Backward oversmoothing: why is it hard to train deep GNNs ?

Nicolas Keriven

CNRS, IRISA

UMR IRISA

European Research Council

Established by the European Commission

cnrs

Forward oversmoothing

Vanilla GNN: $X^{(k)} = \rho\left(PX^{(k-1)}W^{(k)}\right)$

Stochastic propagation matrix

exemple: $P = D^{-1}A$

Forward oversmoothing

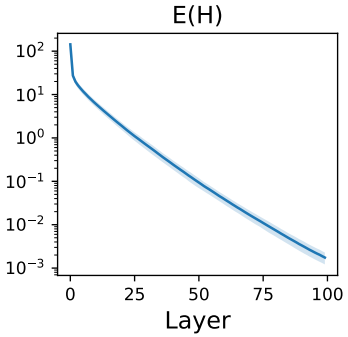
Vanilla GNN: $X^{(k)} = \rho\left(PX^{(k-1)}W^{(k)}\right)$

Stochastic propagation matrix

exemple: $P = D^{-1}A$

Oversmoothing: Assuming $\|W^{(k)}\|_{op} \leq s$

Distance to constant node representations → $\mathcal{E}_1(X^{(k)}) \lesssim (\lambda s)^k$



Forward oversmoothing

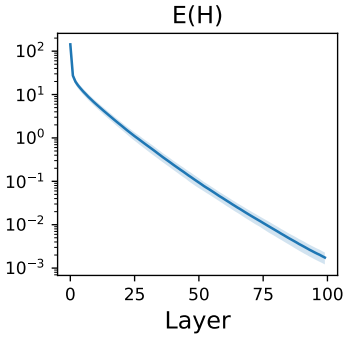
Vanilla GNN: $X^{(k)} = \rho\left(PX^{(k-1)}W^{(k)}\right)$

Stochastic propagation matrix

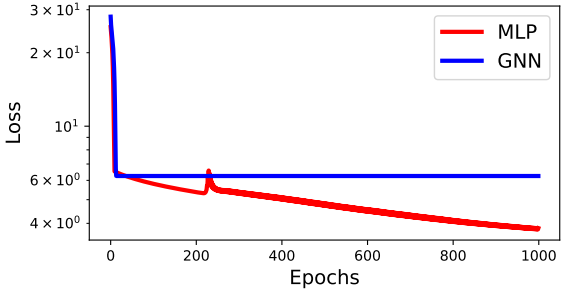
exemple: $P = D^{-1}A$

Oversmoothing: Assuming $\|W^{(k)}\|_{op} \leq s$

Distance to constant node representations → $\mathcal{E}_1(X^{(k)}) \lesssim (\lambda s)^k$



- GNN could learn to anti-oversmooth with large s
- But they don't in practice. Why?



Backward oversmoothing

Gradients depend on forward and ***backward*** graph signal

$$B^{(k)} = \rho'(PX^{(k-1)}W^{(k)}) \odot (S^\top B^{(k+1)} {W^{(k+1)}}^\top)$$

Almost constant *when oversmoothed*

Backward oversmoothing

Gradients depend on forward and ***backward*** graph signal

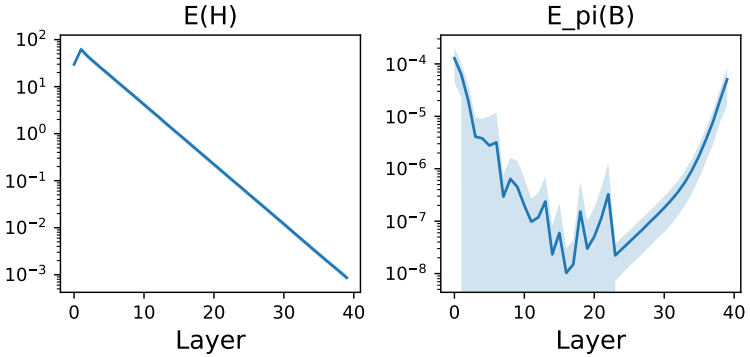
$$B^{(k)} = \rho'(PX^{(k-1)}W^{(k)}) \odot (S^\top B^{(k+1)} W^{(k+1)\top})$$

Almost constant *when oversmoothed*

Thm $\mathcal{E}_\pi(B^{(k)}) \lesssim \lambda^k + (s\lambda)^{L-k}$

Distance to left-stationary distribution of P

Back. oversmoothing *at the middle layers*



Consequence?

- Backward oversmoothing is “intuitively” bad. ***But why?***
- Provably, it is not (necessarily) “vanishing gradients”

Consequence?

- Backward oversmoothing is "intuitively" bad. ***But why?***
- Provably, it is not (necessarily) "vanishing gradients"
- But:

Thm (informal)

If the layer L is at an approximate ***stationary point****, then the GNN is at an approximate global stationary point*

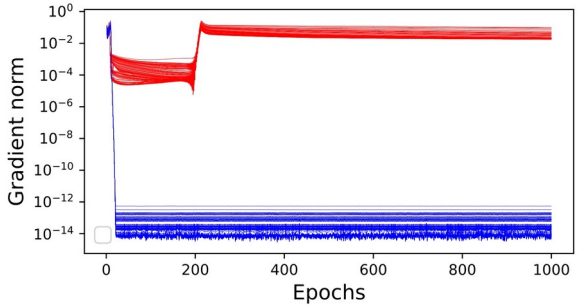
Consequence?

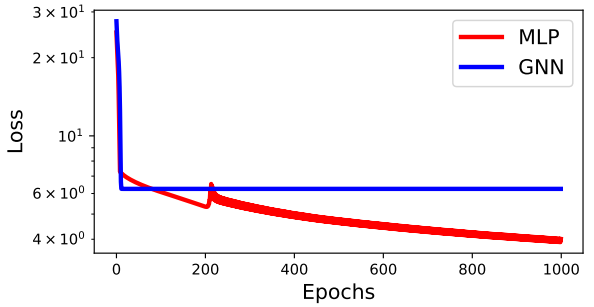
- Backward oversmoothing is "intuitively" bad. ***But why?***
- Provably, it is not (necessarily) "vanishing gradients"
- But:

Thm (informal)

If the layer L is at an approximate ***stationary point***, then *the GNN is at an approximate global stationary point*

As soon as the last layer is trained, ***all gradients vanish***





Remarks

- The proof is more interesting than the result!
  - ***Interaction*** forward/backward oversmoothing
- Lemma: existence of bad stationary points with high loss
- Lemma: this is ***specific to GNNs*** (counter-example MLP)
- Outlook: skip connection, normalization, attention-based models...

POSTER | Thu, Jul 9, 2026 • 7:30 AM – 9:15 AM CEST | Coex: HALL A