Dynamics of continuous-time gated recurrent neural networks

Tankut Can[1*], Kamesh Krishnamurthy[2**], David J. Schwab[1]
[1]Initiative for the Theoretical Sciences, CUNY Graduate Center; [2]Princeton University
*tankut.can@gmail.com; **kameshk@princeton.edu

THE GRADUATE CENTER CITY UNIVERSITY OF NEW YORK — PRINCETON UNIVERSITY — Center for the Physics of Biological Function

Continuous-time Gated RNN

- Equations of motion: $\mathbf{h}, \mathbf{z}, \mathbf{r} \in \mathbb{R}^N$

$$\frac{d\mathbf{h}}{dt} = \sigma(\mathbf{z}) \odot \Big[-\mathbf{h} + g_h J^h \Big(\phi(\mathbf{h}) \odot \sigma(\mathbf{r})\Big)\Big], \qquad \phi(x) = \tanh(x) \qquad \sigma(x) = 1/\left(1 + e^{-x}\right)$$

$$\tau_z \frac{d\mathbf{z}}{dt} = -\mathbf{z} + \alpha_z J^z \phi(\mathbf{h}), \qquad \tau_r \frac{d\mathbf{r}}{dt} = -\mathbf{r} + \alpha_r J^r \phi(\mathbf{h}), \qquad J_{ij}^{h,z,r} \sim \mathcal{N}(0, 1/N)$$

update gate — **reset gate** — analogous gates in GRU [4]

- Hyperparameters are α_z and α_r for gates, and g_h for neuronal activation function
- Dynamical mean-field theory (DMFT) — autocorrelation is order parameter

$$\frac{dh}{dt} = \sigma(z)\left(-h + g_h \eta_h\right), \quad \frac{dz}{dt} = -z + \alpha_z \eta_z, \quad \frac{dr}{dt} = -r + \alpha_r \eta_r \qquad C_{\varphi(\mathbf{x})}(t, t') = \mathbb{E}_{\mathbf{x}}\left[\varphi(\mathbf{x}(t))\varphi(\mathbf{x}(t'))\right]$$

- Mean-field equation for fixed point, can be mapped to GRU MFT in [2].

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

DMFT for Gradients

- DMFT can be developed for adjoint dynamics [1], and used to study gradients

Loss $\mathcal{F}(\mathbf{x}(t), \boldsymbol{\theta}) = \int_0^T dt f(\mathbf{x}, \boldsymbol{\theta}, t)$ **w/ state** $\mathbf{x} = (\mathbf{h}, \mathbf{z}, \mathbf{r})$ **& parameters** $\boldsymbol{\theta} = (J^h, J^z, J^r)$

Adjoint dynamics $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_h, \boldsymbol{\lambda}_z, \boldsymbol{\lambda}_r)$

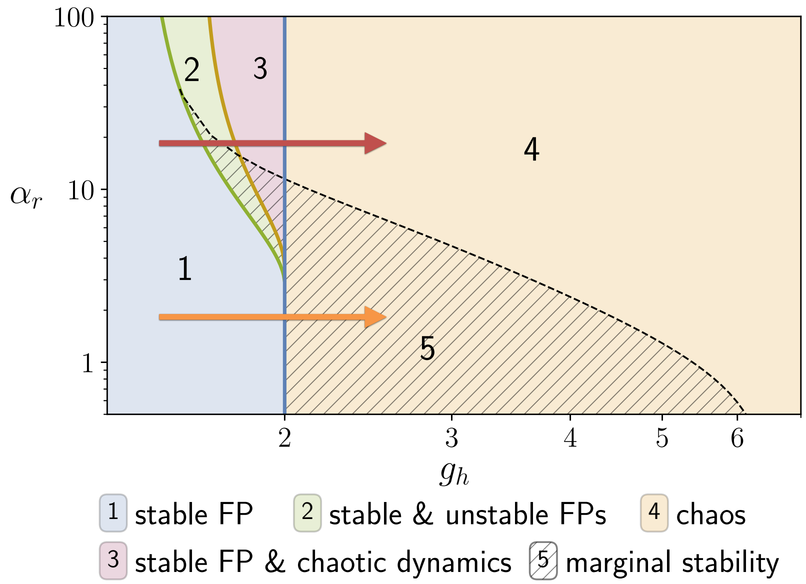
$$\frac{d\boldsymbol{\lambda}}{dt} = -\mathcal{D}(t)^T \boldsymbol{\lambda} + \frac{\partial f}{\partial \mathbf{x}}, \quad \boldsymbol{\lambda}(T) = 0$$

Gradient norm via DMFT

$$\left\langle \left\| \frac{d\mathcal{F}}{dJ^h} \right\|_2 \right\rangle = \int dt dt' C_{\lambda_h}(t, t') C_{\phi}(t, t') C_{\sigma_r}(t, t')$$

- Backpropagation of gradients is closely related to forward propagation via **Jacobian** $\mathcal{D}(t)$ (see e.g. [5]) ⇒ close relationship between network dynamics (forward propagation) and trainability (backpropagation).

Hyperparameter Phase Diagram for Gated RNN

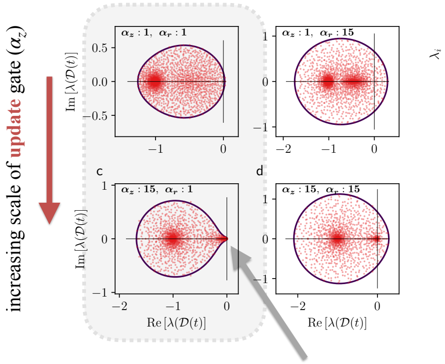


Main Takeaways

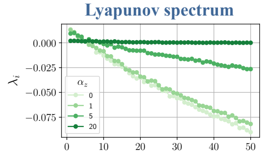
- **Complete dynamical phase diagram to guide hyperparameter initialization in GRUs**. Suggests interesting unexplored regions, e.g. near marginal stability with update gate effectively more switch-like. Also regions to avoid, e.g. near first-order transition to chaos with more switch-like reset gate.
- Gated RNNs have **robust line attractors** for a wide range of hyperparameters at initialization. Beneficial for training by mitigating exploding/vanishing gradients [5]. Also, can serve a computational purpose for certain tasks [7].
- **Rethinking edge-of-chaos initialization** in light of first-order (discontinuous) transition to chaos. Refined heuristic: initialize on *critical* transitions to chaos where timescales diverge. We show that a critical transition to chaos does not occur for certain hyperparameters.
- **DMFT for adjoint dynamics provides a theory of gradients** at initialization. Also opens up analysis for neural tangent kernel of RNNs.

Emergence of Marginal Stability and Line Attractors

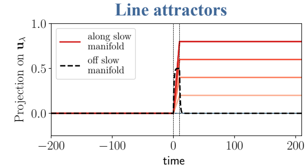
Jacobian Eigenvalues $\lambda(\mathcal{D}(t))$



Lyapunov spectrum



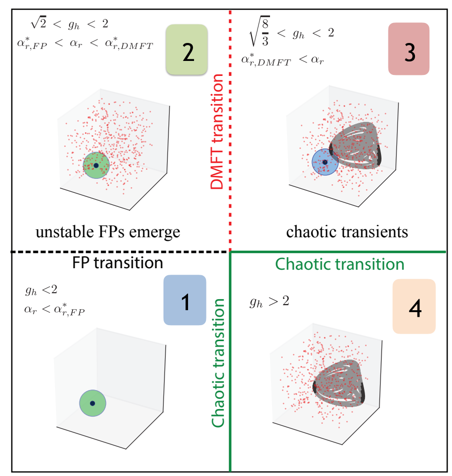
Line attractors



- In marginal stability phase (Region 5), **pinching of spectrum of the instantaneous Jacobian** $\mathcal{D}(t)$ at zero occurs with increasing α_z.
- Unstable nonzero FPs become **marginally stable** in this region
- Marginal stability persists asymptotically (long times), with the **Lyapunov spectrum flattening** out and showing an extensive number of Lyapunov exponents λ_i close to zero for large α_z.
- Nonzero solutions to MFT in this region indicate presence of many fixed-points (FPs), which implies the existence of approximate **line attractors at initialization**

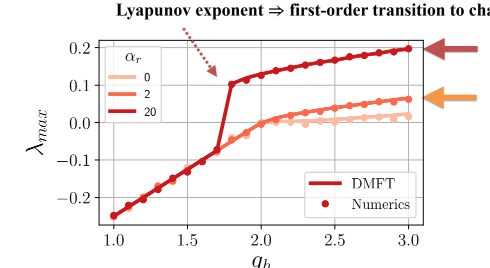
> - Marginal stability helps trainability by controlling gradients
> - Line attractors shown to emerge as mechanism for computation in dynamical systems [7]
>
> **⇒ large α_z in region 5 should be a good init**

Novel Discontinuous Transition to Chaos



1 → 2 Spontaneous appearance of (unstable) fixed points (FPs) via bifurcation. **No dynamical signatures.**

2 → 3 Spontaneous appearance of long chaotic transients with lifetime scaling with N. **Discontinuous jump in max Lyapunov exponent ⇒ first-order transition to chaos**



3 → 4 zero FP becomes unstable, chaos reigns

> - "Edge-of -chaos" init. implicitly assumes a *critical* transition to chaos (see e.g. [9]), which **does not occur** for large α_r
> - Proliferation of fixed-points does not coincide with transition to chaos - in contrast to RNN without gating [8]

[1] K. Krishnamurthy, T. Can, and D. J. Schwab, arXiv:2007.14823, 2020.
[2] T. Can, K. Krishnamurthy, and D. J. Schwab, PMLR, vol 107, pg 476-511, 2020.
[3] S. Hochreiter and J. Schmidhuber, Neural computation 9, 1735–1780, 1997.
[4] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, EMNLP 2014, 1724–1734, 2014.
[5] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, In A Field Guide to Dynamical Recurrent Neural Networks, pgs 237–243. Wiley-IEEE Press, 2001.
[6] Sutskever, I., J. Martens, G. Dahl, and G. Hinton. ICML 2013 pp. 1139-1147.
[7] N. Maheswaranathan, A. H. Williams, M. D. Golub, S. Ganguli, and D. Sussillo, ICML 2019.
[8] G. Wainrib and J. Touboul, Phys. Rev. Lett. **110**, 118101, 2013.
[9] N. Bertschinger and T. Natschlger. Neural Computation, 16,1413–1436, 2004.