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

Nicolas Keriven
CNRS, GIPSA-lab

NeurIPS 2022 (Oral)
LoG 2022 (extended abstract, spotlight)
Graph Neural Networks (GNNs) work mostly by **Message-Passing**:

\[
z_i^{(k)} = \text{AGG}_{\theta_k}(z_i^{(k-1)}, \{z_j^{(k-1)}\}_{j \in \mathcal{N}_i})
\]
Graph Neural Networks (GNNs) work mostly by **Message-Passing**:

\[
Z_i^{(k)} = \text{AGG}_{\theta_k}(Z_i^{(k-1)}, \{Z_j^{(k-1)}\}_{j \in \mathcal{N}_i})
\]

Here we use classic **mean aggregation**:

\[
Z_i^{(k)} = \frac{1}{\sum_j a_{i,j}} \sum_j a_{i,j} \Psi_{\theta_k}(Z_j^{(k-1)})
\]

Note that this is just \(Z^{(k)} = L\Psi_{\theta_k}(Z^{(k-1)})\) with \(L = D^{-1}A\).
Oversmoothing is a well-studied phenomenon “preventing” GNNs from being “too deep” in practice. E.g., for mean aggregation: $L^k Z \xrightarrow{k \to \infty} c_1 n$
Oversmoothing is a well-studied phenomenon “preventing” GNNs from being “too deep” in practice. E.g., for mean aggregation:

\[ L^k Z \xrightarrow{k \to \infty} c1_n \]

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

(\textit{not} for mean aggregation, obviously)

- sufficiently deep GNNs are “Weisfeiler-Lehman” powerful \cite{xu2019powerful}
- some GNNs model a \textit{diffusion process} that separates well data, etc

Figure 7. Sheaf diffusion process disentangling the \( C = 3 \) classes over time. The nodes are coloured by their class.
Oversmoothing vs Sufficient depth

Can “good smoothing” and oversmoothing co-exist? Why?

Middle regime?
Oversmoothing vs Sufficient depth

Can “good smoothing” and oversmoothing co-exist? Why?

Take-home message: smoothing collapses node features, but not everything collapses at the same speed
Model of random graph

Random graph model:

\[(x_i, y_i) \sim P, \quad a_{i,j} = W(x_i, x_j), \quad z_i = Mx_i\]

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

\[(x_i, y_i) \sim P, \quad a_{ij} = W(x_i, x_j), \quad z_i = Mx_i\]

With \(M \in \mathbb{R}^{p \times d}, \quad p < d\)

\[W(x, x') = e^{-\|x-x\|^2} + \epsilon\]

No Johnson-Lindenstrauss here. There is **loss of information** in the node features.
Model of random graph

Random graph model:

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

With \[M \in \mathbb{R}^{p \times d}, \ p < d\]

\[W(x, x') = e^{-\|x - x'\|^2} + \epsilon\]

No Johnson-Lindenstrauss here. There is loss of information in the node features.

Can mean aggregation recover some of the information before oversmoothing occurs?
Settings: Ridge Regression and SSL

- **Linear GNN** *(also called SGC [Wu et al. 2019])*

\[ \hat{Y} = Z^{(k)} \beta \text{ with } Z^{(k)} = L^k Z \]
Settings: Ridge Regression and SSL

- **Linear GNN**  *(also called SGC [Wu et al. 2019])*
  \[
  \hat{Y} = Z^{(k)} \beta \text{ with } Z^{(k)} = L^k Z
  \]

- **Semi-Supervised Learning**  \( n_{tr}, n_{te} \sim n \)
Settings: Ridge Regression and SSL

- **Linear GNN** *(also called SGC [Wu et al. 2019])*
  \[
  \hat{Y} = Z^{(k)} \beta \text{ with } Z^{(k)} = L^k Z
  \]

- **Semi-Supervised Learning** \(n_{tr}, n_{te} \sim n\)

- **Ridge Regression**
  \[
  \beta^{(k)} = \arg \min_\beta \frac{1}{n_{tr}} \| Z^{(k)}_{tr} \beta - Y_{tr} \|^2 + \lambda \| \beta \|^2
  \]
Settings: Ridge Regression and SSL

- **Linear GNN** *(also called SGC [Wu et al. 2019]*)

\[ \hat{Y} = Z^{(k)} \beta \text{ with } Z^{(k)} = L^k Z \]

- **Semi-Supervised Learning** \( n_{tr}, n_{te} \sim n \)

- **Ridge Regression**

\[ \beta^{(k)} = \arg \min_{\beta} \frac{1}{n_{tr}} \| Z_{tr}^{(k)} \beta - Y_{tr} \|^2 + \lambda \| \beta \|^2 \]

- **Test risk**

\[ R^{(k)} = \frac{1}{n_{te}} \| Y_{te} - Z_{te}^{(k)} \beta^{(k)} \|^2 \]
Settings: Ridge Regression and SSL

- **Linear GNN** (also called SGC [Wu et al. 2019])
  \[ \hat{Y} = Z^{(k)} \beta \] with \( Z^{(k)} = L^k Z \)

- **Semi-Supervised Learning** \( n_{tr}, n_{te} \sim n \)

- **Ridge Regression**
  \[ \beta^{(k)} = \arg \min_{\beta} \frac{1}{n_{tr}} \| Z_{tr}^{(k)} \beta - Y_{tr} \|^2 + \lambda \| \beta \|^2 \]

- **Test risk**
  \[ \mathcal{R}^{(k)} = \frac{1}{n_{te}} \| Y_{te} - Z_{te}^{(k)} \beta^{(k)} \|^2 \]

**Thm:** Oversmoothing
\[ Z_{te}^{(k)} \beta^{(k)} \xrightarrow{k \to \infty} C1 n_{te} \]
Settings: Ridge Regression and SSL

- **Linear GNN** (also called SGC [Wu et al. 2019])
  \[ \hat{Y} = Z^{(k)} \beta \text{ with } Z^{(k)} = L^k Z \]
- **Semi-Supervised Learning** \( n_{tr}, n_{te} \sim n \)
- **Ridge Regression**
  \[ \beta^{(k)} = \arg \min_{\beta} \frac{1}{n_{tr}} \| Z_{tr}^{(k)} \beta - Y_{tr} \|^2 + \lambda \| \beta \|^2 \]
- **Test risk**
  \[ R^{(k)} = \frac{1}{n_{te}} \| Y_{te} - Z_{te}^{(k)} \beta^{(k)} \|^2 \]

**Thm:** Oversmoothing
\[ Z_{te}^{(k)} \beta^{(k)} \xrightarrow{k \to \infty} C1n_{te} \]

**Goal:** show there is \( k^* \) s.t.
\[ R^{(k^*)} < \min(R^{(0)}, R^{(\infty)}) \]
Regression

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

**Thm:** if \( \Sigma, \beta^*, M \) are “well-aligned” and \( n \) is large enough, \( k^* \) exists.
Regression

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

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

**Intuition:** $L^k X$ behaves “almost” as

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

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

**Thm:** If \( \Sigma, \beta^*, M \) are “well-aligned” and \( n \) is large enough, \( k^* \) exists.

**Intuition:** \( L^k X \) behaves “almost” as
\[
\mathcal{N}(0, (\text{Id} + \Sigma^{-1})^{-k} \Sigma)
\]

- The small eigenvalues shrink **faster** than the large ones
  \[ \lambda_i \leftarrow \lambda_i/(1 + 1/\lambda_i)^k \]
Regression

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

**Thm:** if \( \Sigma, \beta^*, M \) are “well-aligned” and \( n \) is large enough, \( k^* \) exists.

**Intuition:** \( L^k X \) behaves “almost” as 
\[
\mathcal{N}(0, (\text{Id} + \Sigma^{-1})^{-k} \Sigma)
\]

- The small eigenvalues shrink faster than the large ones \( \lambda_i \leftarrow \lambda_i/(1 + 1/\lambda_i)^k \)
- If well-aligned ("homophily"), smoothing helps
- If inversely aligned ("heterophily"), smoothing never helps
Regression

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

**Thm:** if \( \Sigma, \beta^*, M \) are “well-aligned” and \( n \) is large enough, \( k^* \) exists.

**Intuition:** \( \mathcal{L}^k X \) behaves “almost” as \( \mathcal{N}(0, (\text{Id} + \Sigma^{-1})^{-k} \Sigma) \)

- The small eigenvalues shrink faster than the large ones \( \lambda_i \leftarrow \lambda_i / (1 + 1/\lambda_i)^k \)
- If well-aligned (“homophily”), smoothing helps
- If inversely aligned (“heterophily”), smoothing never helps
- Proof not that simple: for \( k > 0 \), dependent rows of \( Z \)
Classification

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

**Thm:** if \(||\mu||, n\) are large enough and \(||M\mu|| > 0\), \(k^*\) exists.
Thm: if $\|\mu\|, n$ are large enough and $\|M\mu\| > 0$, $k^*$ 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. As expected, there are links with heterophily/homophily.

Outlooks

- Take inspiration to “combat” oversmoothing less indiscriminatively?

- How to better describe and exploit the interactions between labels, node features and graph structure?

Keriven N. Not too little, not too much: a theoretical analysis of graph (over)smoothing. NeurIPS 2022 (Oral)

gipsa-lab