right] = C_{\phi(h)}(t, t')$$

The z-gate appears as a multiplicative term in the equation for $h(t)$, causing it to be non-Gaussian. This complicates the analysis of the *dynamical* mean field theory (DMFT) in the presence of nonzero α_z. However, Gaussianity is restored in the static limit $dh/dt = 0$, corresponding to fixed points of the dynamics. Here, the static mean-field equations become

$$C_h = g_h^2 C_{\phi(h)} C_{\sigma(r)}, \quad C_r = \alpha_r^2 C_{\phi(h)}, \tag{4}$$

We note that these static mean-field equations for our continuous-time gated RNN are the same as those for the GRU found in [1].

We find that the important dynamical regimes, and the transitions between them, are determined only on the $g_h - \alpha_r$ plane. However, α_z essentially being a dynamical time constant, influences the stability properties, which we discuss further in Sec. (3.3)

3 Phase Diagram for the Gated RNN

The hyperparameter phase diagram for our gated RNN is presented in Figure (1), the essential aspects of which we describe in this section.

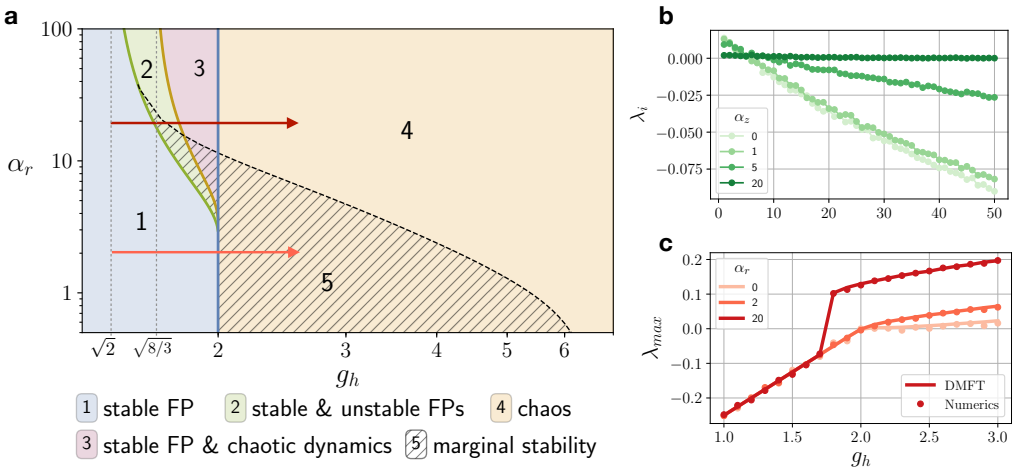

Figure 1: (**a**) Hyperparameter phase diagram for network (1-2) showing the different dynamical regimes in the $\alpha_r - g_h$ plane; (**b**) Lyapunov spectrum flattens with increasing α_z, showing emergence of marginal stability; (**c**) Maximum Lyapunov exponent as a function of g_h for large α_r (along red arrow in (**a**)), small α_r (along orange arrow in (**a**)), and $\alpha_r = 0$. Full network simulations (circles) agree well with DMFT prediction (solid line).

3.1 Fixed points from mean-field theory

The first crucial ingredient in constructing our phase diagram is a map of the fixed-point (FP) solutions to the mean-field theory (4). Finding non-zero solutions in the MFT typically implies that the microscopic equations (1-2) have a large (likely exponential in N [10]) number of critical points (i.e. saddles) which sculpt the dynamics [8]. Furthermore, if these FPs are unstable, the dynamics is typically chaotic. We find a notable exception to this, discussed in Sec. (3.5). We summarize the nature of the fixed-point solutions below.

In the absence of biases, $C_h = 0$ is always a solution. In addition to this trivial solution, we find the following regimes:

- For $g_h < \sqrt{2}$, $C_h = 0$ is the *only* solution to (4) for all α_r, and the zero fixed-point is **stable**.
- For $\sqrt{2} < g_h < 2$, there exists a bifurcation at a critical a_r^* where the number of fixed points jumps from one to three, with exactly two FPs on the critical curve. The bifurcation curve appears as the green line separating regions 1 and 2. The zero FP remains stable in this region, whereas the newly appearing non-zero FPs are unstable.

  The asymptotes of the bifurcation curve are given by the following: $\alpha_r^* \to \infty$ as $g_h \to \sqrt{2}^+$, whereas $\alpha_r^* \to \sqrt{8}$ as $g_h \to 2^-$. Therefore, with α_r below a certain threshold, the regions 2 and 3 will not be observed.
- For $g_h > 2$, there are two fixed points (including the zero FP), and both are **unstable**.

3.2 Jacobian stability analysis for fixed points

We study the stability properties of the fixed-points using the Jacobian spectrum. The instantaneous Jacobian evaluated near a fixed point can be treated as a *structured* random matrix, whose spectrum can be computed by the method of Hermitian reduction [4, 2]. In the large N limit, we evaluate the resolvent using the Dyson equation within the self-consistent Born approximation (c.f. [1]). The main result of this analysis is an expression for the curve describing the boundary of the Jacobian spectral density as a function of the network statistics, which is given by $\lambda \in \mathbb{C}$ satisfying

$$g_h^2 C_{\phi'(h)} \left(C_{\sigma(r)} + \frac{\alpha_r^2 C_{\phi(h)} C_{\sigma'(r)}}{|\lambda + 1|^2} \right) F(\lambda, \bar{\lambda}) = 1, \quad \text{where} \quad F(\lambda, \bar{\lambda}) = \mathbb{E}\left[\frac{\sigma(z)^2}{|\lambda + \sigma(z)|^2} \right]. \quad (5)$$

From this, we can determine the condition for stability. We ask what is the condition for $\lambda = 0$ to be on this boundary curve. This determines the FP instability transition, quoted in the previous section (3.1). In the absence of biases, the zero FP is the only stable FP. With biases it is possible to have stable nonzero FPs.

3.3 Marginal stability and line attractors

Despite being formally unstable, there is a region of phase space where the dynamics can exhibit a whole spectrum of very slow modes by increasing α_z. In the limit $\alpha_z = \infty$, $\sigma(z)$ becomes bimodal and $F(\lambda, \bar{\lambda}) \to \frac{1}{2|\lambda+1|^2}$. Using this in (5), we find that the FPs in the hatched region 5 of the phase diagram becomes *marginally stable*: the bulk of Jacobian eigenvalues has negative real part, while an extensive number of eigenvalues (approaching $N/2$) remain precisely at zero.

A careful analysis of the spectral curve shows that the leading edge of the support approaches zero exponentially in α_z, indicating that for even modestly large $\alpha_z \sim O(10)$, we can reasonably expect very slow dynamics and approximate marginal stability. This property of the instantaneous Jacobian is mirrored in the asymptotic (late-time) stability of the network. Asymptotically, the appropriate object of study is the *Lyapunov spectrum*, which gives a fine-grained breakdown of the distribution of stable and unstable (i.e. chaotic) directions in phase space, in which perturbations either decay or grow, respectively. Fig. (1b) shows that a significant fraction of the Lyapunov spectrum becomes nearly flat and concentrated near zero already for $\alpha_z = 20$.

3.3.1 Trainability near marginal stability

A practical consequence of this proximity to marginal stability is that the network supports approximate line attractors *without fine-tuning*, and which persist for very long timescales. In tasks which utilize line attractors for computation (e.g. sentiment classification [7]), we conjecture that initializing the network in this region will lead to more efficient training.

3.4 Transition to chaos

We find that for $\alpha_r < \sqrt{8}$, the fixed-point becomes unstable (blue (1) to yellow (4)) precisely when the dynamics continuously transitions from stable to chaotic. In fact, the two phenomena are tightly linked across this transition: 1) transition from one FP to many (unstable) FPs, which is a transition in the *topological complexity*, and 2) transition from negative Lyapunov exponent (stable FP) to positive Lyapunov exponent (chaotic attractor), which is a transition in the *dynamical complexity* [10].

This tight coupling between the topological and dynamical complexity can be broken in a novel chaotic transition in the region $g_h < 2$, which we now explore.

3.5 Discontinuous transition to chaos

Region 1 only supports a single fixed-point which is the global attractor of dynamics. Region 2 supports many more fixed points, but all of them are unstable; moreover, from region 1 to 2, there are no discernible dynamical consequences. A significant *dynamical* transition occurs crossing from Region 2 to 3, for $\sqrt{8/3} < g_h < 2$. Once in region 3, *chaotic transients* appear. Furthermore, we find that the lifetime of chaotic transients scales *extensively* with N, indicating, as $N \to \infty$, the emergence of a stable chaotic attractor *in addition to* the stable zero FP. This is reflected in the DMFT by the existence of a time-dependent solution to the auto-correlation function $C_h(t, t')$ in region 3.

Another novel aspect of this transition is the discontinuous jump in the maximal Lyapunov exponent when crossing the red line separating regions 2 and 3, shown in Fig. 1c and illustrated by a red arrow in Fig. 1a. In contrast, the transition to chaos along the orange arrow is continuous, with the Lyapunov exponent (Fig. 1c) passing through zero. In the continuous case, the transition line is critical in the sense that the timescale for relaxation of auto-correlation functions diverges as the transition is approached. In contrast, the discontinuous transition resembles a first-order phase transition, and there is strong chaotic activity right from the start.

4 Conclusion

In summary, we have shown how gates shape the dynamics of a RNN by mapping out a phase diagram, which additionally serves as a guide for hyperparameter choices. We show that gates can robustly produce line-attractor dynamics, which are useful mechanisms for tasks involving long-time dependencies. Gating can also produce a novel, discontinuous transition to chaos which is likely detrimental to training. In ongoing work, we investigate how gates modulate the interaction between inputs and intrinsic dynamics.

Broader Impact

Our work provides a principled way to assess how architectural choices shape the behavior of RNNs. This can be used to improve trainability of RNNs on tasks such as NLP and time-series prediction. Our work also shows how analytical techniques popular in theoretical physics can be leveraged to understand the behavior of machine learning models.

Acknowledgments and Disclosure of Funding

We have benefited greatly from conversations with William Bialek, Giulio Biroli, Jonathan Cohen, Andrea Crisanti, Rainer Engelken, Moritz Helias, Jonathan Kadmon, Louis Kang, Jimmy Kim, Itamar Landau, Wave Ngampruetikorn, Jeffrey Pennington, Katherine Quinn, Friedrich Schuessler, James Sethna, Julia Steinberg and Merav Stern. KK is supported by a C.V. Starr Fellowship and a CPBF Fellowship (through NSF PHY-1734030). DJS was supported by the NSF through the CPBF (PHY-1734030) and by a Simons Foundation fellowship for the MMLS. This work was partially supported by the NIH under award number R01EB026943. KK & DJS thank the Simons Institute for the Theory of Computing at U.C. Berkeley, where part of the research was conducted.

A Mapping to GRU

The Gated Recurrent Unit (GRU) [3] in the absence of input and bias is described by a hidden state variable $\mathbf{x}_t \in \mathbb{R}^N$, with two 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

$$\mathbf{z}_t = \sigma\left(\alpha_z J_z \mathbf{x}_{t-1}\right), \quad \text{update} \tag{6}$$

$$\mathbf{r}_t = \sigma\left(\alpha_r J_r \mathbf{x}_{t-1}\right), \quad \text{reset} \tag{7}$$

$$\mathbf{y}_t = J_h(\mathbf{r}_t \odot \mathbf{x}_{t-1}), \tag{8}$$

$$\mathbf{x}_t = \mathbf{z}_t \odot \mathbf{x}_{t-1} + (1 - \mathbf{z}_t) \odot \phi(g_h \mathbf{y}_t). \tag{9}$$

In the (static) mean-field theory for the fixed-point variances, y and x are Gaussian random variables with zero mean and variances C_y and C_x, respectively. The mean-field equations for the GRU are then [1]

$$C_y = C_x C_{\sigma(\alpha_r x)}, \quad C_x = C_{\phi(g_h y)}. \tag{10}$$

These equations can be seen as a single nonlinear implicit equation for either C_x or C_y. Comparing to (4), there is an exact mapping in which $g_h^2 C_y = C_h$, so that the MFT for both the continuous-time network and the GRU have the same space of FP solutions.

References

[1] Tankut Can, Kamesh Krishnamurthy, and David J. Schwab. Gating creates slow modes and controls phase-space complexity in GRUs and LSTMs. In *Proceedings of Machine Learning Research*, volume 107, pages 476–511, 2020.

[2] John T Chalker and Bernhard Mehlig. Eigenvector statistics in non-hermitian random matrix ensembles. *Physical review letters*, 81(16):3367, 1998.

[3] 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.

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

[5] Sepp Hochreiter and J Schmidhuber. Long Short-term Memory. *Neural Computation*, 9:1735–1780, 1997.

[6] Kamesh Krishnamurthy, Tankut Can, and David J. Schwab. Theory of gating in recurrent neural networks. *arXiv:2007.14823*.

[7] 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.

[8] David Sussillo and Omri Barak. Opening the Black Box : Low-Dimensional Dynamics in High-Dimensional Recurrent Neural Networks. *Neural Computation*, 25:626–649, 2013.

[9] 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.

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