Optimization of Graph Neural Networks: Implicit Acceleration by Skip Connections and More Depth

Keyulu Xu, Mozhi Zhang, Stefanie Jegelka, Kenji Kawaguchi

Background: theory of GNNs

Optimization

Can gradient descent find a global minimum for GNNs?

What affects the speed of convergence?

Expressive Power

(Xu et al. 2019, Sato et al 2020, Chen et al 2019, 2020, Maron et al 2019, Keriven et al 2019, Loukas 2020, Balcilar et al 2021, Morris et al 2020, Azizian et al 2021, Vignac et al 2020)

Generalization

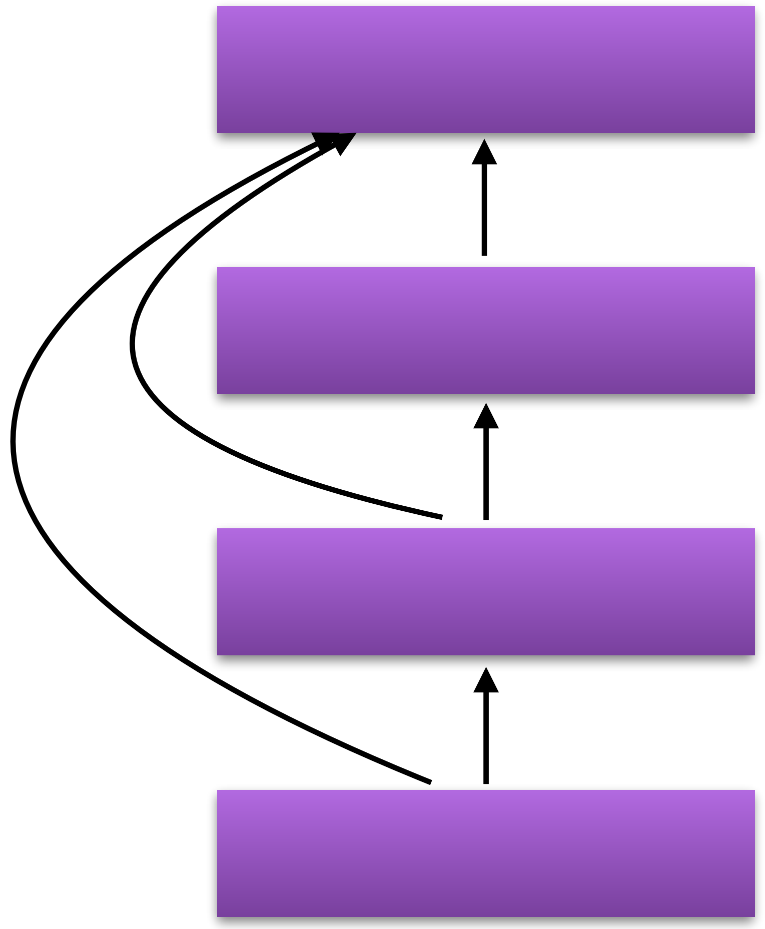
(interpolation & extrapolation)

(Scarselli et al. 2018, Verma et al 2019, Du et al 2019, Garg et al 2020, Xu et al 2020, 2021)

Analysis of gradient dynamics

Linearized GNNs with and without skip connections *(non-convex)*:

$$f(X, W, B) = \sum_{l=0}^{H} W_{(l)} X_{(l)},$$
$$X_{(l)} = B_{(l)} X_{(l-1)} S.$$



JK-Net

(Xu et al 2018)

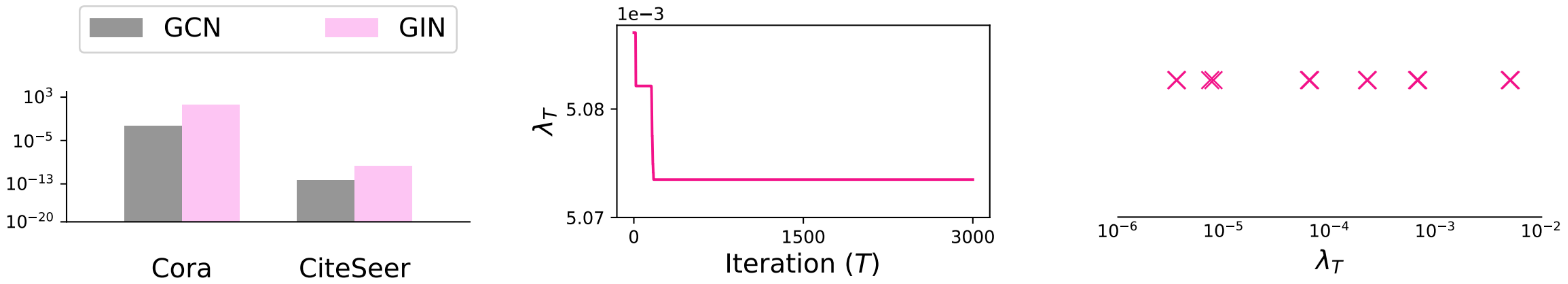
Trajectory of gradient descent (flow) training:

$$\frac{d}{dt} W_t = -\frac{\partial L}{\partial W}(W_t, B_t), \quad \frac{d}{dt} B_t = -\frac{\partial L}{\partial B}(W_t, B_t)$$

Global convergence

Theorem (XZJK’21)
Gradient descent training of a linearized GNN, with or without skip connections, converges to a *global minimum* at a linear rate.

Rates (with vs. without skip): $e^{-\boxed{4\lambda_T^{(1:H)}}\boxed{\sigma^2_{\min}((G_H)_{*\mathcal{I}})T}}$ $e^{-4\lambda_T^{(H)}\sigma^2_{\min}(X(S^H)_{*\mathcal{I}})T}$



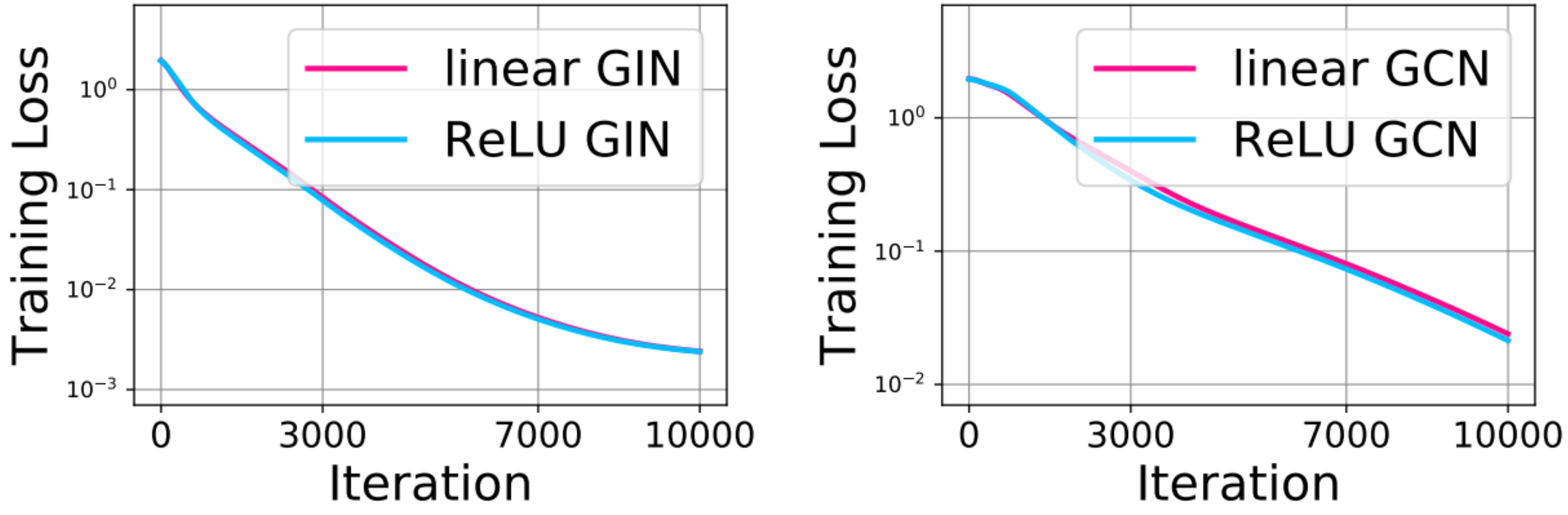
(a) Graph $\sigma^2_{\min}(X(S^H)_{*\mathcal{I}})$

(b) Time-dependent $\lambda_T^{(H)}$

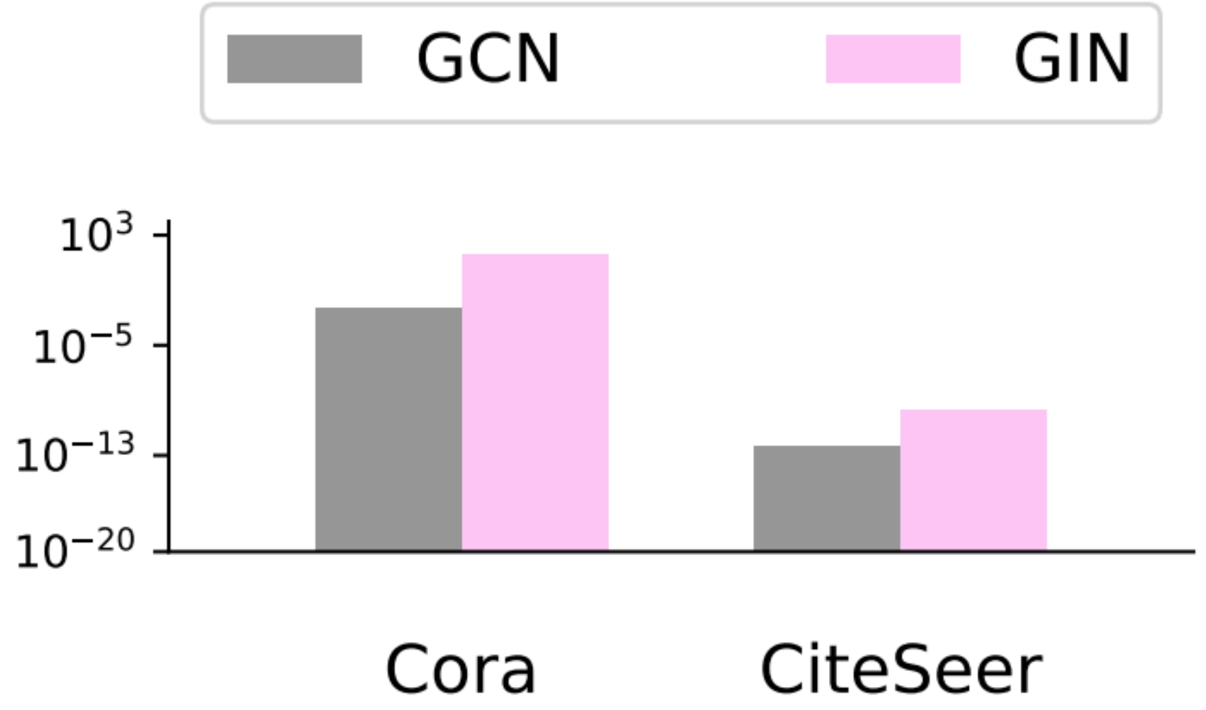
(c) $\lim_{T\to\infty}\lambda_T^{(H)}$ across training settings

Convergence in practice

Linear convergence for linearized and ReLU GNNs:



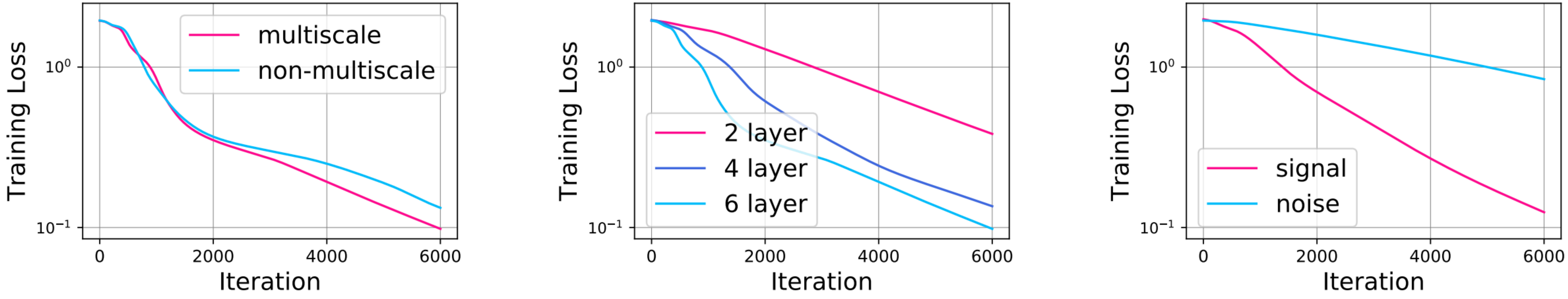
Rates for GIN vs. GCN:



Implicit acceleration

Theorem (XZJK'21)

GNN training is implicitly accelerated by skip connections, depth, and/or a good label distribution.



(a) Multiscale vs. non-multiscale.

(b) Depth.

(c) Signal vs. noise.

Implications

1. Global convergence implies more powerful GNNs better fit the training data

2. Deep GNNs with skip connections are promising:
    - more expressive & helps with over-smoothing
    - faster convergence

More information:

https://people.csail.mit.edu/keyulux/