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

3/6/23

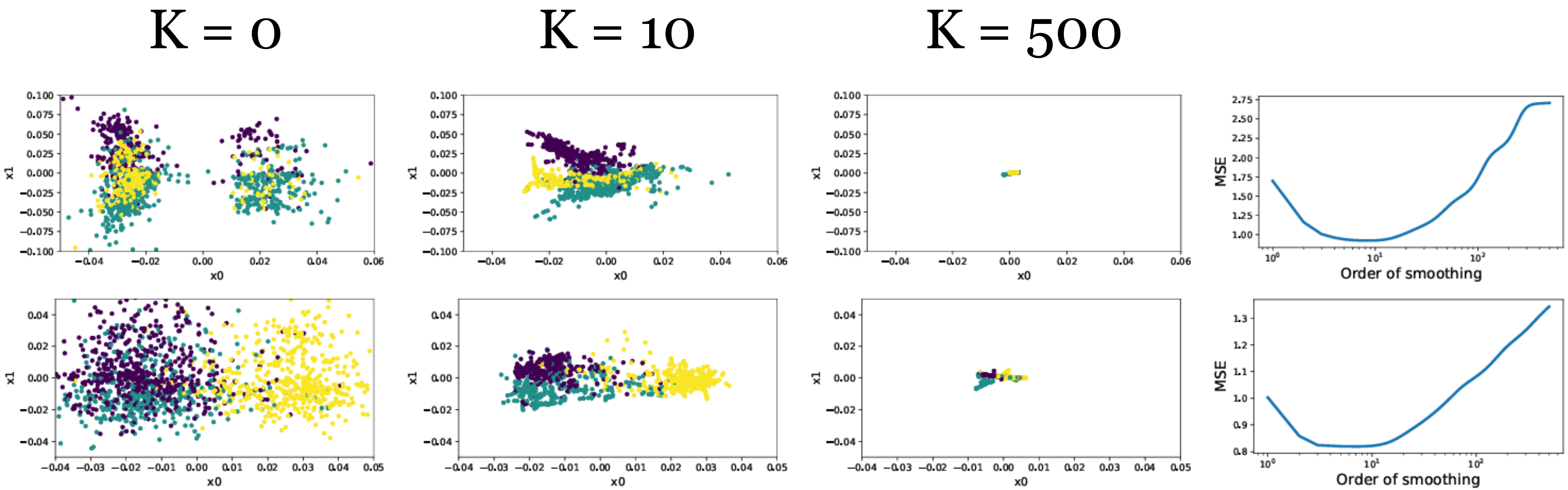
Presenter: Junhao Lin
Paper author: Nicolas Keriven

UNIVERSITY OF WATERLOO

What Is The Problem?

- Some variant of Message-Passing(MP) with repeated aggregation, may suffer from over-smoothing
- For mean aggregation, for connected graphs, the node features become constant
- A finite number of rounds of MP can improve performance, but a lack of theoretical research showing that some smoothing is useful for learning and explaining why it is beneficial

Visualization of Over-smoothing



UNIVERSITY OF WATERLOO

Solution Proposed By The Author

- Conduct theoretical research based on linear GNNs and random graphs
- Rigorously analyze two examples: one regression and one classification
- Prove that a finite number of mean aggregation steps improves the learning performance, before over-smoothing kicks in

Related Work

- Applying graph smoothing operators induces convergence of the node features: *Graph Neural Networks Exponentially Lose Expressive Power for Node Classification*
- Residual mechanisms: *Simple and deep graph convolutional networks*
- Randomly dropping connections: *Tackling Over Smoothing for General Graph Convolutional Networks*
- Introducing local jumps: *Representation learning on graphs with jumping knowledge networks*

SETTING

Semi-Supervised Learning

- Observe a weighted adjacency matrix $A = [a_{ij}]_{i,j=1}^{n} \in \mathbb{R}_{+}^{n \times n}$
- Observe node features $Z \in \mathbb{R}^{n \times p}$ of the graph
- Observe some labels Y_{tr} at training time and aim to predict the remaining labels Y_{te}

Architecture and Loss

- We will focus on Linear GCN with Mean Square Error (MSE)
- The input feature after k rounds of mean aggregation is

$$Z^{(k)} = L^k Z$$

- Learning with MSE loss and Ridge regularization

$$\hat{\beta}^{(k)} \overset{\text{def.}}{=} \operatorname{argmin}_{\beta} \frac{1}{2n_{\text{tr}}} \left\| Y_{\text{tr}} - Z_{\text{tr}}^{(k)} \beta \right\|^2 + \lambda \left\| \beta \right\|^2 = \left(\frac{(Z_{\text{tr}}^{(k)})^\top Z_{\text{tr}}^{(k)}}{n_{\text{tr}}} + \lambda \mathrm{Id} \right)^{-1} \frac{(Z_{\text{tr}}^{(k)})^\top Y_{\text{tr}}}{n_{\text{tr}}}$$

Test Risk

$$\mathcal{R}^{(k)} \stackrel{\text{def.}}{=} n_{\text{te}}^{-1} \left\| Y_{\text{te}} - \hat{Y_{\text{te}}}^{(k)} \right\|^2 \quad \text{where } \hat{Y_{\text{te}}}^{(k)} = Z_{\text{te}}^{(k)} \hat{\beta}^{(k)}$$

- $\mathcal{R}^{(0)}$ is the risk for directly performing linear regression without smoothing
- $\mathcal{R}^{(\infty)}$ is the asymptotic test risk as k -> ∞
- Over-smoothing: $\mathcal{R}^{(0)} < \mathcal{R}^{(\infty)}$
- Goal: $\mathcal{R}^{(1)} < \mathcal{R}^{(0)}$, and therefore $\mathcal{R}^{(k^\star)} < \min(\mathcal{R}^{(0)}, \mathcal{R}^{(\infty)})$

Latent Space Random Graphs

- Unobserved latent variable x_i with dimension d
- Node features z_i with dimension p are linear projection of x_i , d >= p
- Edge between x_i and x_j is denoted by $a_{ij} = W(x_i, x_j)$
- W is a connectivity kernel (Gaussian kernel used in this paper)

$$\forall i, j, \quad (x_i, y_i) \overset{iid}{\sim} P, \quad z_i = M^\top x_i, \quad a_{ij} = W(x_i, x_j)$$

$$W(x, y) = \varepsilon + W_g(x, y) \quad \text{where } W_g(x, y) \overset{\text{def.}}{=} e^{-\frac{1}{2}\|x-y\|^2}$$

UNIVERSITY OF WATERLOO

Mean Aggregation

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

$$z_i^{(k)} = \frac{1}{\sum_j a_{ij}} \sum_j a_{ij} \Psi\left(z_j^{(k-1)}\right)$$

- $z_i^{(k)}$ are the smoothed features after k steps of mean aggregation
- a_{ij} are the entries of the adjacency matrix, and Ψ is some function (usually a Multi-Layer Perceptron).

Mathematical Explanation for Over-smoothing

Theorem 1 (Ergodic theorem for stochastic matrices, e.g. [2, Thm. 4.2].). *Recall that d_A is the vector of degrees, let $\bar{d}_A = d_A / d_A^\top 1_n$. We have*

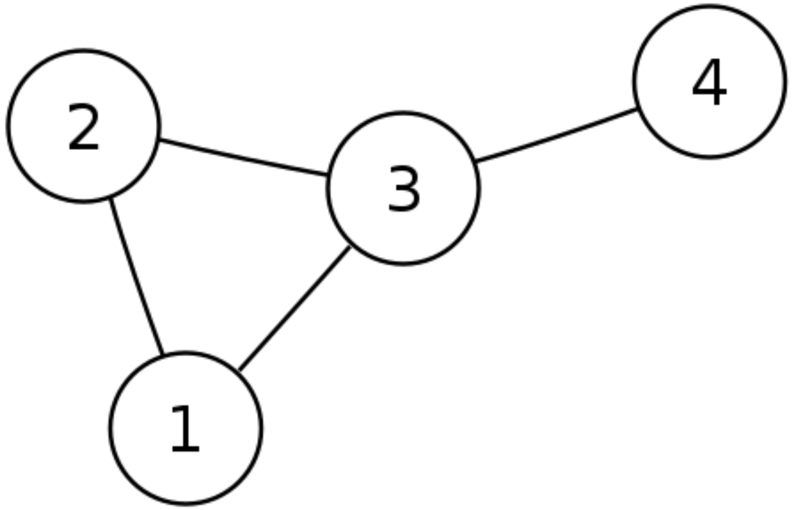
$$L^k \xrightarrow[k \to \infty]{} 1_n \bar{d}^\top \tag{7}$$

- For for an irreducible and aperiodic stochastic matrix P, there exists a unique probability vector π such that $\pi = \pi P$
- For certain types of stochastic matrices, repeatedly applying the matrix to a probability vector will eventually converge to a unique stationary distribution.

UNIVERSITY OF WATERLOO

Mathematical Explanation for Over-smoothing

$$P = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 & 0 \end{bmatrix}$$



$$\pi = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{3}{8} & \frac{1}{8} \end{bmatrix}$$

$$\pi \cdot P = \begin{bmatrix} \frac{1}{8}+\frac{1}{8} & \frac{1}{8}+\frac{1}{8} & \frac{1}{8}+\frac{1}{8}+\frac{1}{8} & \frac{1}{8} \end{bmatrix} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{3}{8} & \frac{1}{8} \end{bmatrix} = \pi$$

Mathematical Explanation for Over-smoothing

Corollary 1. *We have the following*

$$\hat{Y}_{\text{te}}^{(k)} \xrightarrow[k\to\infty]{} \left(\frac{\|v\|^2}{\lambda+\|v\|^2}\bar{y}_{\text{tr}}\right) 1_{n_{\text{te}}} \tag{8}$$

where $v = Z^\top \bar{d}$ *and* $\bar{y}_{\text{tr}} = n_{\text{tr}}^{-1}\sum_{i=1}^{n_{\text{tr}}} y_i$.

Average of the training labels

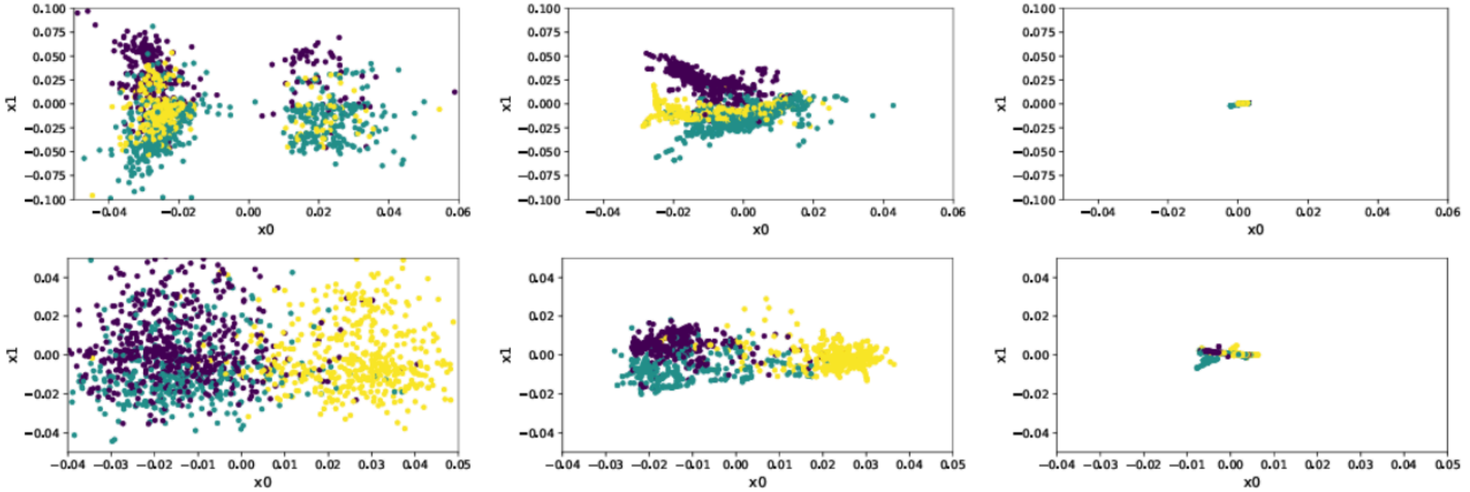
As a result $\mathcal{R}^{(\infty)} \approx \mathrm{Var}(y) + \mathcal{O}\left(1/\sqrt{n}\right)$

UNIVERSITY OF WATERLOO

RESULTS

How to prove that “beneficial” smoothing exists?

- Deriving an explicit equation for risk after smoothing is hard
- Hints from the visuals of smoothing: variance of samples decrease
- Can we show that the variance of samples decrease after smoothing?
- Can we show that a (relatively) lower variance leads to lower risk?



Finite Smoothing: Linear Regression

$x \sim \mathcal{N}_{0,\Sigma}$, without noise for simplicity, $y = x^\top \beta^\star$

Step 1: Get an estimation of risk in terms of variance

$$R_{\text{reg.}}(S) \overset{\text{def.}}{=} (\Sigma^{\frac{1}{2}}\beta^\star)^\top \left(\text{Id} - S^{\frac{1}{2}} M(\lambda\text{Id} + M^\top S M)^{-1} M^\top S^{\frac{1}{2}}\right)^2 (\Sigma^{\frac{1}{2}}\beta^\star) \in \mathbb{R}_+$$

Assumption (but not always true!):

$$R_{\text{reg.}}(\Sigma) > R_{\text{reg.}}((\text{Id} + \Sigma^{-1})^{-2}\Sigma)$$

Finite Smoothing: Linear Regression

$$d(x) = |\mathrm{Id} + \Sigma|^{-\frac{1}{2}} \, e^{-\frac{1}{2}\|x\|^2_{(\mathrm{Id}+\Sigma)^{-1}}}$$

$$\varphi_{\text{reg.}}(x) = \frac{d(x)}{d(x)+\varepsilon}(\Sigma^{-1} + \mathrm{Id})^{-1} x$$

Step 2 : Construct a variable which behaves like the samples after one step of mean aggregation

Lemma 1. *With probability at least* $1-\rho$, *for all* $i = 1, \ldots, n$:

$$\left.\begin{array}{r} \left\| x_i^{(1)} - \varphi_{\text{reg.}}(x_i) \right\|_{\Sigma^{-1}} \\ \left\| \Sigma^{-\frac{1}{2}} \left(x_i^{(1)} (x_i^{(1)})^\top - \varphi_{\text{reg.}}(x_i)\varphi_{\text{reg.}}(x_i)^\top \right) \Sigma^{-\frac{1}{2}} \right\| \end{array}\right\} \lesssim \frac{C \log n (\sqrt{d + \log(1/\rho)})}{\sqrt{n}}$$

where $C = \mathrm{poly}(\varepsilon^{-1}, \|\Sigma\|, |\mathrm{Id}+\Sigma|)$.

With high probability, the constructed variable behaves like the samples after one step of smoothing within some error term

UNIVERSITY OF WATERLOO

Finite Smoothing: Linear Regression

Difference between the estimated risk and the true risk:

$$\mathcal{R}^{(0)} = R_{\text{reg.}}(\Sigma) + \mathcal{O}\left(\frac{\|\Sigma\| \, \|\beta^\star\|^2 \, d\sqrt{\log(1/\rho)}}{(\lambda + \lambda_{\min})\sqrt{n}}\right)$$

Error term can be ignored when n is sufficiently large

$$\mathcal{R}^{(1)} = R_{\text{reg.}}(\Sigma^{(1)}) + \mathcal{O}\left(C\varepsilon^{1/5}\right) + \mathcal{O}\left(\frac{C' \log n \sqrt{d + \log(1/\rho)}}{(\lambda + \lambda_{\min})\sqrt{n}}\right)$$

Error terms can be ignored when ε is sufficiently small, n is sufficiently large

Recall that we have the assumption: $R_{\text{reg.}}(\Sigma) > R_{\text{reg.}}((\mathrm{Id} + \Sigma^{-1})^{-2}\Sigma)$

Therefore, $\mathcal{R}^{(1)} < \mathcal{R}^{(0)}$

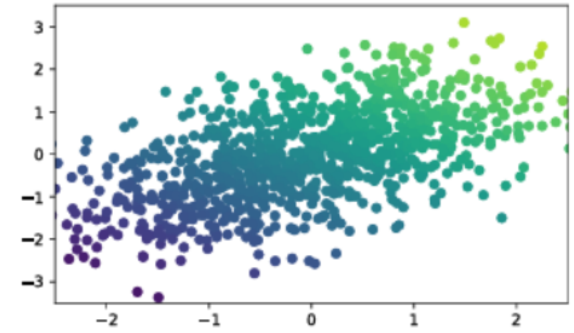
Smoothing Shrinks The Directions of The Small Eigenvalues Faster

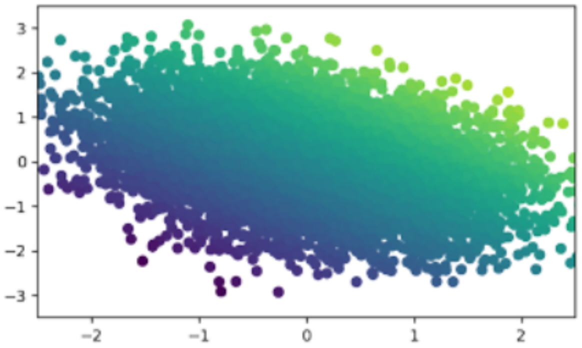
$x^{(k)}$ behaves like $(\mathrm{Id} + \Sigma^{-1})^{-k} x$

Eigenvalues becomes $\lambda_i^{(k)} = (1 + 1/\lambda_i)^{-2k} \lambda_i.$

As a result, when $\lambda_i \gg 1$ $\lambda_i^{(1)} \sim \lambda_i$

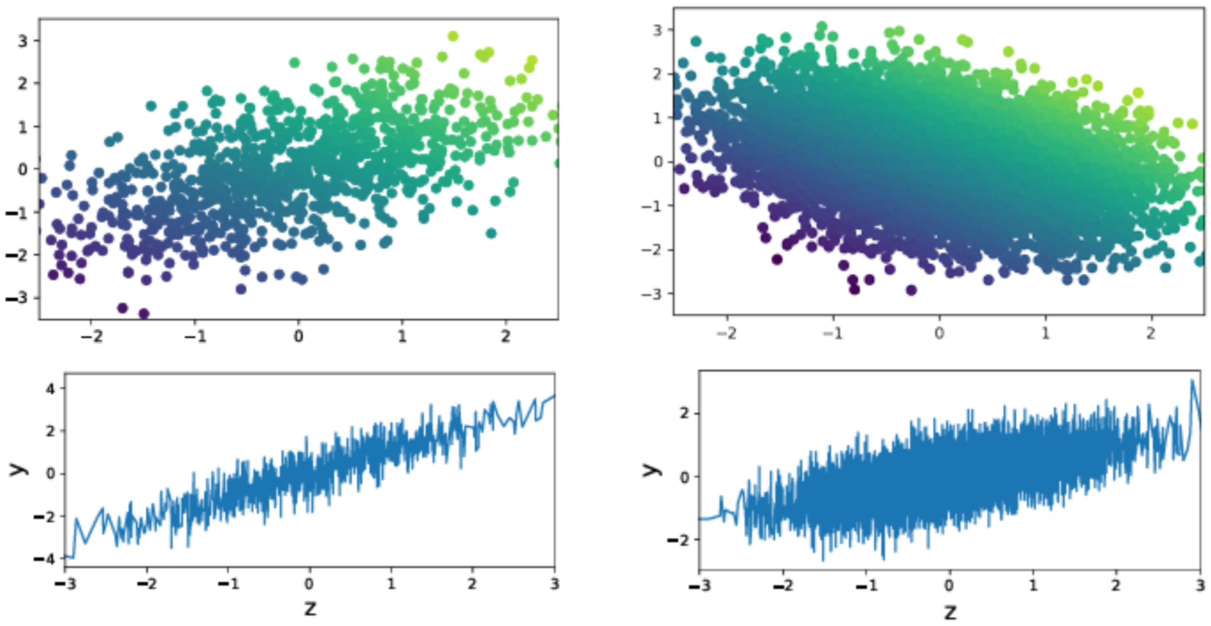
when $\lambda_i \ll 1$ $\lambda_i^{(1)} \sim \lambda_i^{2k+1}$





Examples when d = 2

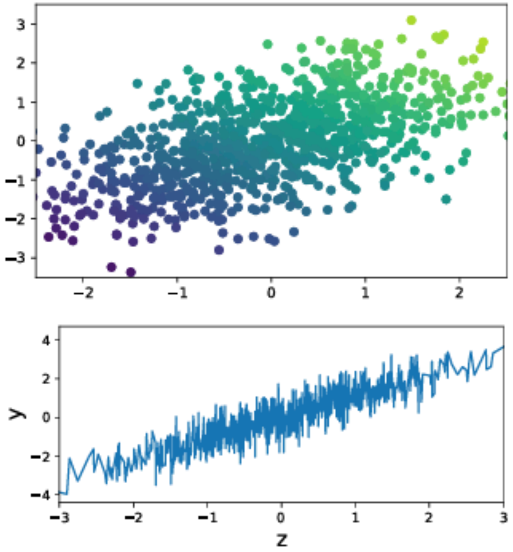
- d = 2, p = 1
- Σ has two eigenvalues λ_1 and λ_2
- $\beta^\star = bu_1$ (aligned with first eigenvectors, so u_1 is the useful information and u_2 is noise)
- M = [1,0] (projection on the first coordinate)



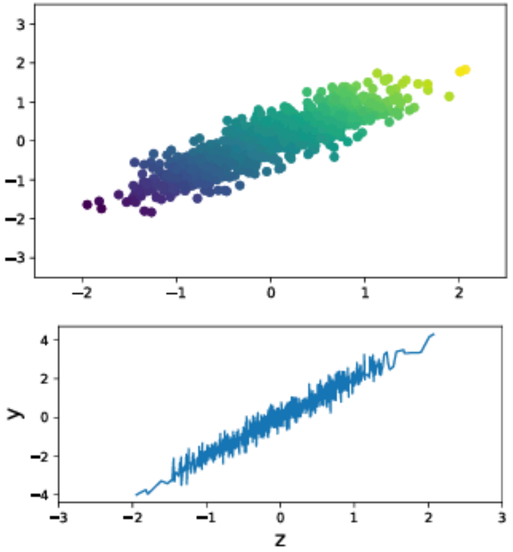
UNIVERSITY OF WATERLOO

Smoothing Improves Performance

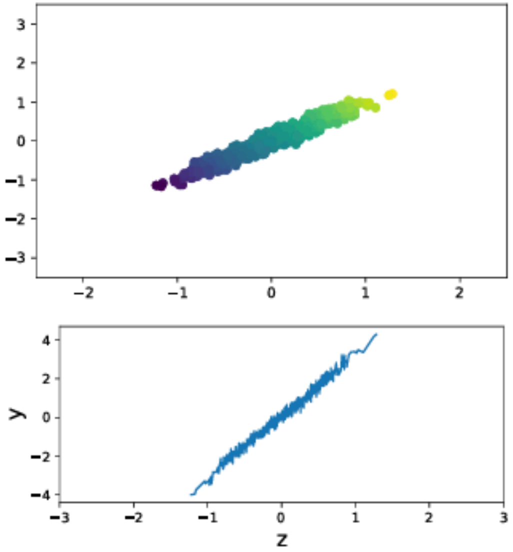
K = 0

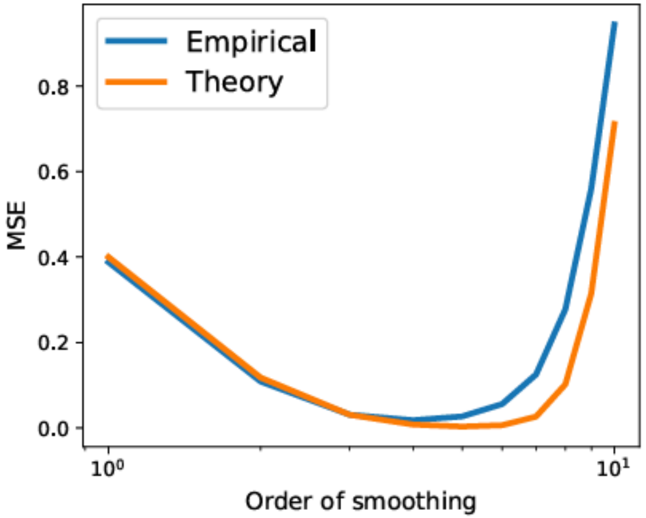


K = 1



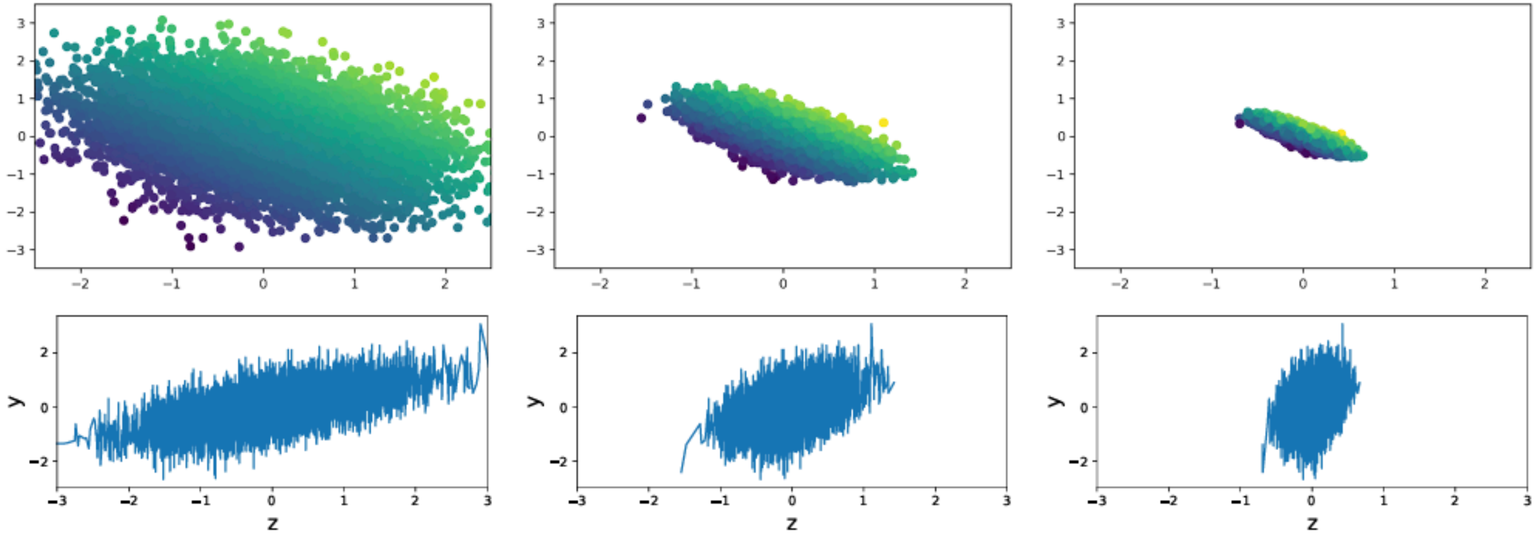
K = 2

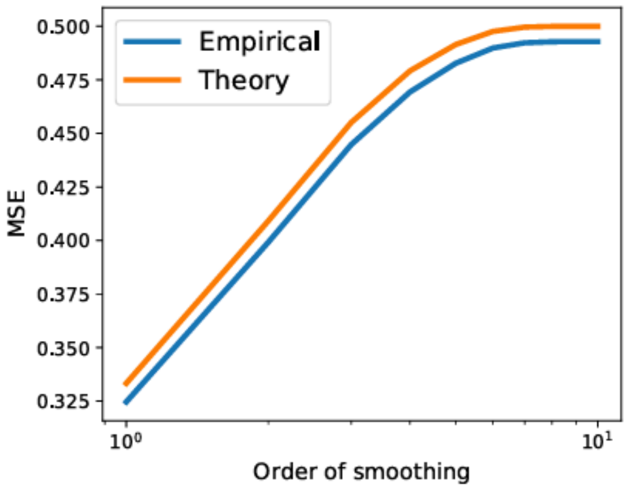




Smoothing Does Not Improve Performance

K = 0 K = 1 K = 2



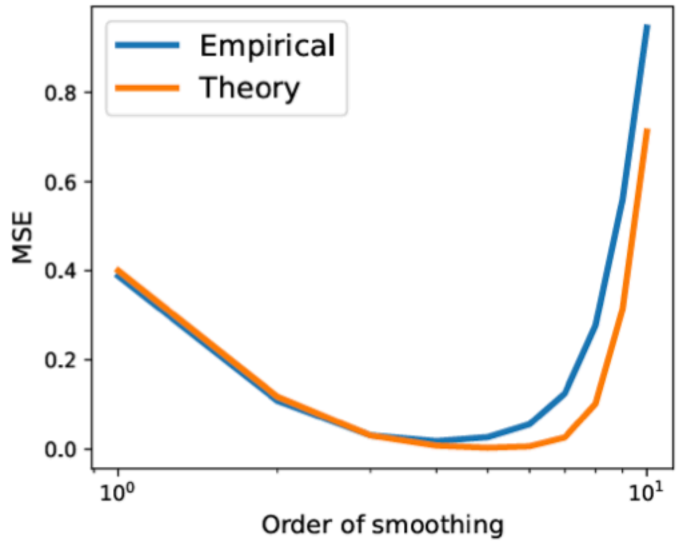


What we hope is a thin straight line after smoothing, but in this case, the situation is even worse after smoothing!

Explicit Risk Expression in terms of Eigenvalues

- d = 2, p = 1
- Σ has two eigenvalues λ1 >> 1 and λ2 << 1
- Eigenvectors: $u_1 = [1,1]/\sqrt{2}$ and $u_2 = [-1,1]/\sqrt{2}$
- $\beta^\star = b u_1$
- M = [1,0] (projection on the first coordinate)

$$\mathcal{R}^{(k)} \approx R_{\text{reg.}}(\Sigma^{(k)}) = \lambda_1 b^2 \frac{(2\lambda + \lambda_2^{(k)})^2 + \lambda_2^{(k)}\lambda_1^{(k)}}{(2\lambda + \lambda_1^{(k)} + \lambda_2^{(k)})^2}$$



UNIVERSITY OF WATERLOO

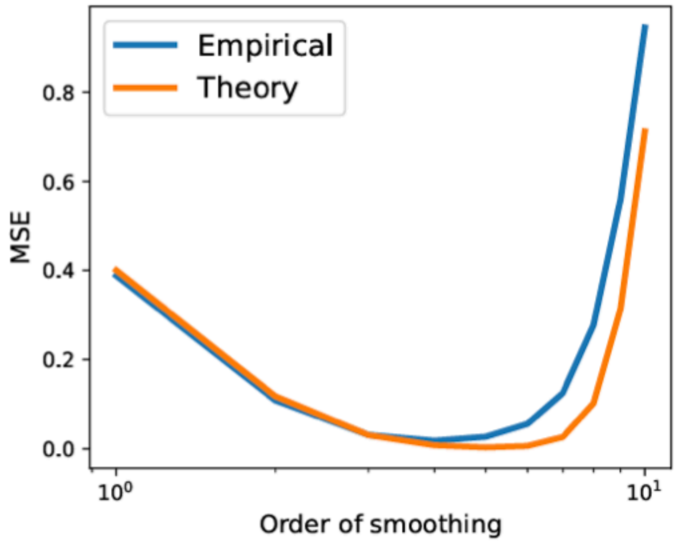
Explicit Risk Expression in terms of Eigenvalues

$$\mathcal{R}^{(k)} \approx R_{\text{reg.}}(\Sigma^{(k)}) = \lambda_1 b^2 \frac{(2\lambda + \lambda_2^{(k)})^2 + \lambda_2^{(k)}\lambda_1^{(k)}}{(2\lambda + \lambda_1^{(k)} + \lambda_2^{(k)})^2}$$

- When λ2 << 1 << λ1, λ2 decreases much faster
- The risk will first decreases to a minimum

$$\lambda_1 b^2 \left(\frac{2\lambda}{2\lambda+\lambda_1^{(k^\star)}}\right)^2$$

λ2 is close to 0



- Then, it will increase to $\lambda_1 b^2 = \|\beta^\star\|_\Sigma^2 = \lim_{n\to\infty} \mathcal{R}^{(\infty)}$

Finite Smoothing: Classification

Latent variables and labels $(x, y) \sim (1/2)(\mathcal{N}_\mu \otimes \{1\} + \mathcal{N}_{-\mu} \otimes \{-1\})$

Two balanced classes Gaussian distribution with identity covariance

In this case z_i are also Gaussian, with mean $\nu \stackrel{\text{def.}}{=} M^\top \mu$ or $-\nu$

The loss function is still MSE, although it is not the best method for classification

UNIVERSITY OF WATERLOO

Finite Smoothing: Classification

$$d_\mu(x) \overset{\text{def.}}{=} 2^{-d/2} e^{-\frac{\|x-\mu\|^2}{4}}$$

$$\varphi_{\text{cl.}}(x) = \frac{d_\mu(x)\left(\frac{x+\mu}{2}\right) + d_{-\mu}(x)\left(\frac{x-\mu}{2}\right)}{2\varepsilon + d_\mu(x) + d_{-\mu}(x)}$$

Step 1 : Construct a variable which behaves like the samples after one step of mean aggregation

Lemma 2. *With probability at least* $1-\rho$,

$$\left.\begin{array}{r} \sup\limits_{i=1,\dots,n} \left\| x_i^{(1)} - \varphi_{\text{cl.}}(x_i) \right\| \\ \sup\limits_{i=1,\dots,n} \left\| x_i^{(1)} (x_i^{(1)})^\top - \varphi_{\text{cl.}}(x_i)\varphi_{\text{cl.}}(x_i)^\top \right\| \end{array}\right\} \lesssim \frac{\text{poly}(\varepsilon^{-1}) \log n(\sqrt{d} + \sqrt{\log(1/\rho)})}{\sqrt{n}}$$

With high probability, the constructed variable behaves like the samples after one step of smoothing within some error term

UNIVERSITY OF WATERLOO

Finite Smoothing: Classification

$$R_{\text{cl.}}(s) = \frac{(s+\lambda)^2 + s\,\|\nu\|^2}{(s+\lambda+\|\nu\|^2)^2}$$

Step 2: Get an estimation of risk in terms of variance and mean

$$\mathcal{R}^{(0)} = R_{\text{cl.}}(1) + \mathcal{O}\left(\frac{\|\nu\|^4\, p\sqrt{\log(1/\rho)}}{\sqrt{n}}\right)$$

Error term can be ignored when n is sufficiently large

$$\mathcal{R}^{(1)} = R_{\text{cl.}}(1/4) + \mathcal{O}\left(C\left(\varepsilon^{\frac{1}{4}} + \frac{1}{\varepsilon^3} e^{-\frac{\|\mu\|^2}{4}}\right)\right) + \mathcal{O}\left(\frac{C'(\log n)(\sqrt{d+\log(1/\rho)})}{\sqrt{n}}\right)$$

The second error term is due to communities getting closer to each other

Error terms can be ignored when ε is sufficiently small, n is sufficiently large, and μ is sufficiently large

Finite Smoothing: Classification

$$R_{\text{cl.}}(s) = \frac{(s+\lambda)^2 + s\left\|\nu\right\|^2}{(s+\lambda+\left\|\nu\right\|^2)^2}$$

An estimation of risk in terms of variance and mean

$$\frac{dR}{ds} = \frac{(s+3\lambda)\cdot\|\nu\|^2 + \|\nu\|^4}{(s+\lambda+\|\nu\|^2)^3} > 0 \Rightarrow R \text{ increases as } s \text{ increases.}$$

$$\frac{dR}{d\|\nu\|} = -\frac{2\lambda(\lambda+s)\cdot\|\nu\|}{(s+\lambda+\|\nu\|^2)^3} < 0 \Rightarrow R \text{ decreases as } \|\nu\| \text{ increases.}$$

How About When K becomes larger?

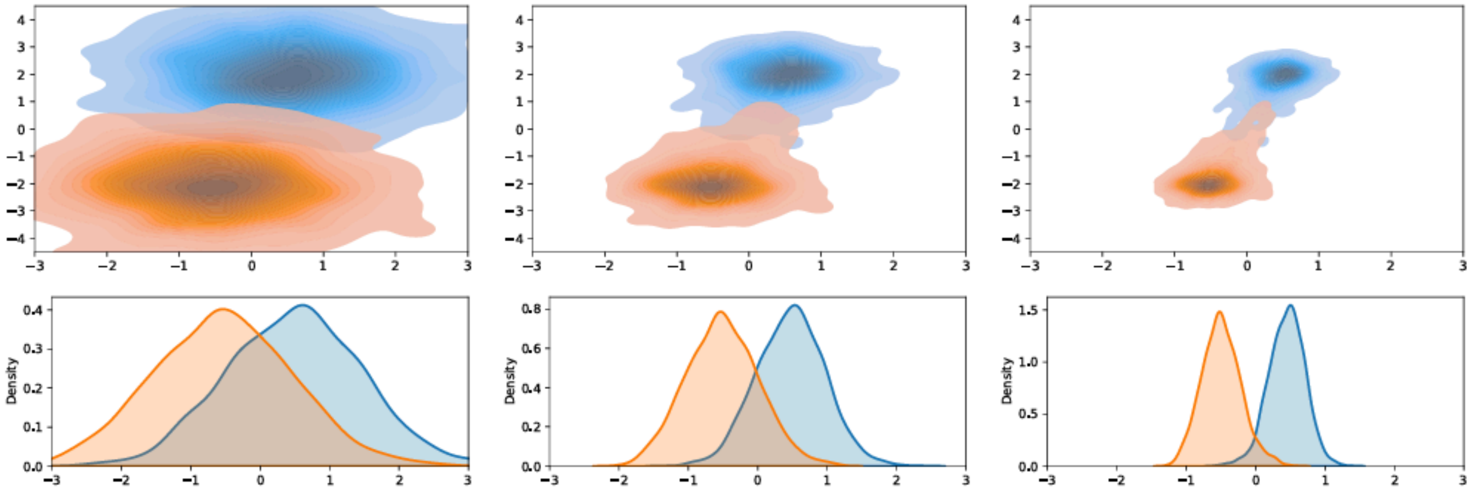
$$\varphi_{\text{cl.}}(x) = \frac{d_\mu(x)\left(\frac{x+\mu}{2}\right) + d_{-\mu}(x)\left(\frac{x-\mu}{2}\right)}{2\varepsilon + d_\mu(x) + d_{-\mu}(x)}$$

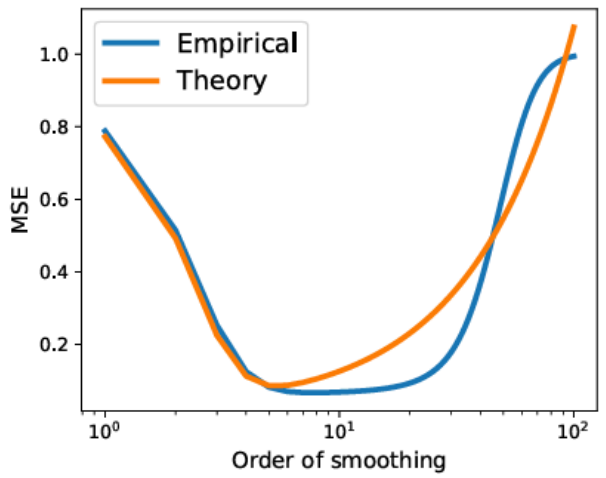
After one step of mean aggregation, the mean for one community does not change while the variance decreases by a quarter

$$\mathcal{R}^{(k)} \approx R_{\text{cl.}}(4^{-k}) + \mathcal{O}\left(\sum_{\ell=0}^{k-1} e^{-\frac{\|\mu\|^2}{2(1+4^{-\ell})}}\right)$$

When k increases, the first term decreases, but error terms become dominant, however, the author mentioned that the error term was not accurate enough

Finite Smoothing: Classification





UNIVERSITY OF WATERLOO

Conclusion

- A limited number rounds of mean aggregation can improve the performance
- The label should align with the large principal directions (in reality, we usually assume this is true?) so that smoothing can improve performance
- Mean aggregation tends to shrink noisy principal components (the ones with smaller eigenvalues) faster than meaningful ones
- Mean aggregation tends to shrink communities faster than they collapse together

Discussion

- The theoretical analysis aims to find the relationship between risk and the variance of the samples (mean aggregation can reduce variance)
- This paper illustrates the underlying logic of mean aggregation: shrinking the eigenvalues (with different rates depending on the values)
- A good approximation for linear regression case
- More work need to be done for classification case
- Only analyze the risk after one step of smoothing

Future Works

- Extend the theory to other models (rather than linear GNNs) and other loss functions
- Extend the theory to other types of aggregations
- Get an explicit risk expression in terms of the rounds of smoothing

THANK YOU!