Graph Denoising for Molecular Imaging

Cédric Mendelin <cedric.mendelin@stud.unibas.ch>

Department of Mathematics and Computer Science, University of Basel

29.06.2022

Outline

Outline

Cryo-Electron Microscopy (Cryo-EM)

- Major motivation
- Enables observation of molecules in near atomic resolution

Cryo-Electron Microscopy (Cryo-EM)

- Major motivation
- Enables observation of molecules in near atomic resolution
- Observation through an electron microscope
- Frozen state of molecules required for observation
  - Frozen molecules are fragile $\mapsto$ electron microscope low power
  - During freezing, molecules rotate randomly
- Observations can be reconstructed to a 3D model

Cryo-Electron Microscopy (Cryo-EM)

- Major motivation
- Enables observation of molecules in near atomic resolution
- Observation through an electron microscope
- Frozen state of molecules required for observation
  - Frozen molecules are fragile $\mapsto$ electron microscope low power
  - During freezing, molecules rotate randomly
- Observations can be reconstructed to a 3D model

Only single particle cryo-EM is considered.

Cryo-Electron Microscopy (Cryo-EM) - Illustration

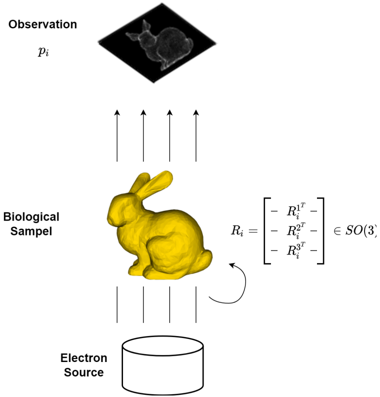


Figure: Cryo-EM overview

Cryo-Electron Microscopy (Cryo-EM) - Illustration

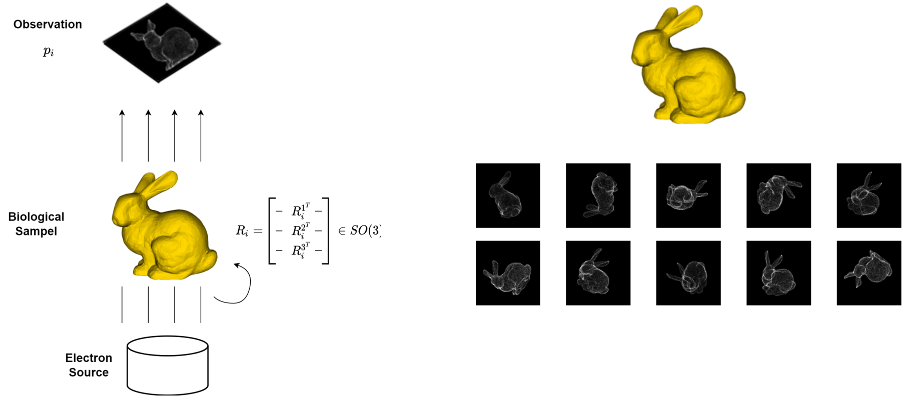
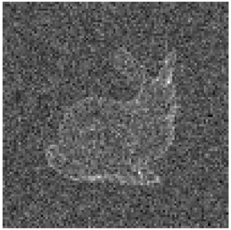


Figure: Cryo-EM overview

Cryo-Electron Microscopy (Cryo-EM) - Challenges

- High level of noise
- Unknown rotation during freezing
- (Structural variety of observations)

(a) Clean micrograph



(b) Noisy micrograph

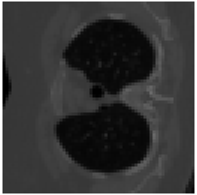
Computed Tomography (CT)

Computed Tomography (CT)

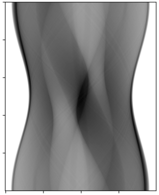
- Similar to Cryo-EM
- Can be seen as a simpler version in 2D with known observation angles
- Good to start with towards a Cryo-EM algorithm

Computed Tomography (CT)

- Similar to Cryo-EM
- Can be seen as a simpler version in 2D with known observation angles
- Good to start with towards a Cryo-EM algorithm



(a) Biological sample



(b) Clean observation (sinogram)

Shared Observation Model

Shared Observation Model

Observation

$$y = p + \eta \tag{1}$$

- y: noisy observation
- p: noiseless observation
- η: noise, assumed $\eta_i \sim \mathcal{N}(0, \sigma^2)$

Shared Observation Model

Observation

$$\begin{aligned} y &= p + \eta \\ y_i[j] &= p_i[j] + \eta_i[j] \quad \text{with } 1 \leq i \leq N, 1 \leq j \leq M \end{aligned} \tag{1}$$

- y: noisy observation
- p: noiseless observation
- η: noise, assumed $\eta_i \sim \mathcal{N}(0, \sigma^2)$
- N: number of observations
- M: observation dimension

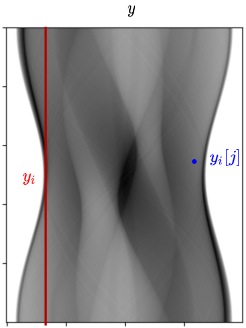
Shared Observation Model

Observation

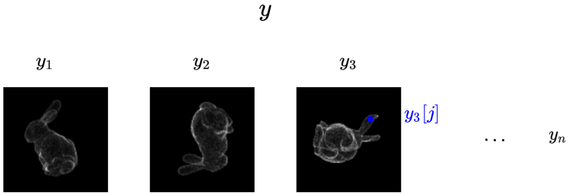
$$\begin{aligned} y &= p + \eta \\ y_i[j] &= p_i[j] + \eta_i[j] \qquad \text{with } 1 \leq i \leq N, 1 \leq j \leq M \\ y_i &= A(x, \theta_i) + \eta_i \end{aligned} \tag{1}$$

- y: noisy observation
- p: noiseless observation
- η: noise, assumed $\eta_i \sim \mathcal{N}(0, \sigma^2)$
- x: biological sample
- N: number of observations
- M: observation dimension
- $A : x \mapsto A(x; \theta_i) \in \mathbb{R}^M$: a non-linear operator
- θ_i: observation angle

Observation - Illustration



(a) CT Observation - sinogram



(b) Cryo-EM Observation - micrographs

Reconstruction

Reconstruction

Reconstruction

$$Recon : \mathbb{R}^{M \times N} \to \mathbb{R}^{M \times M} \quad y \mapsto Recon(y; \theta) \tag{2}$$

Signal-to-noise-ration (SNR)

Signal-to-noise-ration (SNR)

- SNR is a measure, which compares the power of an input signal to the power of the undesired noise
- Typically given in decibel (dB)
- SNR ≤ 0 dB indicated more noise than signal.

Signal-to-noise-ration (SNR)

- SNR is a measure, which compares the power of an input signal to the power of the undesired noise
- Typically given in decibel (dB)
- SNR ≤ 0 dB indicated more noise than signal.

SNR_y is used to define the level of noise in an observation.

Signal-to-noise-ration (SNR)

- SNR is a measure, which compares the power of an input signal to the power of the undesired noise
- Typically given in decibel (dB)
- SNR ≤ 0 dB indicated more noise than signal.

SNR_y is used to define the level of noise in an observation.

SNR is used as a metric for the quality of reconstructions.

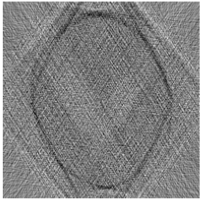
Reconstruction - Computed Tomography

- Filter Backprojection (FBP)
  - Can be considered historical approach
  - Enables reconstruction for moderate noise

Reconstruction - Computed Tomography

- Filter Backprojection (FBP)
  - Can be considered historical approach
  - Enables reconstruction for moderate noise

(a) Reconstruction clean: $Recon(p, \theta) \approx x$

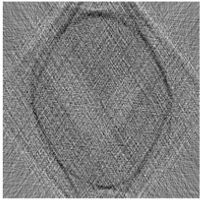


(b) Reconstruction noisy with SNR_y 0 dB: $Recon(y, \theta) \not\approx x$

Reconstruction - Computed Tomography

- Filter Backprojection (FBP)
  - Can be considered historical approach
  - Enables reconstruction for moderate noise
- Neural Network Approaches
  - Today state-of-the art
  - Using result of FBP and further denoise
  - U-Net Ronneberger, Fischer, and Brox 2015

(a) Reconstruction clean: $Recon(p, \theta) \approx x$



(b) Reconstruction noisy with SNR_y 0 dB: $Recon(y, \theta) \not\approx x$

Problem and Goal

Problem and Goal

Problem

p not observable directly only access to y.

Problem and Goal

Problem

p not observable directly only access to y.

Goal

$$denoiser : y_i = (p_i + \eta) \mapsto p_i^* \approx p_i$$

$$Recon\,(denoiser(y;\theta)) \approx x$$

Outline

Graph - Definitions

Graph - Definitions

Graph Definition

A graph is defined as $G = \langle V, E \rangle$, where V is a set of nodes and E is a set of edges.

Graph - Definitions

Graph Definition

A graph is defined as $G = \langle V, E \rangle$, where V is a set of nodes and E is a set of edges.

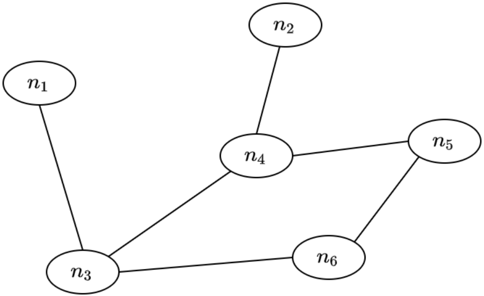


Figure: Sample graph

Graph - Definitions

Graph Definition

A graph is defined as $G = \langle V, E \rangle$, where V is a set of nodes and E is a set of edges.

Nodes

$(n_1, n_2, \dots) \in \mathbb{R}^F$, with F as node feature dimensions

Edges

Edges are defined as a set of tuples (i, j), where i and j determine the index of the nodes.

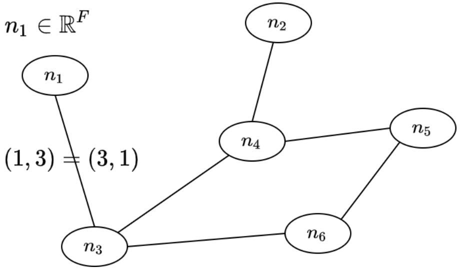


Figure: Sample graph

Graph - Definitions - Adjacency Matrix

Graph - Definitions - Adjacency Matrix

Adjacency Matrix

The binary adjacency matrix of graph $G = \langle V, E \rangle$ is defined as:

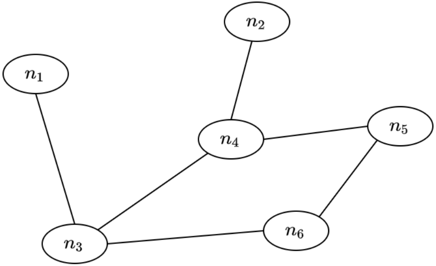
$$A_{ij} = \begin{cases} 1 & \text{if } (i, j) \in E \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Graph - Definitions - Adjacency Matrix

Adjacency Matrix

The binary adjacency matrix of graph $G = \langle V, E \rangle$ is defined as:

$$A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0, & \text{otherwise} \end{cases} \tag{3}$$



$$A = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

Figure: Sample graph

Graph - Definitions - Degree

Graph - Definitions - Degree

Degree of a node

The *degree* of a node is defined as the number of (incoming) edges.

Graph - Definitions - Degree

Degree of a node

The *degree* of a node is defined as the number of (incoming) edges.

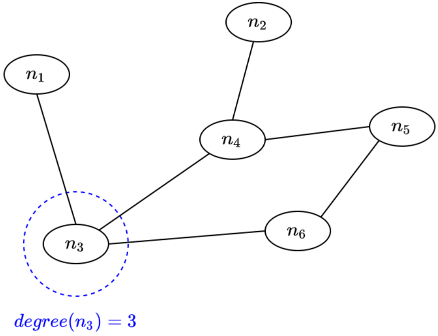


Figure: Sample graph

Graph - Definitions - Degree

Degree of a node

The *degree* of a node is defined as the number of (incoming) edges.

Degree Matrix of Graph G

Is a diagonal matrix with degree of nodes as entries.

$$D_{ii} = degree(n_i)$$

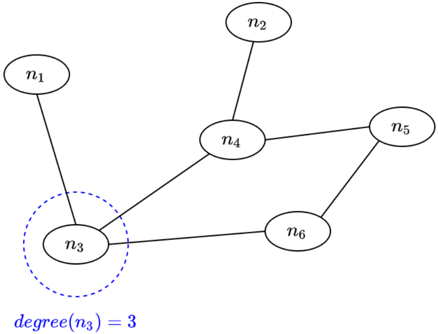


Figure: Sample graph

Graph for Molecular Imaging Observation

How to construct a graph for molecular imaging?

Graph for Molecular Imaging Observation

How to construct a graph for molecular imaging?

- Nodes: Single observation y_i

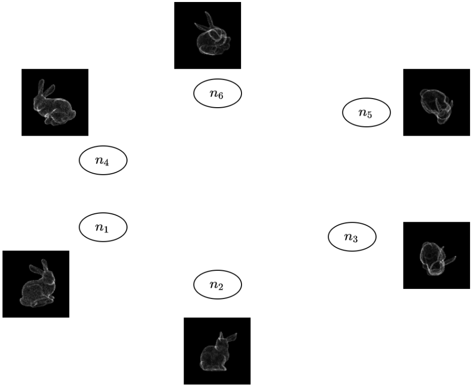


Figure: Sample graph for cryo-EM observation

Graph for Molecular Imaging Observation

How to construct a graph for molecular imaging?

- Nodes: Single observation y_i
- Edges: Use k-nearest neighbours (k-NN) to construct a graph

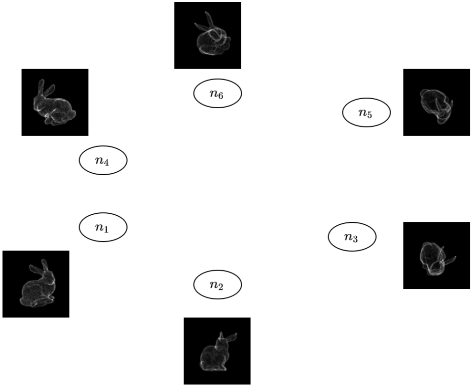


Figure: Sample graph for cryo-EM observation

Graph for Molecular Imaging Observation

How to construct a graph for molecular imaging?

- Nodes: Single observation y_i
- Edges: Use k-nearest neighbours (k-NN) to construct a graph
  - Define similarity measure for nodes: $\ell 2$-norm

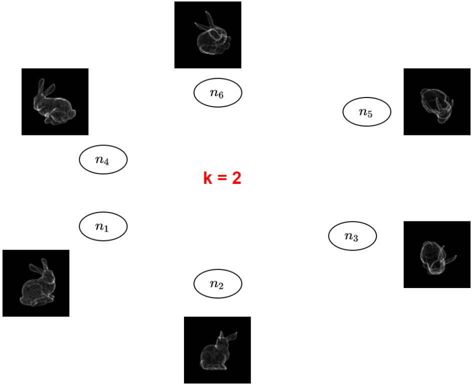


Figure: Sample graph for cryo-EM observation

Graph for Molecular Imaging Observation

How to construct a graph for molecular imaging?

- Nodes: Single observation y_i
- Edges: Use k-nearest neighbours (k-NN) to construct a graph
  - Define similarity measure for nodes: $\ell 2$-norm

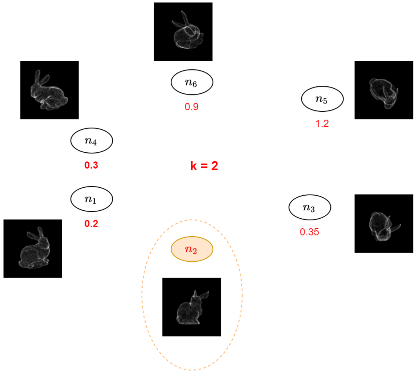


Figure: Sample graph for cryo-EM observation

Graph for Molecular Imaging Observation

How to construct a graph for molecular imaging?

- Nodes: Single observation y_i
- Edges: Use k-nearest neighbours (k-NN) to construct a graph
  - Define similarity measure for nodes: ℓ2-norm

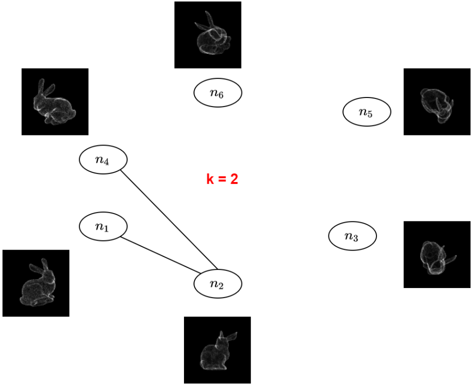


Figure: Sample graph for cryo-EM observation

Graph for Molecular Imaging Observation

How to construct a graph for molecular imaging?

- Nodes: Single observation y_i
- Edges: Use k-nearest neighbours (k-NN) to construct a graph
  - Define similarity measure for nodes: $\ell 2$-norm

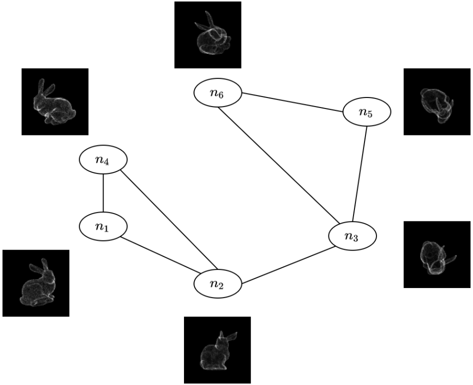


Figure: Sample graph for cryo-EM observation

Graph for Molecular Imaging Observation - Noise

What happens with our noisy observations?

Graph for Molecular Imaging Observation - Noise

What happens with our noisy observations?

- With noise, graph will capture neighborhood inaccurately.

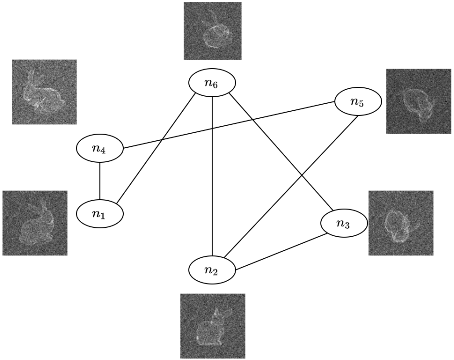


Figure: Sample graph for noisy cryo-EM observation

Graph Laplacian (GL)

What can we use this graph for?

Graph Laplacian (GL)

What can we use this graph for?

- Coifman et al. 2008 used it to approximate angles for CT:

Graph Laplacian (GL)

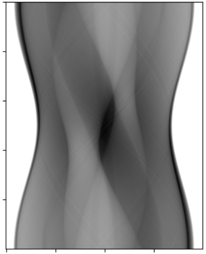
What can we use this graph for?

- Coifman et al. 2008 used it to approximate angles for CT:

Low-dimensional Embedding

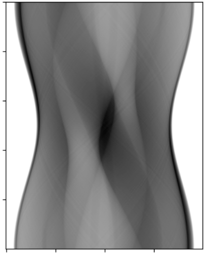
1. Construct a k-NN graph from observations.
2. Calculate $L = D - A$
3. Get 2nd and 3rd smallest eigenvalue with corresponding eigenvectors.

Low-dimensional Embedding for Computed Tomography

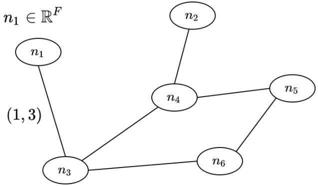


(a) Clean CT observation

Low-dimensional Embedding for Computed Tomography

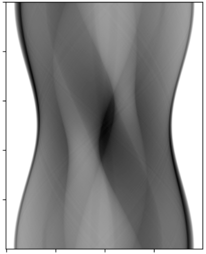


(a) Clean CT observation

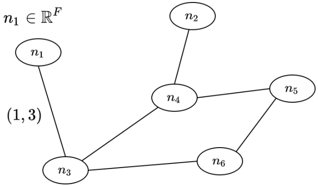


(b) Building k-NN graph with $k = 2$

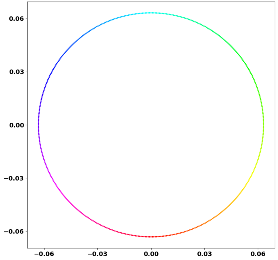
Low-dimensional Embedding for Computed Tomography



(a) Clean CT observation



(b) Building k-NN graph with $k = 2$



(c) 2_{nd} and 3_{rd} smallest eigenvectors of $L = D - A$

Computed Tomography with unknown angles

Computed Tomography with unknown angles

- GL-Embedding estimates observation angles

Computed Tomography with unknown angles

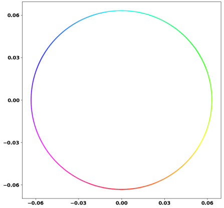
- GL-Embedding estimates observation angles

(a) Reconstruction known angles

Computed Tomography with unknown angles

- GL-Embedding estimates observation angles

(a) Reconstruction known angles

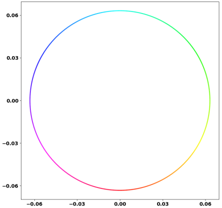


(b) GL-Embedding from $k = 2$

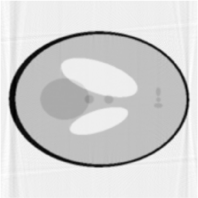
Computed Tomography with unknown angles

- GL-Embedding estimates observation angles

(a) Reconstruction known angles



(b) GL-Embedding from $k = 2$



(c) Reconstruction unknown angles

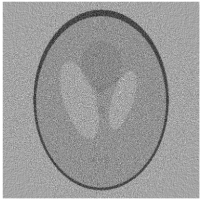
Computed Tomography with unknown angles

- GL-Embedding estimates observation angles

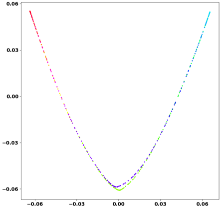
What happens in the noisy case?

Computed Tomography with unknown angles

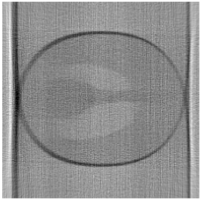
- GL-Embedding estimates observation angles



(a) Reconstruction known angles SNR_y : 10 dB



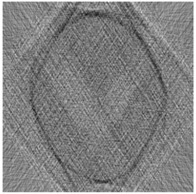
(b) GL-Embedding from $k = 6$ and SNR_y : 10 dB



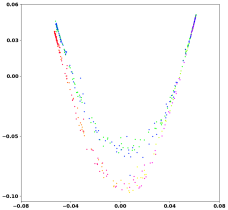
(c) Reconstruction unknown angles SNR_y : 10 dB

Computed Tomography with unknown angles

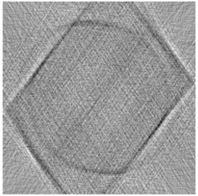
- GL-Embedding estimates observation angles



(a) Reconstruction known angles SNR_y : 0 dB



(b) GL-Embedding from $k = 6$ and SNR_y : 0 dB



(c) Reconstruction unknown angles SNR_y : 0 dB

Observation Denoising

The fewer noise is available in the observation, the better reconstruction is possible.

Observation Denoising

The fewer noise is available in the observation, the better reconstruction is possible.

- Use existing denoising algorithms
  - Block-matching and 3D filtering (BM3D) Dabov et al. 2007
  - Non-local means Buades, Coll, and Morel 2005
  - No graph as data structure
  - But, both exploit neighborhood during averaging

Observation Denoising

The fewer noise is available in the observation, the better reconstruction is possible.

- Use existing denoising algorithms
  - Block-matching and 3D filtering (BM3D) Dabov et al. 2007
  - Non-local means Buades, Coll, and Morel 2005
  - No graph as data structure
  - But, both exploit neighborhood during averaging
- Show potential for graph as a data structure

Observation Denoising

The fewer noise is available in the observation, the better reconstruction is possible.

- Use existing denoising algorithms
  - Block-matching and 3D filtering (BM3D) Dabov et al. 2007
  - Non-local means Buades, Coll, and Morel 2005
  - No graph as data structure
  - But, both exploit neighborhood during averaging
- Show potential for graph as a data structure

Exploit graph as a data structure and the GL-Embedding

Outline

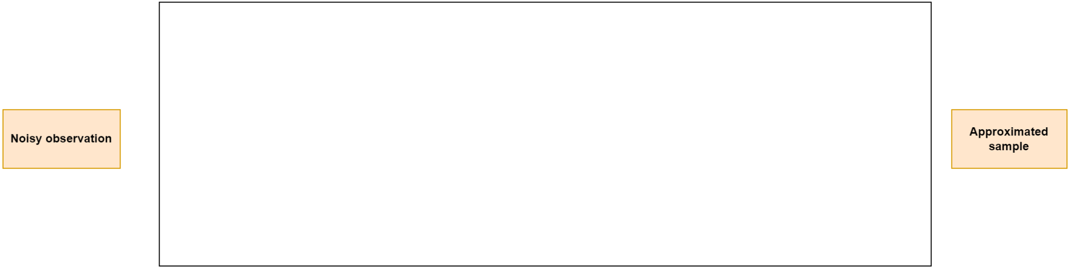
GAT-Denoiser

- GAT-Denoiser is a graph neural network (GNN) to denoise observations.
- Consists of three components:
  - Convolution
  - Graph Attention Network (GAT) Veličković et al. 2017
  - End-to-End Learning

GAT-Denoiser

- GAT-Denoiser is a graph neural network (GNN) to denoise observations.
- Consists of three components:
  - Convolution
  - Graph Attention Network (GAT) Veličković et al. 2017
  - End-to-End Learning

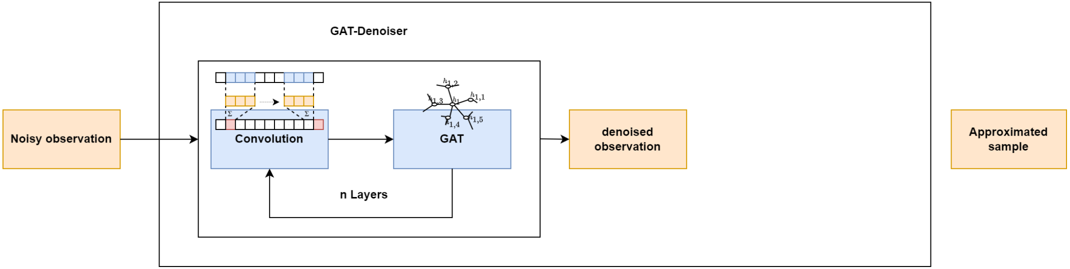
GAT-Denoiser Pipeline



GAT-Denoiser

- GAT-Denoiser is a graph neural network (GNN) to denoise observations.
- Consists of three components:
  - Convolution
  - Graph Attention Network (GAT) Veličković et al. 2017
  - End-to-End Learning

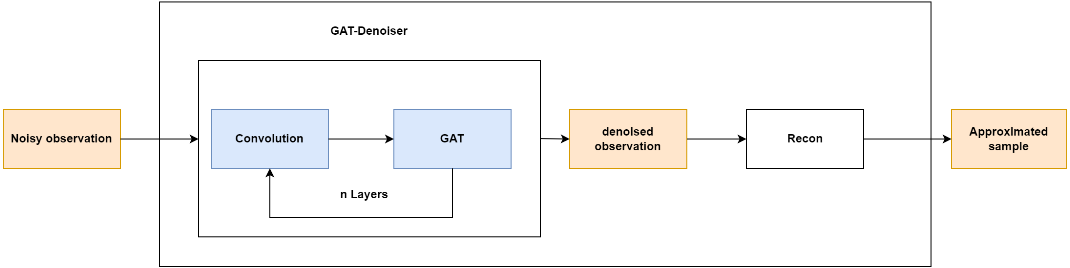
GAT-Denoiser Pipeline



GAT-Denoiser

- GAT-Denoiser is a graph neural network (GNN) to denoise observations.
- Consists of three components:
  - Convolution
  - Graph Attention Network (GAT) Veličković et al. 2017
  - End-to-End Learning

GAT-Denoiser Pipeline



GAT-Denoiser Components

GAT-Denoiser Components

- Convolution
  - Denoise single observation

GAT-Denoiser Components

- Convolution
  - Denoise single observation
- Graph Attention Network (GAT) Veličković et al. 2017
  - Denoise neighboring observation

GAT-Denoiser Components

- Convolution
  - Denoise single observation
- Graph Attention Network (GAT) Veličković et al. 2017
  - Denoise neighboring observation
- End-to-End Learning
  - Optimize for reconstruction quality
  - Loss:

GAT-Denoiser Components

- Convolution
  - Denoise single observation
- Graph Attention Network (GAT) Veličković et al. 2017
  - Denoise neighboring observation
- End-to-End Learning
  - Optimize for reconstruction quality
  - Loss:
    - $\mathcal{L}_{observation} = \| p - GAT\text{-}Denoiser(y) \|_2^2$

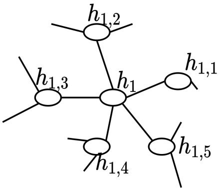
GAT-Denoiser Components

- Convolution
  - Denoise single observation
- Graph Attention Network (GAT) Veličković et al. 2017
  - Denoise neighboring observation
- End-to-End Learning
  - Optimize for reconstruction quality
  - Loss:
    - $\mathcal{L}_{observation} = \| p - GAT\text{-}Denoiser(y) \|_2^2$
    - $\mathcal{L}_{reconstruction} = \| x - Recon(GAT\text{-}Denoiser(y)) \|_2^2$

Graph Attention Network - GAT

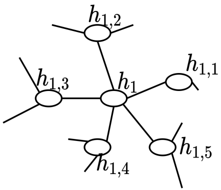
Graph Attention Network - GAT

- Extends Graph Convolution Network with attention (weights)
- Computes new node features
- Averages graph over neighborhood
- Multi-head available, motivated by Vaswani et al. 2017



Graph Attention Network - GAT

- Extends Graph Convolution Network with attention (weights)
- Computes new node features
- Averages graph over neighborhood
- Multi-head available, motivated by Vaswani et al. 2017
- σ: activation function (Exponential Linear Unit)
- W: learnable weight matrix
- α: normalized attention coefficients



$$h'_1 = \sigma\left(\sum_{i=1}^{5} \alpha_i W h_{1,i}\right)$$

Input Graph

Input Graph

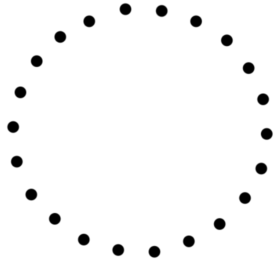
- Exploit information from GL
  - Low-dimensional embedding estimates angles
  - Dominant information in data can be considered observation angles.

Input Graph

- Exploit information from GL
  - Low-dimensional embedding estimates angles
  - Dominant information in data can be considered observation angles.
- Construct graph from observation angles

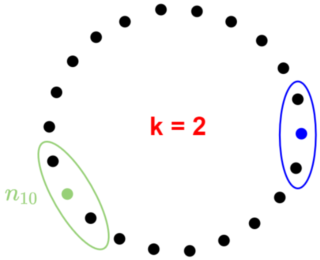
Input Graph

- Exploit information from GL
  - Low-dimensional embedding estimates angles
  - Dominant information in data can be considered observation angles.
- Construct graph from observation angles
  - Map angles to unit-Circle or unit-sphere



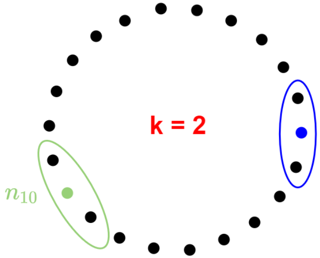
Input Graph

- Exploit information from GL
  - Low-dimensional embedding estimates angles
  - Dominant information in data can be considered observation angles.
- Construct graph from observation angles
  - Map angles to unit-Circle or unit-sphere
  - Use k-NN with great-circle distance



Input Graph

- Exploit information from GL
  - Low-dimensional embedding estimates angles
  - Dominant information in data can be considered observation angles.
- Construct graph from observation angles
  - Map angles to unit-Circle or unit-sphere
  - Use k-NN with great-circle distance



Observation angles θ are assumed to be equally spaced.

GAT-Denoiser Implementation for Computed Tomography

- Use U-Net for reconstruction
- During Trainig, U-Net might be trained jointly

Outline

LoDoPaB-CT dataset

- Dataset for low-dose Computed Tomography
- 35'820 train samples
- 3'553 test samples
- BM3D as baseline algorithm
- Resolution 64 x 64

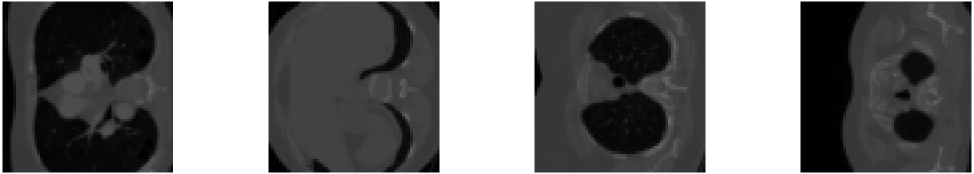


Figure: Some samples from the LoDoPaB-CT dataset.

Evaluation

- Small Scale Experiments
  - 1024 train samples
  - 100 test samples
  - 200 epochs
- Large Scale Experiments
  - Complete LoDoPaB-CT dataset
  - 20 - 40 epochs

Evaluation

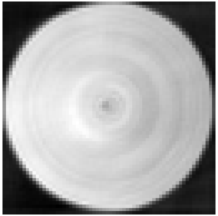
- Small Scale Experiments
  - 1024 train samples
  - 100 test samples
  - 200 epochs
  - Goal: Find most promising architecture
- Large Scale Experiments
  - Complete LoDoPaB-CT dataset
  - 20 - 40 epochs
  - Goal: Find best model

Training

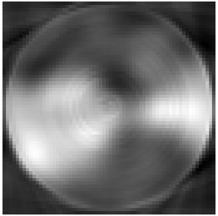
- U-Net used for reconstruction
  - Pre-trained with complete dataset and SNR_y in [-10, 0] for 200 epochs
- Mini-batch gradient descent with batch size 64
- Adam optimizer
- Joint U-Net training possible

Small Scale Experiments - Overall Results

- Learning fails with random graph
- Learning succeeds with defined input graph



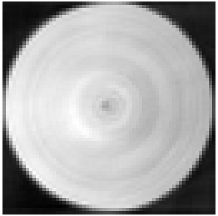
(a) Reconstruction with random input graph



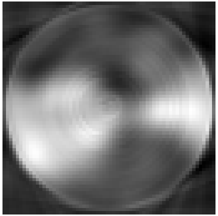
(b) Reconstruction with k-NN input graph $k = 10$

Small Scale Experiments - Overall Results

- Learning fails with random graph
- Learning succeeds with defined input graph
- Components contribute to success of GAT-Denoiser
- Best model with joint U-Net training

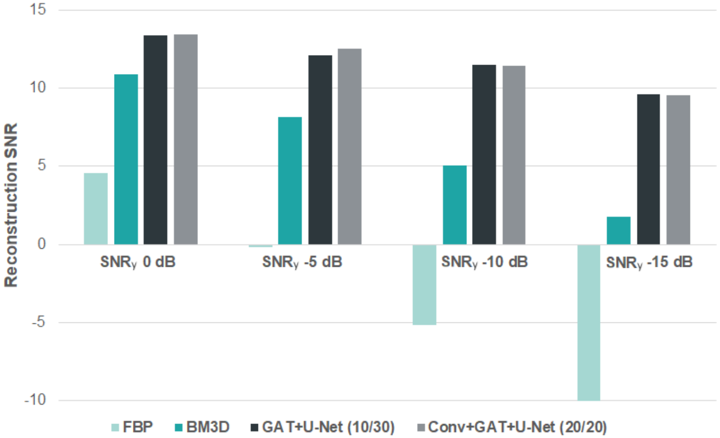


(a) Reconstruction with random input graph

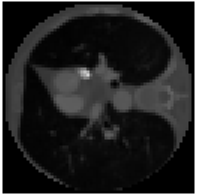


(b) Reconstruction with k-NN input graph $k = 10$

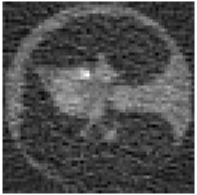
Large Scale Experiment - SNR Results



Large Scale Experiment - Visual Results - SNR_y 0 dB

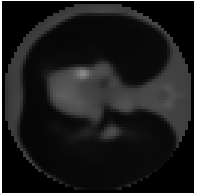


(a) Clean sample

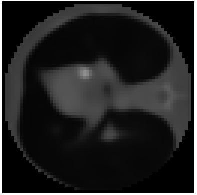


(b) FBP

(c) BM3D



(d) *GAT+U-Net(10/30)*



(e) *Conv+GAT+U-Net(20/20)*

Figure: Large Scale Experiment: Visual results for SNR_y 0 dB.

Large Scale Experiment - Visual Results - SNR_y 0 dB

(a) Clean sample

(b) FBP

(c) BM3D

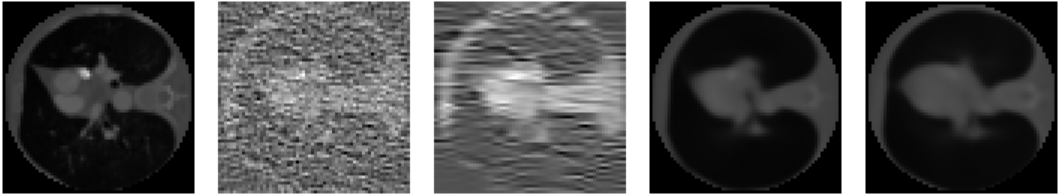
(d) *GAT+U-Net(10/30)*

(e) *Conv+GAT+U-Net(20/20)*

Figure: Large Scale Experiment: Visual results for SNR_y 0 dB.

GAT-Denoiser improves BM3D for SNR_y 0 dB by 27.6%.

Large Scale Experiment - Visual Results - SNR_y -10 dB



(a) Clean sample

(b) FBP

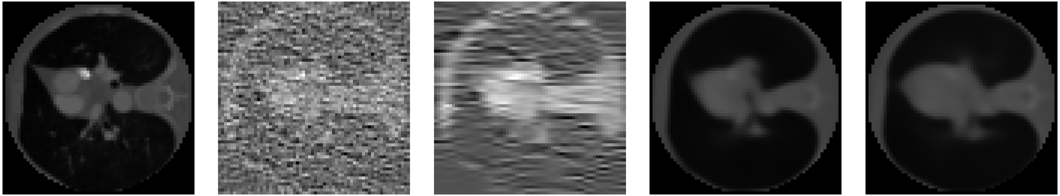
(c) BM3D

(d) *GAT+U-Net(10/30)*

(e) *Conv+GAT+U-Net(20/20)*

Figure: Large Scale Experiment: Visual results for SNR_y -10 dB.

Large Scale Experiment - Visual Results - SNR_y -10 dB

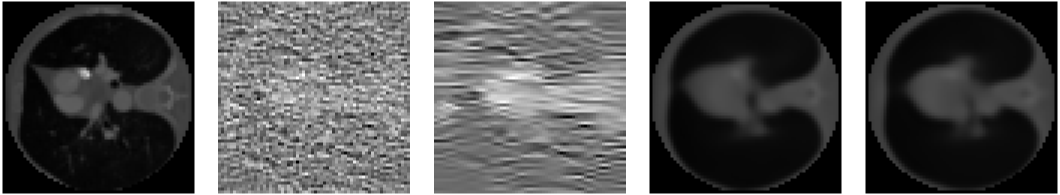


(a) Clean sample (b) FBP (c) BM3D (d) *GAT+U-Net(10/30)* (e) *Conv+GAT+U-Net(20/20)*

Figure: Large Scale Experiment: Visual results for SNR_y -10 dB.

GAT-Denoiser improves BM3D for SNR_y 0 dB by 126.0%.

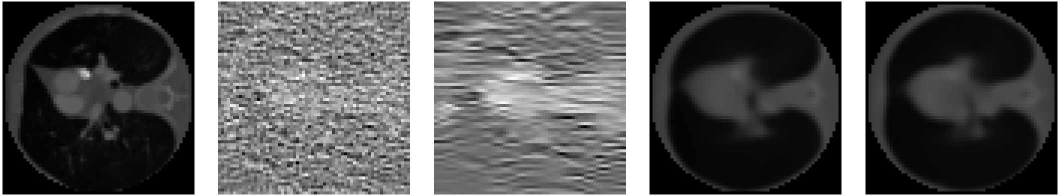
Large Scale Experiment - Visual Results - SNR_y -15 dB



(a) Clean sample (b) FBP (c) BM3D (d) *GAT+U-Net(10/30)* (e) *Conv+GAT+U-Net(20/20)*

Figure: Large Scale Experiment: Visual results for SNR_y -15 dB.

Large Scale Experiment - Visual Results - SNR_y -15 dB



(a) Clean sample (b) FBP (c) BM3D (d) *GAT+U-Net(10/30)* (e) *Conv+GAT+U-Net(20/20)*

Figure: Large Scale Experiment: Visual results for SNR_y -15 dB.

GAT-Denoiser improves BM3D for SNR_y 0 dB by 379.9%.

Outline

Summary

- GAT-Denoiser enables denoising of observations
  - Convolution
  - GAT
  - End-To-End Learning
  - Joint U-Net training boost performance

Summary

- GAT-Denoiser enables denoising of observations
  - Convolution
  - GAT
  - End-To-End Learning
  - Joint U-Net training boost performance
- Evaluated on LoDoPaB-CT dataset
  - Outperformed baseline BM3D by up to 379.9 %

Future Work

- Improve current GAT-Denoiser
- Derive GAT-Denoiser for 3D
- Make it work for unknown angles

Future Work

- Improve current GAT-Denoiser
- Derive GAT-Denoiser for 3D
- Make it work for unknown angles
- Cryo-EM
  - Known angles
  - Unknown angles
  - Work with structural variety in observations

Outline

Questions

References

- Buades, Antoni, Bartomeu Coll, and J-M Morel (2005). "A non-local algorithm for image denoising". In: **2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)**. Vol. 2. IEEE, pp. 60–65. DOI: 10.1109/CVPR.2005.38.
- Coifman, Ronald R et al. (2008). "Graph Laplacian tomography from unknown random projections". In: **IEEE Transactions on Image Processing** 17.10, pp. 1891–1899. DOI: 10.1109/TIP.2008.2002305.
- Dabov, Kostadin et al. (2007). "Image denoising by sparse 3-D transform-domain collaborative filtering". In: **IEEE Transactions on Image Processing** 16.8, pp. 2080–2095. DOI: 10.1109/TIP.2007.901238.
- Leuschner, Johannes et al. (2019). "The lodopab-ct dataset: A benchmark dataset for low-dose ct reconstruction methods". In: **arXiv preprint arXiv:1910.01113**. DOI: 10.1038/s41597-021-00893-z.
- Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). "U-net: Convolutional networks for biomedical image segmentation". In: **International Conference on Medical image computing and computer-assisted intervention**. Springer, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.
- Vaswani, Ashish et al. (2017). "Attention is all you need". In: **Advances in neural information processing systems** 30. DOI: 10.48550/arXiv.1706.03762.