Dynamics of continuous-time gated recurrent neural networks

Tankut Can\textsuperscript{1}, Kamesh Krishnamurthy\textsuperscript{2*}, David J. Schwab\textsuperscript{1}
\textsuperscript{1}\text{\small Initiative for the Theoretical Sciences, CUNY Graduate Center;} \textsuperscript{2}\text{\small Princeton University}
*tankut.can@gmail.com; **kameshk@princeton.edu

\textbf{Continuous-time Gated RNN}

- Equations of motion: \( h, x, r \in \mathbb{R}^D \)
  \[ \begin{align*}
  \dot{h}_i &= \sigma(x)_i - h_i + y_i \alpha \left( \langle h_i \rangle + \langle e_i \rangle \right), \\
  \dot{r}_i &= -r_i + \alpha \langle h_i \rangle
  \end{align*} \]

- \( \alpha \) is the gate control, and \( \langle \cdot \rangle \) denotes the state-averaged value of \( \cdot \).

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

\textbf{DMF for Gradients}

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

\[ \begin{align*}
  \text{Loss} &= \int \left( \frac{\partial f}{\partial \theta} \right) \cdot \theta \, w/ \text{state} \, x = (h, x, r) \quad \text{\& parameters} \, \theta = (J_0, J_1, J_2) \\
  \text{Adjoint dynamics} \, \lambda = (\lambda_x, \lambda_h, \lambda_r) \\
  \text{Gradient norm via DMFT} \\
  \frac{\partial \theta}{\partial \lambda} = -D\left( \frac{\partial f}{\partial \theta} \right) \cdot \lambda, \, \lambda(T) = 0 \\
  \text{Backpropagation of gradients is closely related to forward propagation via Jacobian} \, D(t) \text{ (see e.g. [5])} \Rightarrow \text{close relationship between network dynamics (forward propagation) and trainability (backpropagation).}
\]

\textbf{Emergence of Marginal Stability and Line Attractors}

- Jacobian Eigenvalues \( \lambda(D(t)) \)

- Lyapunov spectrum

- Line attractors

- Marginal stability persists asymptotically (long times), with the Lyapunov spectrum flattening out and showing an extensive number of Lyapunov exponents \( \lambda \) close to zero for large \( \alpha \).

- Nonzero solutions to MFT in this region indicate presence of many fixed-point (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]

- \( \lambda \) is large in region 5 should be a good init

- In marginal stability phase (Region 5), pinching of spectrum of the instantaneous Jacobian \( D(t) \) at zero occurs with increasing \( \alpha \).

- Unstable nonzero FPs become marginally stable in this region

- “Edge-of-chaos” init. implicitly assumes a critical transition to chaos (see e.g. [9]), which does not occur for large \( \alpha \).

- Proliferation of fixed-points does not coincide with transition to chaos - in contrast to RNN without gating [8]

\textbf{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.

\begin{itemize}
  \item 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.
  \item Gated RNNs have robust line attractors for a wide range of hyperparameters at initialization. Beneficial for training by mitigating exploding/vanishing gradients. Also, can serve a computational purpose for certain tasks.
  \item 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.
  \item DMFT for adjoint dynamics provides a theory of gradients at initialization. Also opens up analysis for neural tangent kernel of RNNs.
\end{itemize}