A Convergence Analysis of Gradient Descent on Graph Neural Networks

Pranjal Awasthi Abhimanyu Das Sreenivas Gollapudi

Google Research

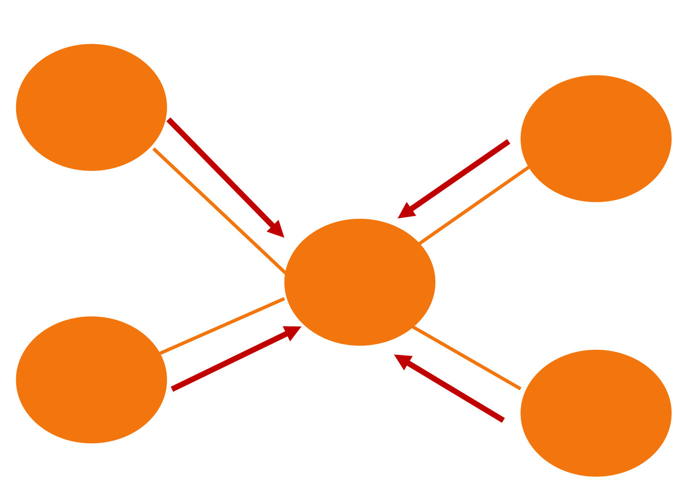
We provide a convergence analysis of gradient descent on graph neural networks.

Motivation

- GNNs are an elegant framework for designing learning algorithms for graphs.
- GNNs are optimized by gradient descent
- However no understanding of the dynamics of GD.
- In contrast, lots of work on fully connected networks.
- Goal: understand the convergence of GD on GNNs.

Model and Preliminaries

- Computation in a GNN proceeds in a message passing manner.
- v_i^h: Computation in a GNN proceeds in a message passing manner.
  - However no understanding of the dynamics of GD.
- Basic operations in a GNN:
- $a_i^h = AGGREGATE(v_j^h : j \in N(i))$.
- $v_i^{h+1} = COMBINE(v_i^h, a_i^h)$.
- To obtain graph embedding at layer H:
- $g^H = READOUT(v_i^H : v_i \in V)$.



Contributions

- We consider two settings
  - ReLU GNNs with one round of message passing.
  - Linear GNNs with multiple rounds of message passing.

Setting 1: ReLU GNNs

- Let G = (V,E) be a graph with degree at most d and n vertices.
- Let $x_i \in \mathbb{R}^r$ be input to node i. Then,
  - $y(W^*) = \sum_i \sigma(W^* \tilde{x}_i)$ where,
  - $\tilde{x}_i = \sum_{j \in N(i)} x_j$ and $x_j \sim N(0, I)$.
- Optimized via GD on loss $L(W) = \mathbb{E}[||y(W) - y(W^*)||^2]$.

Main Result for ReLU GNNs

Theorem. From random initialization ($N(0, \sigma^2 I)$) with an appropriate variance, after T steps of GD, with high probability,

(Loss Bound) $L(W_T) \le \epsilon^2$

(Provided) $d = o(\sqrt{n})$ and $T = \Omega\left(\frac{n^4 r^2 d}{\epsilon^2}\right)$.

Proof Technique

- Show that either W_t is already close to W^* or the PL-condition holds:
  - $||\nabla L(W_t)||^2 \ge \mu L(W_t)$ where,
  - $\mu^* = \Omega(\frac{\epsilon^2}{d r^2 n^2})$
- Challenge: The PL condition may not hold initially. Need to separately argue about Phase I after which the iterates enter the PL-region.

Setting 2: Deep Linear GNNs

- Let G = (V,E) be a graph with degree at most d and n vertices.
- Let $x_i \in \mathbb{R}^r$ be input to node i. Then,
  - $y(W_1^*, W_2^*) = \sum_i x_i^L$ where,
  - $x_i^L = W_1^* x_i^{L-1} + W_2^* \sum_{j \in N(i)} x_j^{L-1}$ and $x_j \in N(0, I)$.
- Optimized via GD on loss $L(W_1, W_2) = \mathbb{E}[||y(W_1, W_2,) - y(W_1^*, W_2^*)||^2]$.

Main Result for Deep Linear GNNs

Theorem. From identity initialization, after T steps of GD,

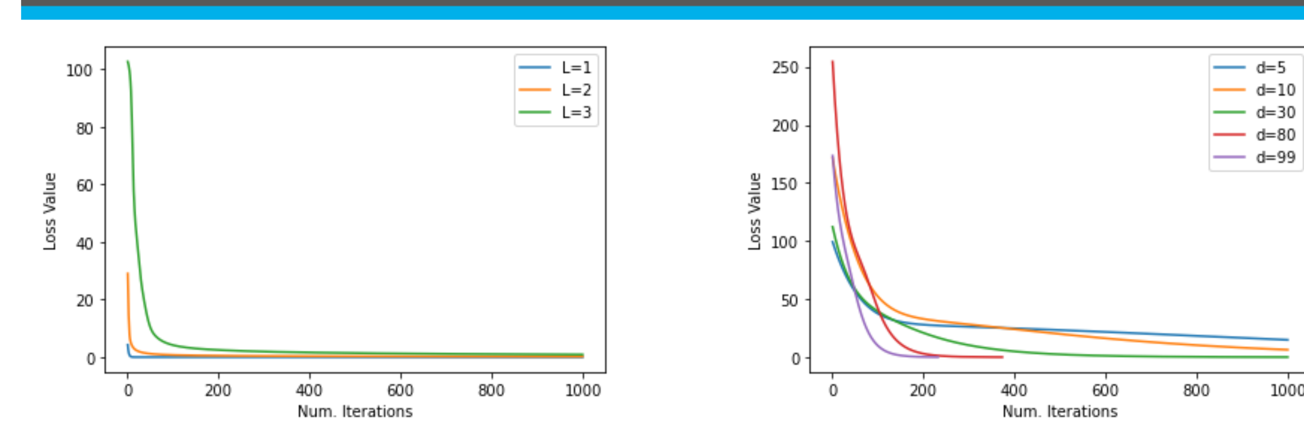
(Loss Bound) $L(W_{1,T}, W_{2,T}) \le \epsilon^2$

(Provided) $T = \Omega\left(2^L d^2 n^2 r^2 \log(\frac{nr}{\epsilon})\right)$.

Proof Technique

- Analyze the evolution of the singular values of W_1, W_2.

Experiments



- Left: convergence for various depths for deep linear GNNs.
- Right: convergence for various degrees for ReLU GNNs.
- Polylogarithmic dependence on ϵ unlikely for ReLU GNNs.

Future Directions

- Remove the dependence on small d for ReLU GNNs
  - Seems to be an artifact of the analysis
- Extend the result for deep linear GNNs to allow different weights per layer.
- Extend the analysis to non-linear GNNs with multiple rounds and to other variants of GNNs.