Research Design on Controllability and Observability Dynamics in Neural Networks

- Task can be spectrally decomposed based on its graph laplacian's eigenbasis
- Local deep learning dynamics may be modeled by a snapshot of linearized system where parameter is input, latent is state and loss is output
- This induces controllability and observability gramian-based matrices at each training step whose spectral statistics may be analyzed in relation to training dynamics

1 Task

Task Graph Let the finite state set be $S = \{1, \dots, n\}$ and the task graph $G = (V, E)$ with $V = S$. For an undirected case, $A \in \mathbb{R}^{n \times n}$ is symmetric and nonnegative with $A_{ij} = 1$ if $(i, j) \in E$, else 0. Let $D = \operatorname{diag}(d_1, \dots, d_n)$ with $d_i = \sum_j A_{ij}$ and the combinatorial Laplacian $L = D - A$.

Spectral Decomposition We use the generalized eigen-decomposition

$$Lu_p = \lambda_p D u_p, \quad u_p^\top D u_q = \delta_{pq}, \quad 0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n, \tag{1}$$

so that $\{u_p\}_{p=1}^n$ is a D-orthonormal basis. Small λ_p indicate smooth (low-frequency) graph modes; large λ_p indicate high-frequency modes.

Task Function A task function $f \in \mathbb{R}^n$ is a monotonic and unimodal (i.e., single-goal) function of states that signals reward. Its spectral coefficients in the task-graph eigenbasis are

$$a_p = u_p^\top D f, \qquad f = \sum_{p=1}^{n} a_p \, u_p. \tag{2}$$

The above assumes an undirected, connected graph. If the task connectivity is directed or disconnected, a more suitable Laplacian (e.g., random-walk or symmetrized) may be considered.

2 Dynamics

Linearization Let $\theta \in \mathbb{R}^p$ be trainable parameters, $z(s; \theta) \in \mathbb{R}^d$ the flattened latent for state $s \in S$, and $\ell \in \mathbb{R}$ the training loss. Around a reference (θ_0, z_0, ℓ_0) we consider the standard discrete-time LTV linearization

$$x_{t+1} = A_t x_t + B_t u_t, \tag{3}$$

$$y_t = C_t x_t + D_t u_t, \tag{4}$$

with $x_t := z_t - z_0 \in \mathbb{R}^d$, $u_t := \theta_t - \theta_0 \in \mathbb{R}^p$, and $y_t := \ell_t - \ell_0 \in \mathbb{R}$. The Jacobians at time t are

$$A_t = \frac{\partial z_{t+1}}{\partial z_t} \in \mathbb{R}^{d \times d}, \quad B_t = \frac{\partial z_t}{\partial \theta_t} \in \mathbb{R}^{d \times p}, \quad C_t = \frac{\partial \ell_t}{\partial z_t} \in \mathbb{R}^{1 \times d}, \quad D_t = \frac{\partial \ell_t}{\partial \theta_t} \in \mathbb{R}^{1 \times p}. \tag{5}$$

Note that index t may refer to optimization steps or environmental step, where for the latter, input, state and output will have to be averaged across the optimization steps, after an empirical validation that they are sufficiently static.

Controllability, Observability and Hankel Operator Suppose our interest is local, i.e. set the horizon of the LTV system to be 1. Then the LTV system reduces to a step-dependent LTI system where (A_t,B_t,C_t,D_t) is a snapshot of the LTV system at each step t, with controllability and observability matrices as follows. Omitting the step-t subscript from now on,

$$C_T := \begin{bmatrix} B & AB & \cdots & A^{T-1}B \end{bmatrix} \in \mathbb{R}^{d\times Tp}, \qquad O_T := \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{T-1} \end{bmatrix} \in \mathbb{R}^{T\times d}. \tag{6}$$

The associated finite–horizon Gramians are

$$W_c^{(T)} := C_T C_T^\top = \sum_{k=0}^{T-1} A^k BB^\top (A^\top)^k, \qquad W_o^{(T)} := O_T^\top O_T = \sum_{k=0}^{T-1} (A^\top)^k C^\top C A^k. \tag{7}$$

If A is Schur-stable ($\rho(A) < 1$), T may be replaced by ∞ to deduce infinite-horizon Gramians. Finally, the associated Hankel operator $H : \mathbb{R}^p \to \mathbb{R}$ at step t can be written in terms of Markov parameters as

$$[H]_{i,j} = [O_\infty]_i [C_\infty]_j = CA^{i+j}B \in \mathbb{R}^{1\times p}, \qquad i, j \geq 0, \tag{8}$$

where $[H]_{i,j}$ indicates the effect of an input applied j steps in the past on the output i steps in the future at step t. For a stable, minimal LTI realization and infinite horizon, the nonzero singular values of H equal the square roots of the nonzero eigenvalues of $W_o W_c$.

Degenerate Case (for quick empirical test and interpretation) The choice $A_t = 0$ and $D_t = 0$ yields the degenerate relation

$$y_{t+1} = CBu_t \quad \text{i.e.,} \quad \ell_{t+1} - \ell_0 = \frac{\partial \ell_t}{\partial z_t}\frac{\partial z_t}{\partial \theta_t}(\theta_t - \theta_0) \tag{9}$$

Then the Gramians and Hankel operator for the case $i = j = 0$,

$$W_c^{(T)} = BB^\top = \frac{\partial z}{\partial \theta}\frac{\partial z}{\partial \theta}^\top, \quad W_o^{(T)} = C^\top C = \left(\frac{\partial \ell}{\partial z}\right)^\top \frac{\partial \ell}{\partial z}, \quad [H]_{0,0} = CB = \frac{\partial \ell}{\partial \theta}, \tag{10}$$

for all T through which the singular values of the pairwise products of familiar partial derivatives ($\partial\ell/\partial z,\ \partial z/\partial\theta$) can respectively be interpreted as the extent of controllability, observability and the joint of both (hence the gradient) along the corresponding mode.

3 Energy

Controllability, Observability and Hankel Energy Let $f \in \mathbb{R}^n$ be a task function and $Z \in \mathbb{R}^{d\times n}$ be a mapping with $\|Zf\|_2 = \|f\|_D = 1$ that embeds graph signals such as f to latent space. Given $W_c^{(T)}$ and $W_o^{(T)}$, whether it being finite or infinite-horizon, define induced operators over graph signals as

$$M_c^{(T)} := Z^\top W_c^{(T)} Z \in \mathbb{R}^{n\times n}, \qquad M_o^{(T)} := Z^\top W_o^{(T)} Z \in \mathbb{R}^{n\times n}. \tag{11}$$

Define the controllability/observability energies as

$$C(f,T) := (Zf)^\top W_c^{(T)} (Zf) = f^\top M_c f \qquad O(f,T) := (Zf)^\top W_o^{(T)} (Zf) = f^\top M_o f \tag{12}$$

Finally, define the Hankel energy as

$$H(f,T) := (Zf)^\top (W_o^{(T)} \# W_c^{(T)})^{1/2} (Zf) \leq \sqrt{O(f,T)\, C(f,T)} \tag{13}$$

where the last inequality comes from Cauchy-Schwarz.

Spectral Expansion Recall $f = \sum_{p=1}^{n} a_p u_p$ where $\{u_p\}$ is the eigenbasis of the task graph. Omitting the horizon T superscript,

$$C(f) = \sum_{p,q} a_p a_q \, u_p^\top M_c u_q \qquad O(f) = \sum_{p,q} a_p a_q \, u_p^\top M_o u_q \qquad H(f) = \sum_{p,q} a_p a_q \, u_p^\top (M_o \# M_c)^{1/2} u_q \tag{14}$$

The terms where $p \neq q$ indicate mixing between graph modes caused jointly by Z and (A, B, C). If M_c and M_o are diagonal in $\{u_p\}$, let $c_p := u_p^\top M_c u_p$ and $o_p := u_p^\top M_o u_p$. Then

$$C(f) \approx \sum_p a_p^2 c_p \qquad O(f) \approx \sum_p a_p^2 o_p \qquad H(f) \approx \sum_p a_p^2 \sqrt{c_p o_p} \tag{15}$$

4 Future Plan

- Section 2: Input (parameter), state (latent) and output (loss) are defined with respect to a reference point. In deep neural networks, all inputs, states and outputs are updated from the previous one, meaning the variables defined as change may be more appropriate for interpretation.

References

[1] B. Canatar, S. Bordelon, and C. Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks, Nature Communications, 2021

[2] A. Jacot, F. Gabriel, C. Hongler, Neural Tangent Kernel: Convergence and Generalization in Neural Networks, NeurIPS, 2020