Principles for Initialization and Architecture Selection in Graph Neural Networks

Gage DeZoort, Princeton University

Joint work with Boris Hanin

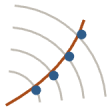
SIMODS Minisymposium · SIAM Conference on Mathematics of Data Science (MDS26) · Salt Lake City · November 19, 2026

Relational data



Particle decays

particles, linked from parent to daughter



Detector hits

measurements, linked when they could share a track



Planetary systems

bodies, linked by gravitational interaction



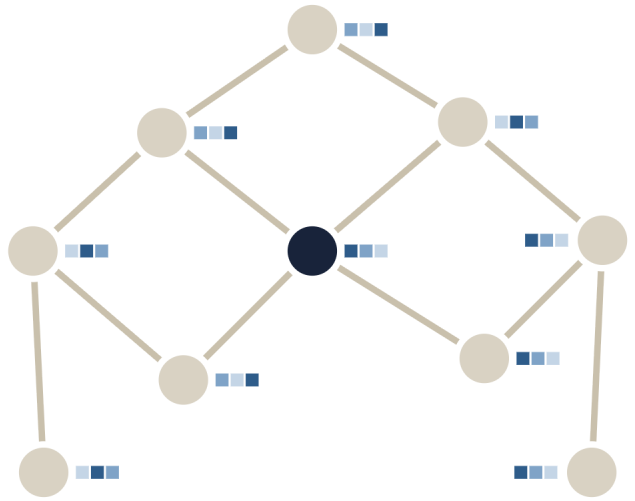
Molecules

atoms, linked by chemical bonds



Citation networks

papers, linked by citations



$G = (V, E)$ nodes V, edges E

$A \in \mathbb{R}^{|V| \times |V|}$ adjacency matrix

$x \in \mathbb{R}^{|V| \times n_0}$ n_0 input features per node

Many hyperparameters, few principles

Architecture

- depth and width
- message passing scheme
- residual connections

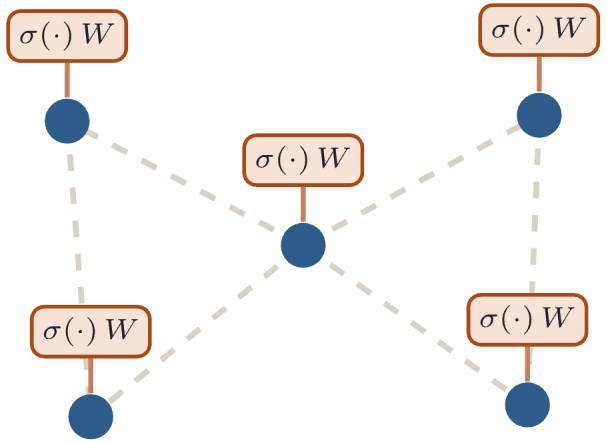
Optimization

- initialization
- learning rate and optimizer
- training schedule

Goal develop principles for selecting hyperparameters

Approach make the forward and backward passes stable at initialization

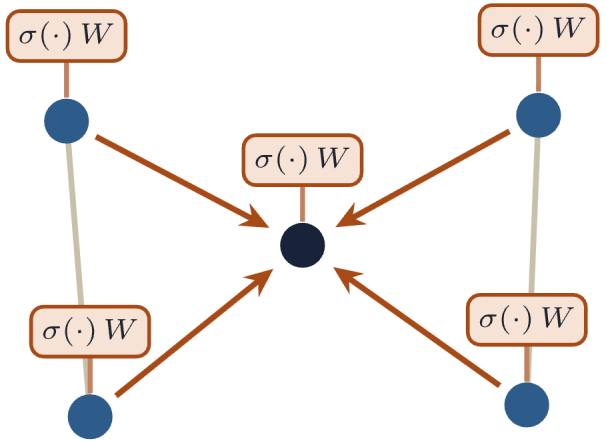
Graph structure in the forward pass



A standard NN on each node

$$x \longmapsto \sigma(x)\,W$$

$$x_u \longmapsto \sigma(x_u)\,W$$



A graph NN mixes across nodes

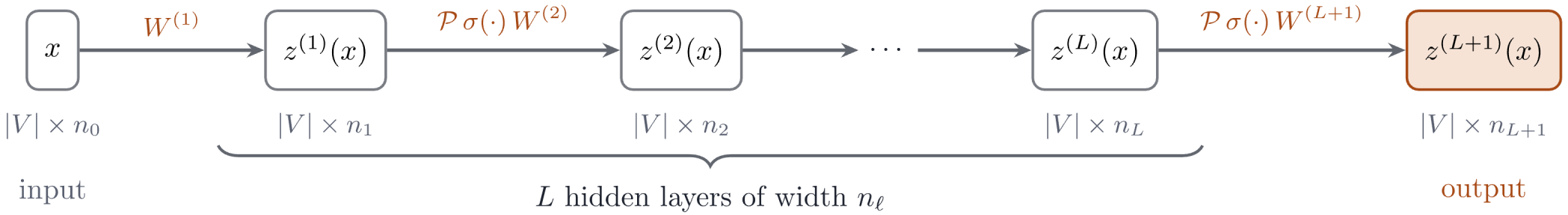
$$x \longmapsto A\,\sigma(x)\,W$$

$$x_u \longmapsto \sum_{v \in V} A_{uv}\,\sigma(x_v)\,W$$

$x_u =$ row u of x, $\quad \sigma(t) = \max(t, 0)$ entrywise, throughout the talk

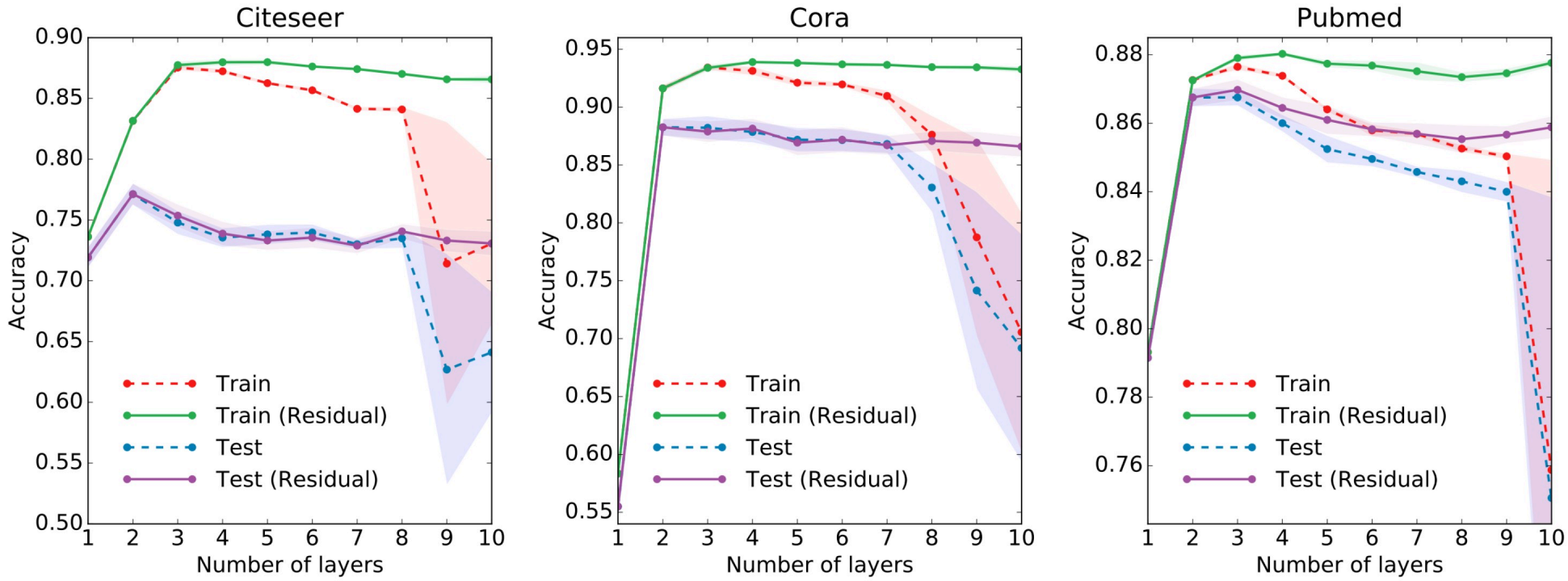
Generalizing the aggregation operator

$$z^{(\ell+1)}(x) = \mathcal{P}\,\sigma\big(z^{(\ell)}(x)\big)\,W^{(\ell+1)}, \qquad z^{(1)}(x) = x\,W^{(1)}$$



Assumption: $\mathcal{P}$ is symmetric, with nonnegative entries

Deep GCNs have been hard to train from the start



Best at 2–3 layers. Past about 7 layers, even *training* accuracy collapses without residual connections.

T. N. Kipf and M. Welling, Semi-supervised classification with graph convolutional networks, ICLR 2017, Appendix B, Fig. 5.

Three ways a deep GNN fails before training starts

FM1 **Exponential outputs:** activations and gradients explode or vanish with depth.

FM2 **Oversmoothing:** features converge to the top eigenspace of the message passing operator.

FM3 **Correlation collapse:** features of different classes become indistinguishable.

FM1: exponential outputs

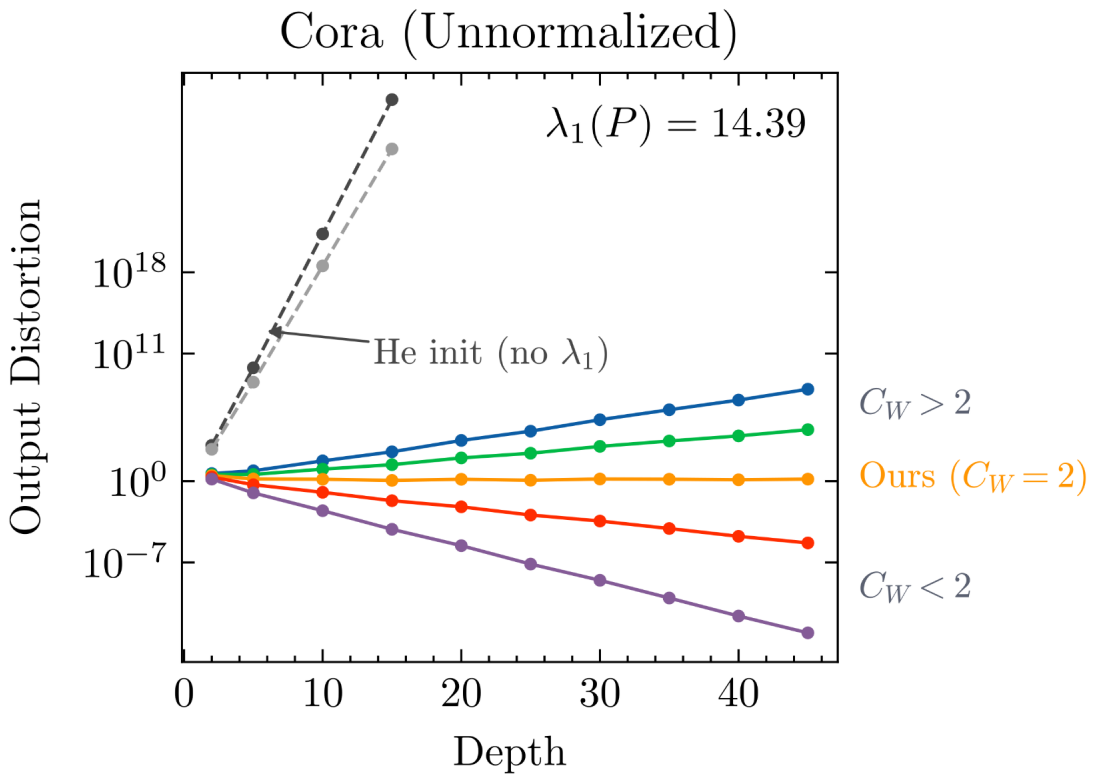
Model $z^{(\ell+1)}(x) = \mathcal{P}\,\sigma\big(z^{(\ell)}(x)\big)\,W^{(\ell+1)}$

Init $W_{ij}^{(\ell)} \sim \mathcal{N}\big(0,\, C_W^{(\ell)}\big)$

Def. output distortion $:= \dfrac{1}{n_L}\left\|z_v^{(L)}(x)\right\|^2 \Big/ \dfrac{1}{n_0}\left\|x_v\right\|^2$

Thm. output distortion $\simeq e^{\,cL}$ unless:

$$C_W^{(\ell)} = \underbrace{\frac{2}{n_{\ell-1}}}_{\text{He init (ReLU)}} \times \underbrace{\frac{1}{\lambda_1(\mathcal{P})^2}}_{\substack{\text{new: normalize} \\ \text{message passing}}}$$



Cora, unnormalized adjacency ($\lambda_1 = 14.39$); width 128, averaged over 100 random initializations. See Fig. 4, backup B10.

Q: How does the GCN fit in? It builds in the normalization

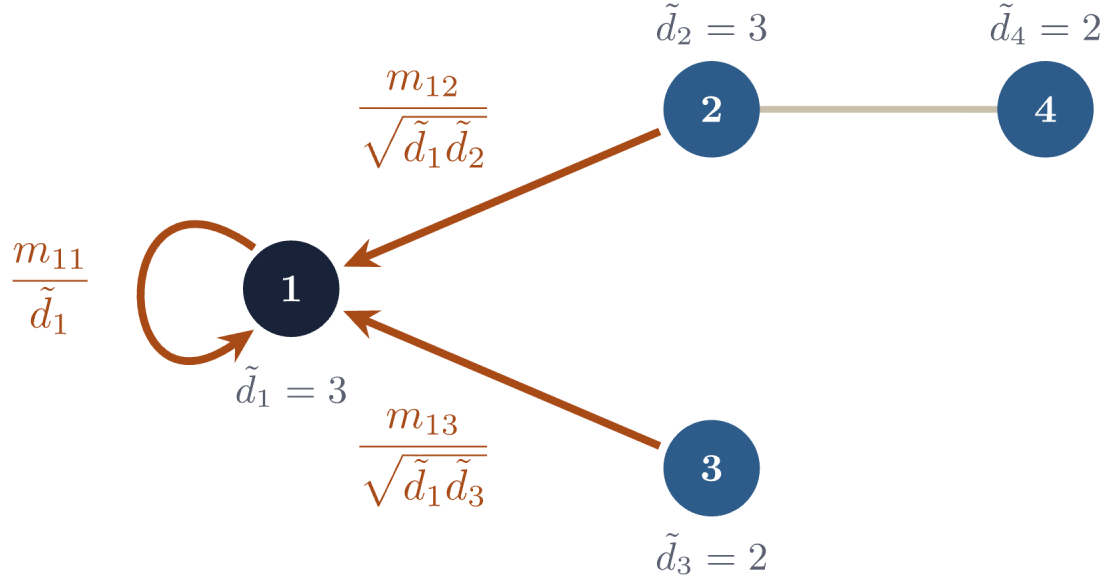
$$x \longmapsto \tilde{A}\,\sigma(x)\,W$$

$$\tilde{A} = \tilde{D}^{-1/2}\,(A + I)\,\tilde{D}^{-1/2}$$

$$\tilde{D} = \mathrm{diag}(\tilde{d}_1, \ldots, \tilde{d}_{|V|}), \quad \text{degrees of } A + I$$

Add a self-loop at every node, then normalize by degree on both sides. The top eigenvalue of Ã is exactly 1.

So plain He init is already right here. But λ_1 only fixes the scale; the rest of the spectrum drives FM2.



FM2: oversmoothing

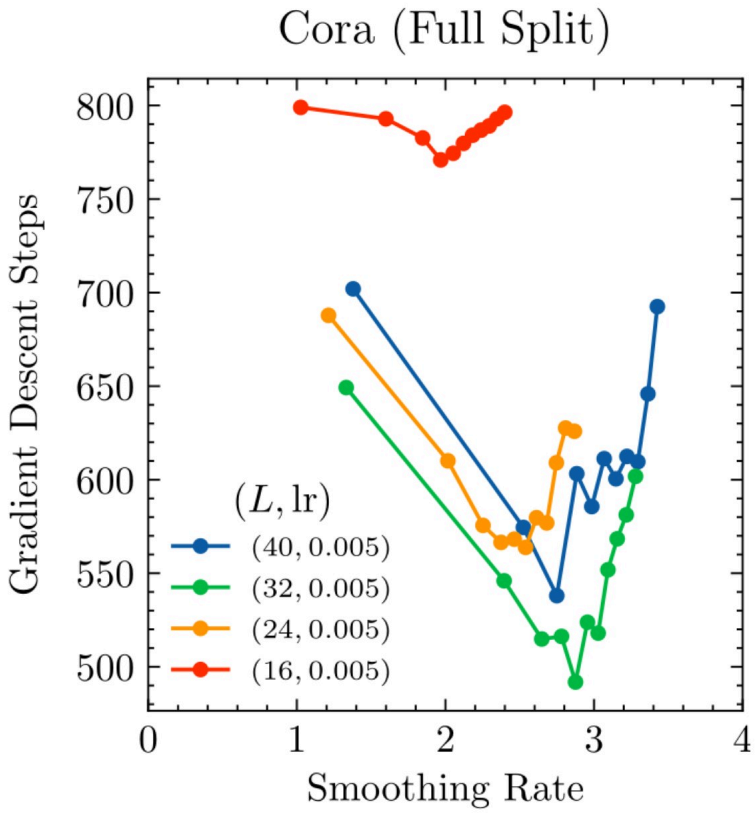
Def. $\Pi_1 =$ projection onto the top eigenspace of $\mathcal{P}$

Def. smoothing ratio: $$r^{(L)} = \frac{\mathbb{E}\left[\left\|\Pi_1 z^{(L)}(x)\right\|^2\right]}{\mathbb{E}\left[\left\|z^{(L)}(x)\right\|^2\right]}$$

As the ratio approaches 1, each feature converges to the top eigenvector of the message passing operator.

Thm. for any fixed $\mathcal{P}$: $$r^{(L)} \geq 1 - \exp\left[-cL\right]$$

Exponential in depth, for any initialization.



smoothing rate $:= -\log\left(1 - r^{(L)}\right)$

FM2: residual aggregation controls the smoothing

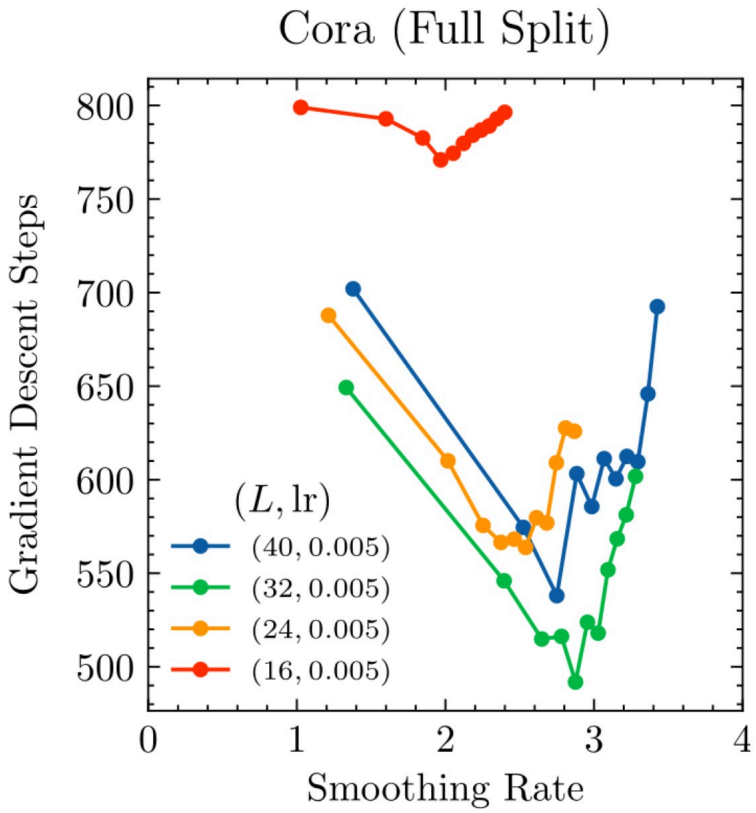
Before for any fixed $\mathcal{P}$: $r^{(L)} \geq 1 - \exp\left[-cL\right]$

With residual aggregation

$$\mathcal{P}_t = \left(1 - \frac{t}{L}\right) I + \frac{t}{L} \mathcal{P}$$

After with $\mathcal{P}_t$: $r^{(L)} \geq 1 - \exp\left[-ct\right]$

Smoothing is set by *t*, not by depth.



$$\text{smoothing rate} := -\log\left(1 - r^{(L)}\right)$$

FM3: correlation collapse

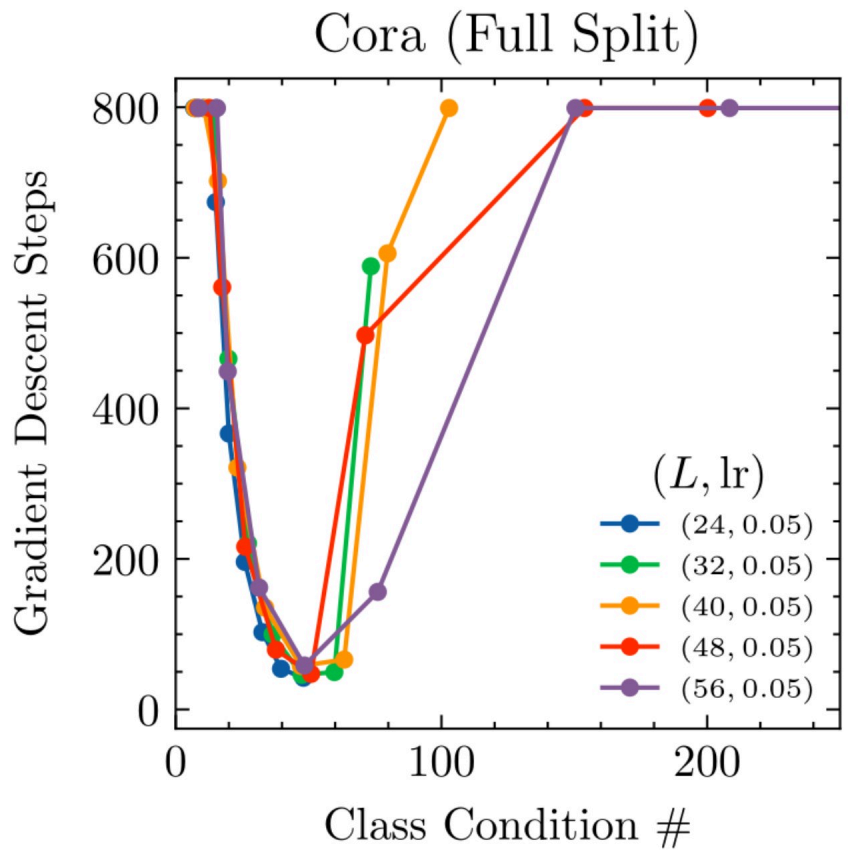
Def. class condition # $:= \kappa(M)$,

$$M = \begin{pmatrix} — & \bar{z}_1^{(L)}(x) & — \\ & \vdots & \\ — & \bar{z}_k^{(L)}(x) & — \end{pmatrix} \in \mathbb{R}^{k \times n_L}$$

$\bar{z}_c^{(L)}(x) =$ mean of $z_v^{(L)}(x)$ over vertices v of class c

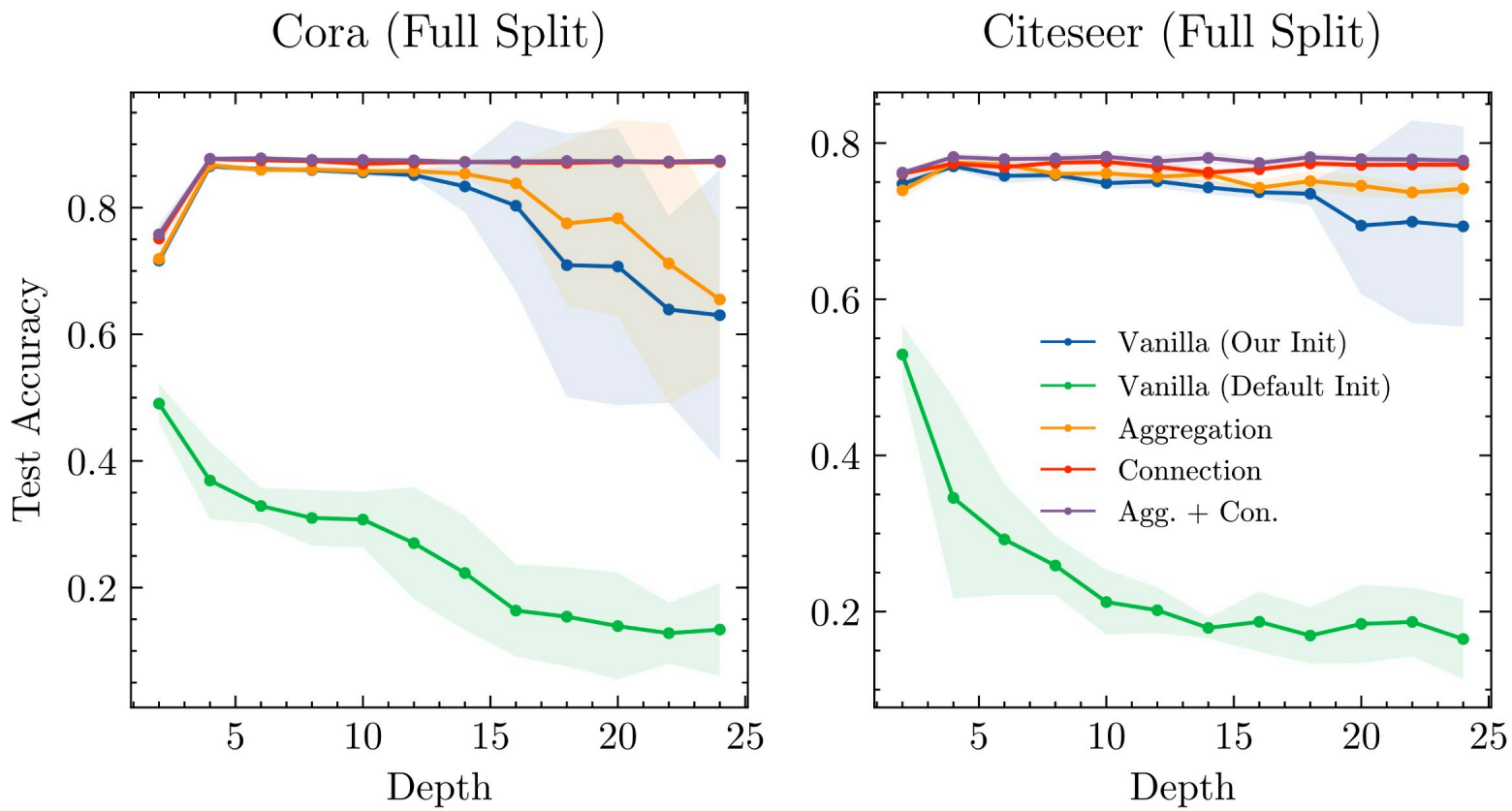
Fix. Residual connections keep class condition numbers moderate:

$$z^{(\ell+1)}(x) = z^{(\ell)}(x) + \frac{\beta}{\sqrt{L}} \mathcal{P}\, \sigma\big(z^{(\ell)}(x)\big)\, W^{(\ell+1)}$$



Cora, node classification. Condition # varied via β ∈ {0.1, ..., 0.9}, with *t* tuned per setting; 25 random inits. Fig. 8, backup B14.

Putting it together: trainable and accurate at depth



Test accuracy vs. depth, node classification (cross-entropy), tuned per depth, 15 inits each. Regression row: Fig. 1 (backup B7); time to train: Fig. 2 (B8).

Proof idea

Recursion

$$K^{(\ell+1)} = C_W^{(\ell+1)}\, n_\ell\, \mathcal{P}\, \mathbb{E}\Big[\sigma\big(z_i^{(\ell)}\big)\, \sigma\big(z_i^{(\ell)}\big)^T\Big]\, \mathcal{P}, \qquad K^{(\ell)} = \mathbb{E}\big[z_i^{(\ell)} z_i^{(\ell)T}\big] \in \mathbb{R}^{|V|\times|V|}$$

Upper

ReLU halves each second moment, and aggregation stretches by at most λ_1:

$$\operatorname{tr} K^{(\ell+1)} \;\le\; a_\ell \operatorname{tr} K^{(\ell)}, \qquad a_\ell := \tfrac{1}{2}\, C_W^{(\ell+1)}\, n_\ell\, \lambda_1(\mathcal{P})^2$$

Lower

The top eigenspace has a nonnegative basis, and ReLU keeps at least half of every correlation (Lemma A.2):

$$\operatorname{tr}\big(K^{(\ell+1)}\Pi_1\big) \;\ge\; a_\ell \operatorname{tr}\big(K^{(\ell)}\Pi_1\big)$$

Result

With the graph-aware variance every factor equals 1 (Theorem 4.1):

$$A_\ell \operatorname{tr}\big(K^{(1)}\Pi_1\big) \;\le\; \operatorname{tr} K^{(\ell+1)} \;\le\; A_\ell \operatorname{tr} K^{(1)}, \qquad A_\ell = \prod_{\ell' \le \ell} a_{\ell'} = 1 \;\text{ when }\; C_W^{(\ell+1)} = \frac{2}{n_\ell\, \lambda_1(\mathcal{P})^2}$$

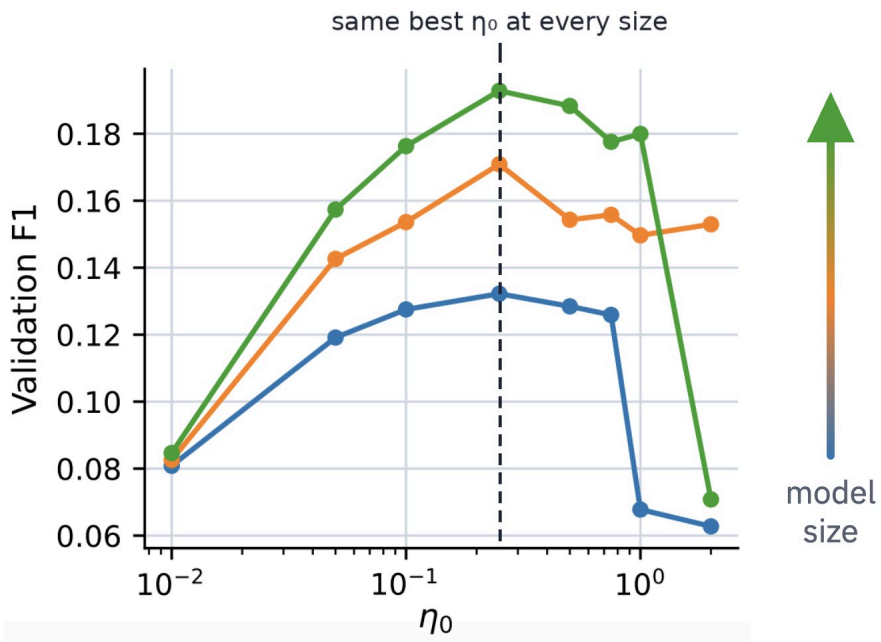
FM2

Divide the two bounds: the variance factor cancels, so for any initialization the smoothing ratio can only grow:

$$r^{(\ell+1)} = \frac{\operatorname{tr}\big(K^{(\ell+1)}\Pi_1\big)}{\operatorname{tr} K^{(\ell+1)}} \;\ge\; \frac{a_\ell \operatorname{tr}\big(K^{(\ell)}\Pi_1\big)}{a_\ell \operatorname{tr} K^{(\ell)}} = r^{(\ell)}$$

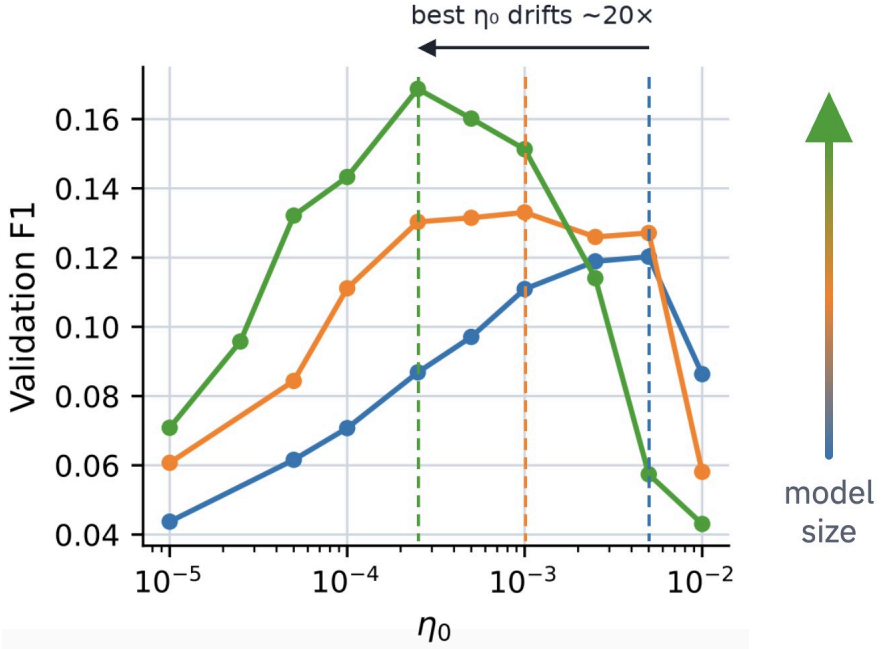
Addendum: hyperparameter transfer for GNNs

Our GNN parameterization



The best learning rate stays put, and bigger models do better.

Standard parameterization



The best learning rate drifts, so every new size needs its own sweep.

PascalVOC-SP, Adam; model sizes (L, D) = (2, 128), (4, 256), (6, 512): width and depth grow together. G. DeZoort and B. Hanin, arXiv:2607.05017 (2026), Fig. 1.

Thank you

Principles for Initialization and Architecture Selection in Graph Neural Networks with ReLU Activations
G. DeZoort and B. Hanin, SIAM J. Math. Data Sci. 7(1), 1–27 (2025)

Hyperparameter Transfer in Graph Neural Networks
G. DeZoort and B. Hanin, arXiv:2607.05017 (2026)

jdezoort@princeton.edu

Funding: DOE DE-SC0007968 · NSF CAREER DMS-2143754, DMS-1855684, DMS-2133806 · ONR MURI