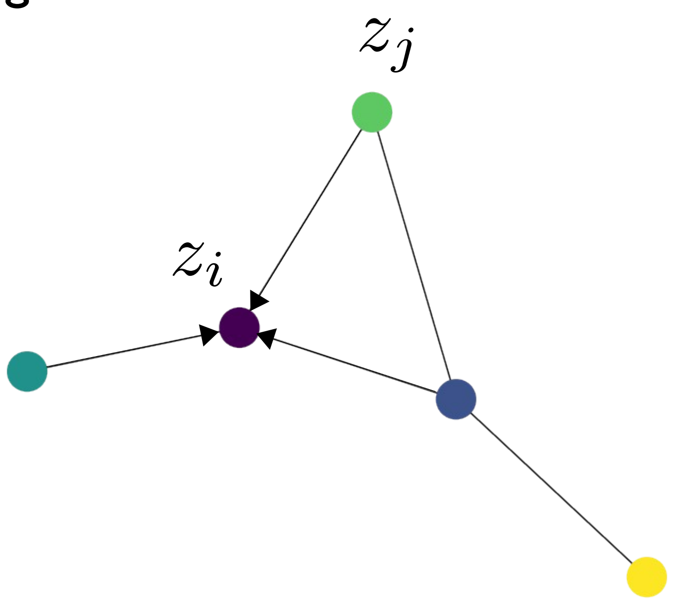
Not too little, not too much:
a theoretical analysis of graph (over)smoothing

Nicolas Keriven
CNRS, GIPSA-lab

Graph Neural Networks: Message-passing

Graph Neural Networks (GNNs) work mostly by **Message-Passing**:

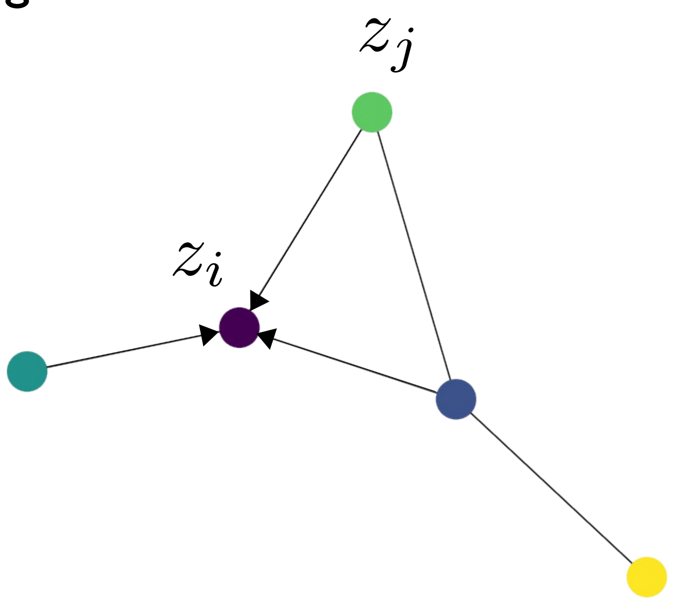
$$z_i^{(k)} = \mathrm{AGG}_{\theta_k}\left(z_i^{(k-1)}, \{z_j^{(k-1)}\}_{j \in \mathcal{N}_i}\right)$$



Graph Neural Networks: Message-passing

Graph Neural Networks (GNNs) work mostly by **Message-Passing**:

$$z_i^{(k)} = \mathrm{AGG}_{\theta_k}(z_i^{(k-1)}, \{z_j^{(k-1)}\}_{j \in \mathcal{N}_i})$$



Here we use classic **mean aggregation**:

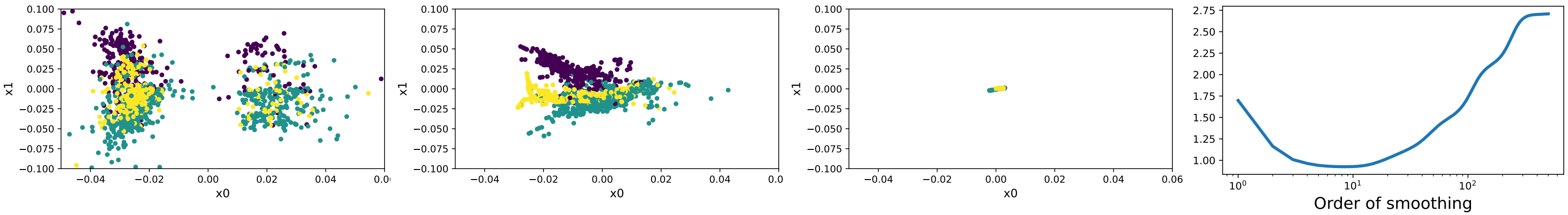
$$z_i^{(k)} = \frac{1}{\sum_j a_{ij}} \sum_j a_{ij} \Psi_{\theta_k}(z_j^{(k-1)})$$

Note that this is just $Z^{(k)} = L\Psi_{\theta_k}(Z^{(k-1)})$ with $L = D^{-1}A$

Overmoothing vs Sufficient depth

Oversmoothing is a well known phenomenon “preventing” GNNs from being “too deep” in practice. E.g., for mean aggregation: $L^k Z \xrightarrow[k\to\infty]{} c1_n$

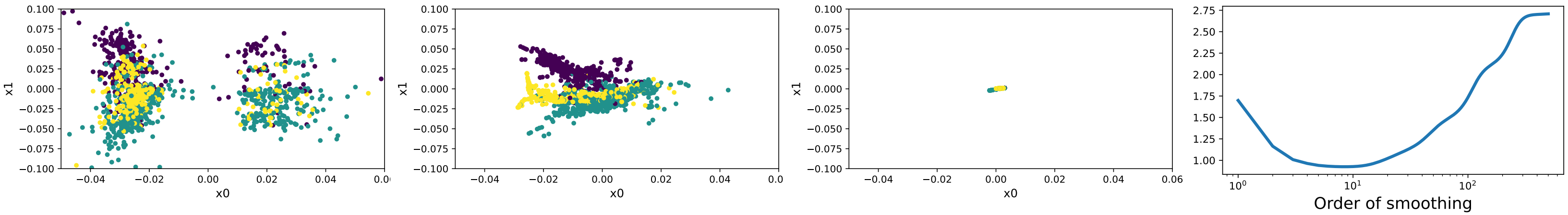
Cora



Overmoothing vs Sufficient depth

Oversmoothing is a well known phenomenon “preventing” GNNs from being “too deep” in practice. E.g., for mean aggregation: $L^k Z \xrightarrow[k\to\infty]{} c1_n$

Cora



But... most analyses showing the power of GNNs **take the limit** $k \to \infty$ **!**

(*not* for mean aggregation, obviously)

- sufficiently deep GNNs are “**Weisfeiler-Lehman**” powerful *[Xu et al. 2019]*

- some GNNs model a **diffusion process** that separates well data, etc

[Bodnar et al. 2022]

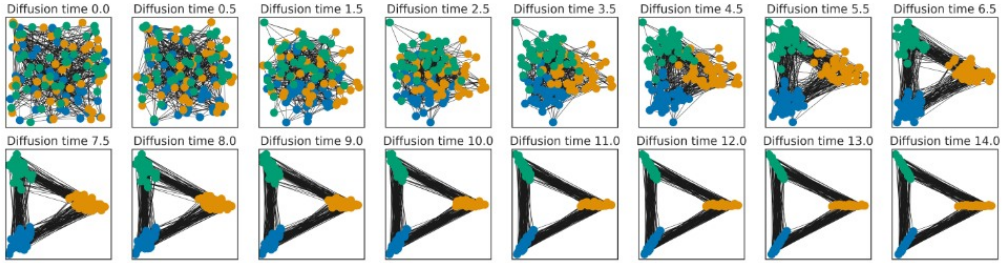
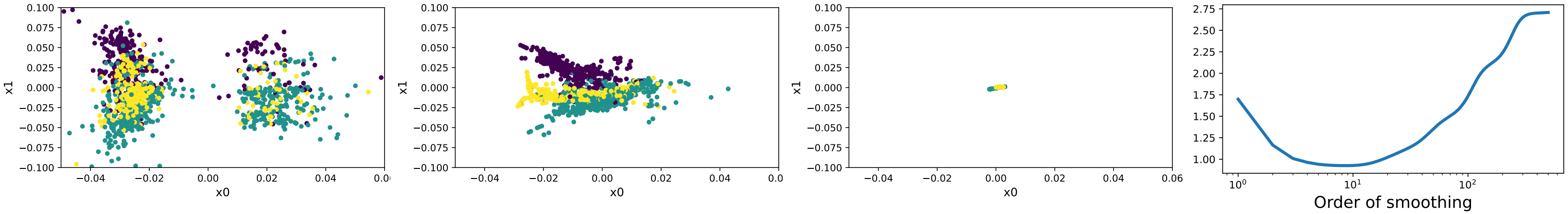


Figure 7. Sheaf diffusion process disentangling the $C = 3$ classes over time. The nodes are coloured by their class.

Oversmoothing vs Sufficient depth

Oversmoothing is a well known phenomenon "preventing" GNNs from being "too deep" in practice. E.g., for mean aggregation: $L^k Z \xrightarrow[k\to\infty]{} c1_n$

Cora



But... most analyses showing the power of GNNs **take the limit** $k \to \infty$ **!**

(*not* for mean aggregation, obviously)

- sufficiently deep GNNs are "**Weisfeiler-Lehman**" powerful *[Xu et al. 2019]*

- some GNNs model a **diffusion process** that separates well data, etc

[Bodnar et al. 2022]

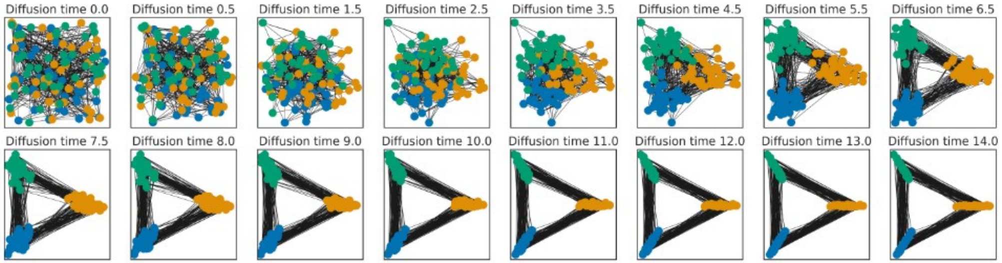


Figure 7. Sheaf diffusion process disentangling the $C = 3$ classes over time. The nodes are coloured by their class.

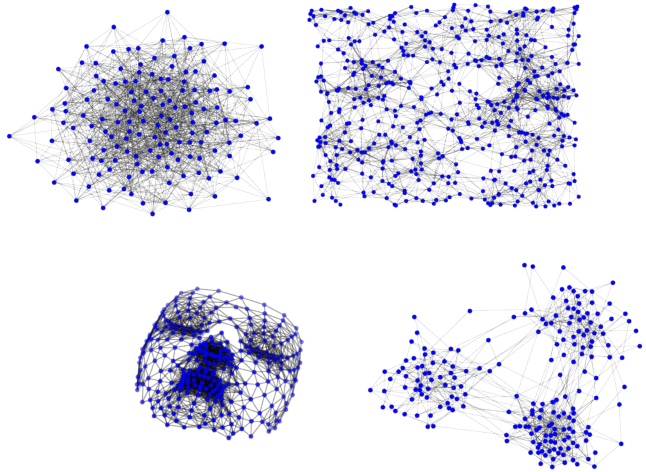
Can "good smoothing" and oversmoothing co-exist?

Settings: Ridge Regression and random graphs

Random graph model:

$$(x_i, y_i) \sim P, \;\; a_{ij} = W(x_i, x_j), \;\; z_i = M x_i$$

With $M \in \mathbb{R}^{p \times d}$, $p < d$ $\quad W(x, x') = e^{-\|x - x'\|^2} + \epsilon$

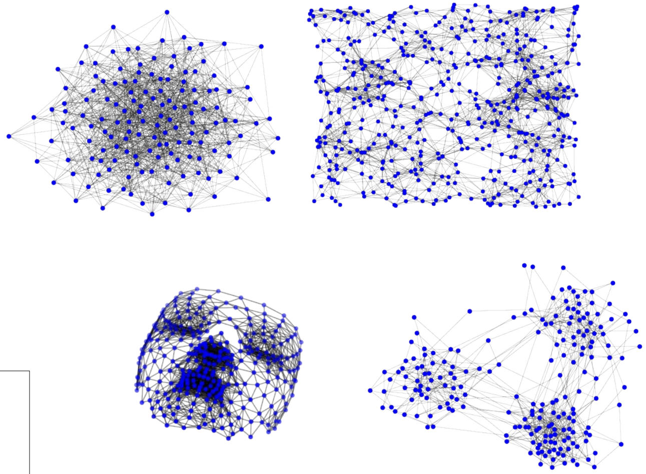


Settings: Ridge Regression and random graphs

Random graph model:

$$(x_i, y_i) \sim P, \;\; a_{ij} = W(x_i, x_j), \;\; z_i = Mx_i$$

With $\quad M \in \mathbb{R}^{p\times d}, \quad p < d \qquad W(x,x') = e^{-\|x-x'\|^2} + \epsilon$

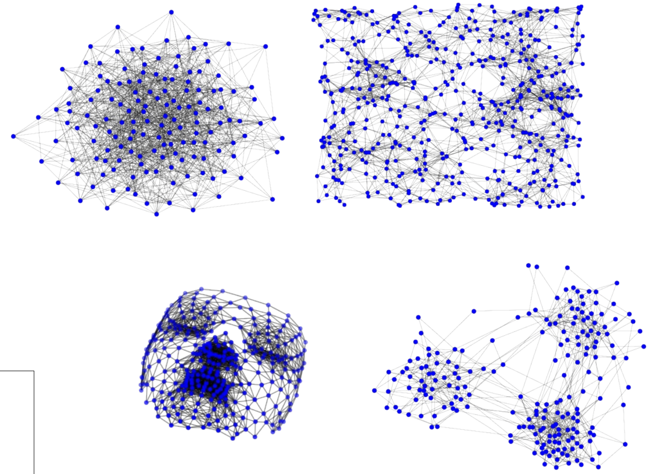


There **is** loss of information in the node features.

Can smoothing recover some of it before oversmoothing occurs ?

Settings: Ridge Regression and random graphs

Random graph model:



$$\boxed{(x_i, y_i) \sim P, \;\; a_{ij} = W(x_i, x_j), \;\; z_i = M x_i}$$

With $\quad M \in \mathbb{R}^{p\times d}, \quad p < d \qquad W(x,x') = e^{-\|x-x'\|^2} + \epsilon$

There **is** loss of information in the node features.

Can smoothing recover some of it before oversmoothing occurs ?

Linear Ridge Regression *(paper: SSL)*

$$\mathcal{R}^{(k)} = \min_\beta \frac{1}{n}\|L^k Z\beta - Y\|^2 + \lambda\|\beta\|^2$$

Goal: show there is $k^\star$ s.t.

$$\mathcal{R}^{(k^\star)} < \min(\mathcal{R}^{(0)}, \mathcal{R}^{(\infty)})$$

Regression

Regression settings: $x \sim \mathcal{N}(0, \Sigma), \quad y = x^\top \beta^\star$

Thm: if $\Sigma, \beta^\star, M$ are "well-aligned" and n is large enough, $k^\star$ exists.

Regression

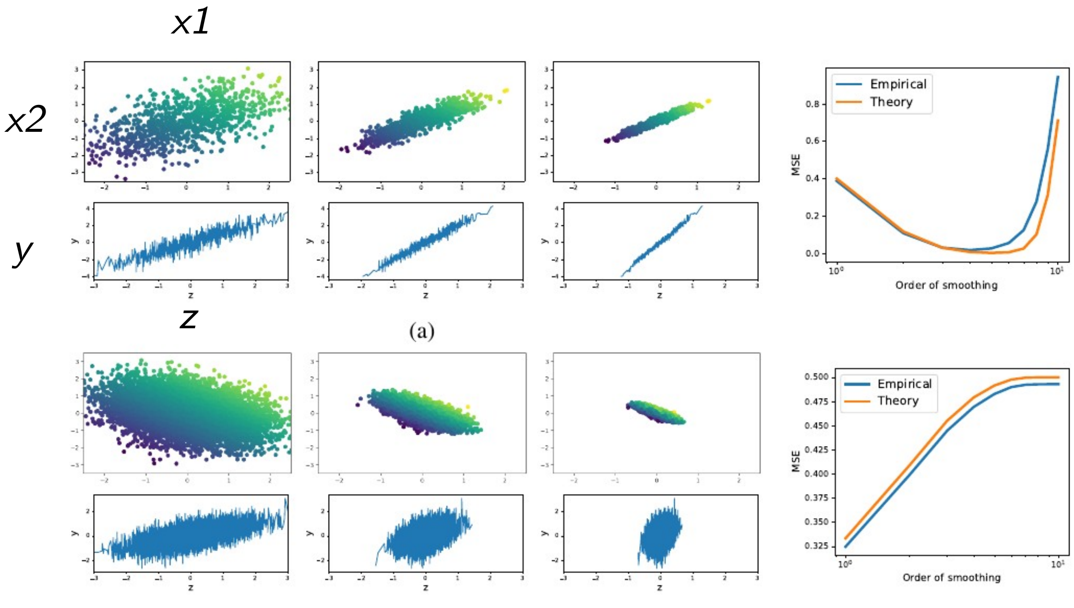
Regression settings: $x \sim \mathcal{N}(0, \Sigma), \quad y = x^\top \beta^\star$

Thm: if $\Sigma, \beta^\star, M$ are "well-aligned" and n is large enough, $k^\star$ exists.

Intuition: $L^k X$ behaves almost as

$$\mathcal{N}(0, (\mathrm{Id} + \Sigma^{-1})^{-k}\Sigma)$$

*The small eigenvalues shrink **faster** than the large ones.*
If $\beta^\star$ is aligned with the large, smoothing reduces noise in z



(a)
(b)

Classification

Classif. settings: $(x, y) \sim \frac{1}{2}\mathcal{N}(\mu, \mathrm{Id}) \otimes \{1\} + \frac{1}{2}\mathcal{N}(-\mu, \mathrm{Id}) \otimes \{-1\}$

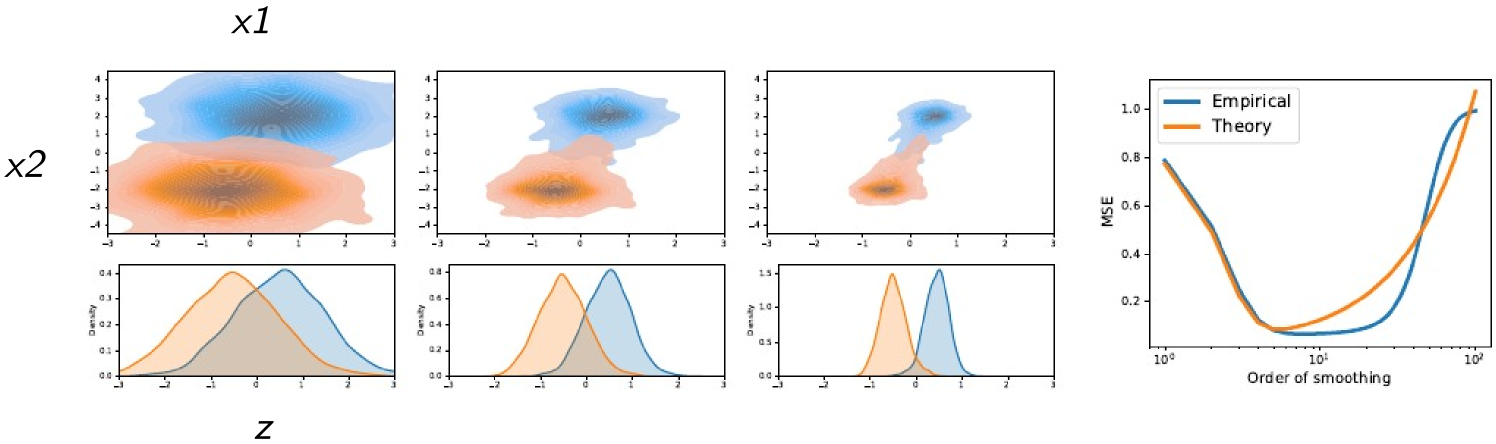
$\boxed{\textbf{Thm}\text{: if } \|\mu\|, n \text{ are large enough and } \|M\mu\| > 0 \,,\; k^\star \text{ exists.}}$

Classification

Classif. settings: $$(x,y) \sim \frac{1}{2}\mathcal{N}(\mu, \mathrm{Id}) \otimes \{1\} + \frac{1}{2}\mathcal{N}(-\mu, \mathrm{Id}) \otimes \{-1\}$$

Thm: if $\|\mu\|, n$ are large enough and $\|M\mu\| > 0$, $k^\star$ exists.

Intuition: *The communities (initially) concentrate faster than they get close to each other.*



Summary, outlooks

*We provided **simple examples** where beneficial smoothing and oversmoothing provably co-exist.*

Outlooks

- More realistic random graphs models
- More complex loss functions, learning methods...
- Actual GNN architectures!
- Link with existing work to combat oversmoothing, improvement?