Chaotic Dynamics are Intrinsic to Neural Network Training with SGD

Anonymous Author(s)
Affiliation
Address
email

Proof for reviewer tUai

Consider a neural network trained by SGD, whose output at time step t is given by a function $\boldsymbol{f}(\boldsymbol{w}^{(t)}; \boldsymbol{z}^{(t)})$ parametrized by weights $\boldsymbol{w}^{(t)}$ from the reals, subject to inputs $\boldsymbol{z}^{(t)}$ from the reals. For more clarity, the inputs $\boldsymbol{z}^{(t)}$ will be omitted in the rest of the proof. Otherwise, we assume the same formalism as in the main paper.

In our paper, we defined the finite-time Lyapunov matrix after a single time step as

$$\boldsymbol{\Lambda}^{(t+1,t)} = \frac{1}{2} \ln \boldsymbol{Y}^{(t+1,t)T} \boldsymbol{Y}^{(t+1,t)}, \tag{1}$$

where $\boldsymbol{Y}^{(t+1,t)}$ is the tangent map after a single time step given by

$$\boldsymbol{Y}^{(t+1,t)} = \boldsymbol{J}_F^{(t)} = \left\{ \frac{\partial w_i^{(t+1)}}{\partial w_j^{(t)}} \right\}_{ij}. \tag{2}$$

The Neural Tangent Kernel is defined as $K^{(t)} = \nabla \boldsymbol{f}(\boldsymbol{w}^{(t)}) \nabla \boldsymbol{f}(\boldsymbol{w}^{(t)})^T$, where

$$\nabla \boldsymbol{f}^{(t)} = \left\{ \frac{\partial f_i^{(t)}}{\partial w_j^{(t)}} \right\}_{ij} \tag{3}$$

Using the chain rule, the expression for the tangent map can be rewritten as

$$Y_{ij}^{(t+1,t)} = \frac{\partial w_i^{(t+1)}}{\partial w_j^{(t)}} = \sum_k \frac{\partial w_i^{(t+1)}}{\partial f_k^{(t)}} \frac{\partial f_k^{(t)}}{\partial w_j^{(t)}} \tag{4}$$

Using the update rule $w_i^{(t+1)} = w_i^{(t)} - \gamma g_i^{(t)}$ for SGD (no momentum), we find

$$Y_{ij}^{(t+1,t)} = \sum_k \frac{\partial}{\partial f_k^{(t)}} (w_i^{(t)} - \gamma g_i^{(t)}) \frac{\partial f_k^{(t)}}{\partial w_j^{(t)}} = \sum_k -\gamma \frac{\partial g_i^{(t)}}{\partial f_k^{(t)}} \frac{\partial f_k^{(t)}}{\partial w_j^{(t)}}. \tag{5}$$

Finally, using the definition of the gradient as $g_i^{(t)} = \frac{\partial \mathcal{L}(w^{(t)})}{\partial w_i^{(t)}}$, we obtain

$$Y_{ij}^{(t+1,t)} = -\gamma \sum_k \frac{\partial^2 \mathcal{L}(\boldsymbol{w}^{(t)})}{\partial w_i^{(t)} \partial f_k^{(t)}} \frac{\partial f_k^{(t)}}{\partial w_j^{(t)}}, \tag{6}$$

or in matrix notation

$$\boldsymbol{Y}^{(t+1,t)} = -\gamma \boldsymbol{H}_{wy}^{(t)} \cdot \nabla \boldsymbol{f}(\boldsymbol{w}^{(t)}), \tag{7}$$

where $\boldsymbol{H}_{wy}^{(t)}$ denotes the Parameter-Output-Hessian at time step t.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.