Research Progress Report

Reporter: Minjie Cheng

July 20, 2023

Overview

Research Topic: **Neural Network Modeling based on Optimal Transport.**

- Revisiting Global Pooling via Regularized OT $\Longrightarrow$ OT for Pooling
- A Quasi-Wasserstein Loss for Learning Graph Neural Networks $\Longrightarrow$ OT for GNNs

Revisiting Global Pooling via Regularized OT

- A Global Pooling outputs **the expectation of samples conditioned on different feature dimensions.**

$$f(\boldsymbol{X}) = (\boldsymbol{X} \odot \underbrace{\mathrm{diag}^{-1}(\overbrace{\boldsymbol{P}\mathbf{1}_N}^{\boldsymbol{p}=[p_d]})\boldsymbol{P}}_{\tilde{\boldsymbol{P}}=[p_{n|d}]})\mathbf{1}_N = \|_{d=1}^{D} \mathbb{E}_{n\sim p_{n|d}}[x_{dn}], \tag{1}$$

- The pooling is determined by **the "sample-feature" distribution**
  - Mean-pooling: $f(\boldsymbol{X}) = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n \quad \Longrightarrow \quad \boldsymbol{P} = [\frac{1}{DN}]$
  - Max-pooling: $f(\boldsymbol{X}) = \|_{d=1}^{D} \max_n \{x_{dn}\}_{n=1}^{N} \quad \Longrightarrow \quad \boldsymbol{P} \in \{0, \frac{1}{D}\}^{D\times N}$
  - Attention-pooling: $f(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{a}_{\boldsymbol{X}} \quad \Longrightarrow \quad \boldsymbol{P} = \frac{1}{D}\mathbf{1}_D\boldsymbol{a}_{\boldsymbol{X}}^T$

Revisiting Global Pooling via Regularized OT

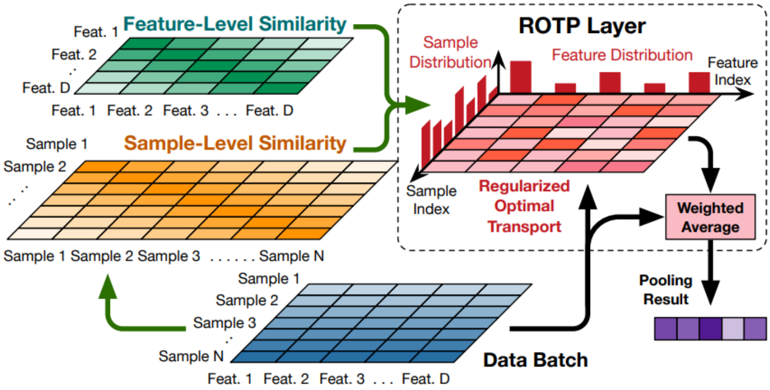
Design a global pooling layer = Determine the "sample-feature dimension" distribution $\boldsymbol{P}$

- Output energy maximization $\Rightarrow$ OT term
- Consider correlation within data $\Rightarrow$ GW structural regularizer
- Avoid sparse distribution $\Rightarrow$ Bregman smoothness regularizer
- Prior of marginal distribution $\Rightarrow$ unbalanced OT regularizer

$$\boldsymbol{P}^*_{\text{rot}}(\boldsymbol{X};\boldsymbol{\theta}) = \arg\min_{\boldsymbol{P}\in\Omega} \overbrace{\underbrace{\langle -\boldsymbol{X},\boldsymbol{P}\rangle}_{\text{OT term}} + \underbrace{\alpha_0\langle C(\boldsymbol{X},\boldsymbol{P}),\boldsymbol{P}\rangle}_{\text{Structural Reg.}}}^{\text{Fused Gromov-Wasserstein discrepancy}} + \underbrace{\alpha_1 \mathrm{R}(\boldsymbol{P})}_{\text{Smoothness Reg.}} + \underbrace{\alpha_2 \mathrm{KL}(\boldsymbol{P}\mathbf{1}_N|\boldsymbol{p}_0) + \alpha_3 \mathrm{KL}(\boldsymbol{P}^T\mathbf{1}_D|\boldsymbol{q}_0)}_{\text{Marginal Reg.}}, \tag{2}$$

$$f_{\text{rot}}(\boldsymbol{X};\boldsymbol{\theta}) = (\boldsymbol{X}\odot \mathrm{diag}^{-1}(\boldsymbol{P}^*_{\text{rot}}(\boldsymbol{X};\boldsymbol{\theta})\mathbf{1}_N)\boldsymbol{P}^*_{\text{rot}}(\boldsymbol{X};\boldsymbol{\theta}))\mathbf{1}_N. \tag{3}$$

Revisiting Global Pooling via Regularized OT



(a) Regularized optimal transport pooling (ROTP) layer

Revisiting Global Pooling via Regularized OT

Theoretical analysis

- Permutation-invariance
- Generalized framework for existing poolings

$f_{rot}(\boldsymbol{X};\boldsymbol{\theta})$	*Mean-pooling*	*Max-pooling*	*Attention-pooling*
α_0	*0*	*0*	*0*
α_1	$\rightarrow\infty$	*0*	$\rightarrow\infty$
α_2	$\rightarrow\infty$	$\rightarrow\infty$	$\rightarrow\infty$
α_3	$\rightarrow\infty$	*0*	$\rightarrow\infty$
$\boldsymbol{p}_0$	$\frac{1}{D}\mathbf{1}_D$	$\frac{1}{D}\mathbf{1}_D$	$\frac{1}{D}\mathbf{1}_D$
$\boldsymbol{q}_0$	$\frac{1}{N}\mathbf{1}_N$	—	$\boldsymbol{a}_{\boldsymbol{X}}$

Revisiting Global Pooling via Regularized OT

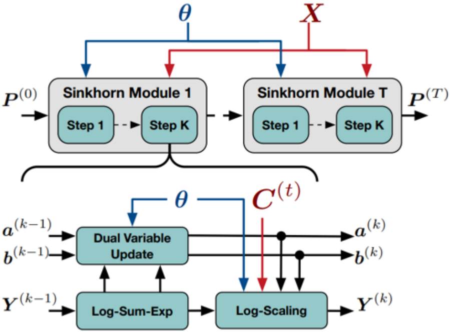
- Discrete Optimal Transport

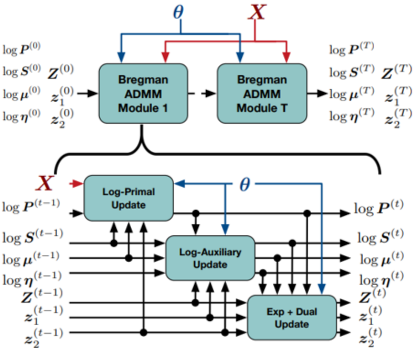
$$W_1(\boldsymbol{\mu}, \boldsymbol{\gamma}) := \min_{\boldsymbol{T} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\gamma})} \langle \boldsymbol{D}, \ \boldsymbol{T} \rangle = \min_{\boldsymbol{T} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\gamma})} \sum_{v, v' \in \mathcal{V} \times \mathcal{V}} t_{vv'} d_{vv'}, \tag{4}$$

  where $\Pi(\boldsymbol{\mu}, \boldsymbol{\gamma}) = \{\boldsymbol{T} \geq \boldsymbol{0} | \boldsymbol{T}\boldsymbol{1}_{|\mathcal{V}|} = \boldsymbol{\mu}, \boldsymbol{T}^{\top}\boldsymbol{1}_{|\mathcal{V}|} = \boldsymbol{\gamma}\}$
- Optimal transportation distances are parameterized distance
- Iterative computation methods: Sinkhorn, Badmm

Revisiting Global Pooling via Regularized OT

Algorithmic modeling: Implement neural network layers by unrolling iterative optimization algorithmic.





Revisiting Global Pooling via Regularized OT

Experiments:

TABLE 1
Comparison on MIL accuracy±Std. (%) for different pooling layers.

Dataset	Messidor	Component	Function
D	687	200	200
#Positive bags	654	423	443
#Negative bags	546	2,707	4,799
#Instances	12,352	36,894	55,536
Min. bag size	8	1	1
Max. bag size	12	53	51
Add	$74.33_{\pm 2.56}$	$93.35_{\pm 0.98}$	$96.26_{\pm 0.48}$
Mean	$74.42_{\pm 2.47}$	$93.32_{\pm 0.99}$	$96.28_{\pm 0.66}$
Max	$73.92_{\pm 3.00}$	$93.23_{\pm 0.76}$	$95.94_{\pm 0.48}$
DeepSet	$74.42_{\pm 2.87}$	$93.29_{\pm 0.95}$	$96.45_{\pm 0.51}$
Mixed	$73.42_{\pm 2.29}$	$\mathbf{93.45_{\pm 0.61}}$	$96.41_{\pm 0.53}$
GatedMixed	$73.25_{\pm 2.38}$	$93.03_{\pm 1.02}$	$96.22_{\pm 0.65}$
Set2Set	$73.58_{\pm 3.74}$	$93.19_{\pm 0.95}$	$96.43_{\pm 0.56}$
Attention	$74.25_{\pm 3.67}$	$93.22_{\pm 1.02}$	$96.31_{\pm 0.66}$
GatedAtt	$73.67_{\pm 2.23}$	$\mathbf{93.42_{\pm 0.91}}$	$\mathbf{96.51_{\pm 0.77}}$
DynamicP	$73.16_{\pm 2.12}$	$93.26_{\pm 1.30}$	$\mathbf{96.47_{\pm 0.58}}$
GNP	$73.54_{\pm 3.68}$	$92.86_{\pm 1.96}$	$96.10_{\pm 1.03}$
OTK	$74.78_{\pm 2.89}$	$93.19_{\pm 0.93}$	$96.31_{\pm 1.02}$
SWE	$74.46_{\pm 3.72}$	$93.32_{\pm 1.26}$	$96.42_{\pm 0.88}$
ROTP_{S}	$\mathbf{75.42_{\pm 2.96}}$	$93.29_{\pm 0.83}$	$\mathbf{96.62_{\pm 0.48}}$
$\text{ROTP}_{\text{B-E}}$ ($\alpha_0 = 0$)	$74.83_{\pm 2.07}$	$93.16_{\pm 1.02}$	$96.17_{\pm 0.43}$
$\text{ROTP}_{\text{B-Q}}$ ($\alpha_0 = 0$)	$75.08_{\pm 2.06}$	$93.13_{\pm 0.94}$	$96.09_{\pm 0.46}$
$\text{ROTP}_{\text{B-E}}$ (learn α_0)	$\mathbf{75.33_{\pm 1.96}}$	$93.16_{\pm 1.08}$	$96.22_{\pm 0.44}$
$\text{ROTP}_{\text{B-Q}}$ (learn α_0)	$\mathbf{75.17_{\pm 2.45}}$	$\mathbf{93.45_{\pm 0.96}}$	$96.22_{\pm 0.48}$

* The top-3 results are bolded and the best result is in red.

TABLE 2
Comparison on graph classification accuracy±Std. (%) for different pooling layers.

Dataset	NCI1	PROTEINS	MUTAG	COLLAB	RDT-B	RDT-M5K	IMDB-B	IMDB-M
#Graphs	4,110	1,113	188	5,000	2,000	4,999	1,000	1,500
Average #Nodes	29.87	39.06	17.93	74.49	429.63	508.52	19.77	13.00
Average #Edges	32.30	72.82	19.79	2,457.78	497.75	594.87	96.53	65.94
#Classes	2	2	2	3	2	5	2	3
Add	$67.96_{\pm 0.43}$	$72.97_{\pm 0.54}$	$\mathbf{89.05_{\pm 0.86}}$	$71.06_{\pm 0.43}$	$80.00_{\pm 1.49}$	$50.16_{\pm 0.97}$	$70.18_{\pm 0.87}$	$47.56_{\pm 0.56}$
Mean	$64.82_{\pm 0.52}$	$66.09_{\pm 0.64}$	$86.53_{\pm 1.62}$	$72.35_{\pm 0.44}$	$83.62_{\pm 1.18}$	$52.44_{\pm 1.24}$	$70.34_{\pm 0.38}$	$48.65_{\pm 0.91}$
Max	$65.95_{\pm 0.76}$	$72.27_{\pm 0.33}$	$85.90_{\pm 1.68}$	$73.07_{\pm 0.57}$	$82.62_{\pm 1.25}$	$44.34_{\pm 1.93}$	$70.24_{\pm 0.54}$	$47.80_{\pm 0.54}$
DeepSet	$66.28_{\pm 0.72}$	$\mathbf{73.76_{\pm 0.47}}$	$87.84_{\pm 0.71}$	$69.74_{\pm 0.66}$	$82.91_{\pm 1.37}$	$47.45_{\pm 0.54}$	$70.84_{\pm 0.71}$	$48.05_{\pm 0.71}$
Mixed	$66.46_{\pm 0.74}$	$72.25_{\pm 0.45}$	$87.30_{\pm 0.87}$	$73.22_{\pm 0.35}$	$84.36_{\pm 2.62}$	$46.67_{\pm 1.63}$	$71.28_{\pm 0.26}$	$48.07_{\pm 0.25}$
GatedMixed	$63.86_{\pm 0.76}$	$69.40_{\pm 1.93}$	$87.94_{\pm 1.28}$	$71.94_{\pm 0.40}$	$80.60_{\pm 3.89}$	$44.78_{\pm 4.53}$	$70.96_{\pm 0.60}$	$48.09_{\pm 0.44}$
Set2Set	$65.10_{\pm 1.12}$	$68.61_{\pm 1.44}$	$87.77_{\pm 0.86}$	$72.31_{\pm 0.73}$	$80.08_{\pm 5.72}$	$49.85_{\pm 2.77}$	$70.36_{\pm 0.85}$	$48.30_{\pm 0.54}$
Attention	$64.35_{\pm 0.61}$	$67.70_{\pm 0.95}$	$88.08_{\pm 1.22}$	$72.57_{\pm 0.41}$	$81.55_{\pm 4.39}$	$51.85_{\pm 0.66}$	$70.60_{\pm 0.38}$	$47.83_{\pm 0.78}$
GatedAtt	$64.66_{\pm 0.52}$	$68.16_{\pm 0.90}$	$86.91_{\pm 1.79}$	$72.31_{\pm 0.37}$	$82.55_{\pm 1.96}$	$51.47_{\pm 0.82}$	$70.52_{\pm 0.31}$	$48.67_{\pm 0.35}$
DynamicP	$62.11_{\pm 0.27}$	$65.86_{\pm 0.85}$	$85.40_{\pm 2.81}$	$70.78_{\pm 0.88}$	$67.51_{\pm 1.82}$	$32.11_{\pm 3.85}$	$69.84_{\pm 0.73}$	$47.59_{\pm 0.48}$
GNP	$\mathbf{68.20_{\pm 0.48}}$	$\mathbf{73.44_{\pm 0.61}}$	$88.37_{\pm 1.25}$	$72.80_{\pm 0.58}$	$81.93_{\pm 2.23}$	$51.80_{\pm 0.61}$	$70.34_{\pm 0.83}$	$48.85_{\pm 0.81}$
ASAP	$68.09_{\pm 0.42}$	$70.42_{\pm 1.45}$	$87.68_{\pm 1.42}$	$68.20_{\pm 2.37}$	$73.91_{\pm 1.50}$	$44.58_{\pm 0.44}$	$68.33_{\pm 2.50}$	$43.92_{\pm 1.13}$
SAGP	$67.48_{\pm 0.65}$	$72.63_{\pm 0.44}$	$87.88_{\pm 2.22}$	$70.19_{\pm 0.55}$	$74.12_{\pm 2.86}$	$46.00_{\pm 1.74}$	$70.34_{\pm 0.74}$	$47.04_{\pm 1.22}$
OTK	$67.96_{\pm 0.55}$	$69.52_{\pm 0.76}$	$86.90_{\pm 1.83}$	$71.35_{\pm 0.91}$	$74.28_{\pm 1.39}$	$50.57_{\pm 1.20}$	$70.94_{\pm 0.79}$	$48.41_{\pm 0.89}$
SWE	$68.06_{\pm 0.98}$	$70.09_{\pm 1.22}$	$85.68_{\pm 2.07}$	$72.17_{\pm 1.29}$	$79.30_{\pm 3.94}$	$51.11_{\pm 1.55}$	$70.34_{\pm 1.05}$	$48.93_{\pm 1.34}$
WEGL	$\mathbf{68.16_{\pm 0.62}}$	$71.58_{\pm 0.94}$	$\mathbf{88.68_{\pm 1.66}}$	$72.55_{\pm 0.69}$	$82.80_{\pm 1.73}$	$52.03_{\pm 0.60}$	$\mathbf{71.94_{\pm 0.75}}$	$49.20_{\pm 0.87}$
ROTP_{S}	$\mathbf{68.27_{\pm 1.06}}$	$\mathbf{73.10_{\pm 0.22}}$	$\mathbf{88.84_{\pm 1.21}}$	$71.20_{\pm 0.55}$	$81.54_{\pm 1.38}$	$51.00_{\pm 0.61}$	$70.74_{\pm 0.80}$	$47.87_{\pm 0.43}$
$\text{ROTP}_{\text{B-E}}$ ($\alpha_0 = 0$)	$66.23_{\pm 0.50}$	$67.71_{\pm 1.70}$	$86.82_{\pm 2.02}$	$73.86_{\pm 0.44}$	$\mathbf{86.80_{\pm 1.19}}$	$52.25_{\pm 0.75}$	$71.72_{\pm 0.88}$	$\mathbf{50.48_{\pm 0.14}}$
$\text{ROTP}_{\text{B-Q}}$ ($\alpha_0 = 0$)	$66.18_{\pm 0.76}$	$69.88_{\pm 0.87}$	$85.42_{\pm 1.10}$	$\mathbf{74.14_{\pm 0.24}}$	$\mathbf{87.72_{\pm 1.03}}$	$\mathbf{52.79_{\pm 0.60}}$	$\mathbf{72.34_{\pm 0.50}}$	$49.36_{\pm 0.52}$
$\text{ROTP}_{\text{B-E}}$ (learn α_0)	$65.90_{\pm 0.94}$	$70.19_{\pm 0.66}$	$88.01_{\pm 1.51}$	$\mathbf{74.05_{\pm 0.34}}$	$86.78_{\pm 1.14}$	$\mathbf{52.77_{\pm 0.69}}$	$71.76_{\pm 0.62}$	$\mathbf{50.28_{\pm 0.86}}$
$\text{ROTP}_{\text{B-Q}}$ (learn α_0)	$65.96_{\pm 0.32}$	$70.12_{\pm 1.17}$	$86.79_{\pm 1.81}$	$\mathbf{74.27_{\pm 0.47}}$	$\mathbf{88.67_{\pm 0.99}}$	$\mathbf{52.84_{\pm 0.60}}$	$\mathbf{71.78_{\pm 1.00}}$	$\mathbf{49.44_{\pm 0.46}}$

* For each dataset, the top-3 results are bolded and the best result is in red.

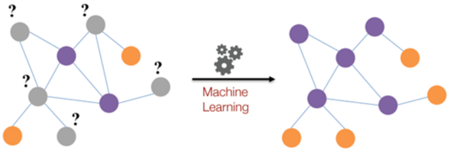
TABLE 3
Comparisons on graph set classification accuracy±Std. (%) for different pooling layers.

Dataset	DECAGON DiBr-APND	DECAGON Anae-Fati	DECAGON PleuP-Diar	FEARS
#Graph sets	6,309	2,922	2,842	6,338
#Positive sets	3,189	1,526	1,422	3,169
Positive label	Difficulty breathing	Anaemia	Pleural pain	Non-myopathy
#Negative sets	3,120	1,396	1,420	3,169
Negative label	Pressure decreased	Fatigue	Diarrhea	Myopathy
Set size	2	2	2	2∼52
Add	$50.86_{\pm 0.97}$	$63.15_{\pm 1.79}$	$\mathbf{62.32_{\pm 1.08}}$	$75.89_{\pm 1.33}$
Mean	$51.10_{\pm 1.09}$	$61.95_{\pm 2.60}$	$61.30_{\pm 2.68}$	$72.42_{\pm 1.51}$
Max	$50.59_{\pm 0.77}$	$61.88_{\pm 2.03}$	$60.11_{\pm 2.03}$	$\mathbf{82.02_{\pm 0.72}}$
DeepSet	$49.83_{\pm 1.07}$	$56.24_{\pm 5.20}$	$51.78_{\pm 3.10}$	$\mathbf{82.40_{\pm 1.56}}$
Mixed	$51.13_{\pm 0.99}$	$\mathbf{63.83_{\pm 1.19}}$	$60.91_{\pm 2.12}$	$81.54_{\pm 1.13}$
GatedMixed	$51.39_{\pm 0.63}$	$61.50_{\pm 1.61}$	$59.12_{\pm 2.12}$	$81.88_{\pm 1.14}$
Set2Set	$50.72_{\pm 1.71}$	$59.35_{\pm 2.04}$	$55.01_{\pm 3.59}$	$79.29_{\pm 0.84}$
Attention	$50.52_{\pm 1.10}$	$61.40_{\pm 2.03}$	$61.33_{\pm 2.40}$	$75.98_{\pm 0.74}$
GatedAtt	$50.74_{\pm 0.61}$	$62.15_{\pm 0.77}$	$58.80_{\pm 1.18}$	$75.84_{\pm 1.29}$
DynamicP	$51.01_{\pm 1.88}$	$55.93_{\pm 1.56}$	$52.58_{\pm 2.91}$	$74.00_{\pm 1.61}$
GNP	$50.00_{\pm 1.88}$	$53.98_{\pm 6.34}$	$52.58_{\pm 4.68}$	$62.71_{\pm 15.55}$
ASAP	$50.89_{\pm 0.82}$	$63.66_{\pm 1.81}$	$60.67_{\pm 2.69}$	$77.15_{\pm 1.13}$
SAGP	$49.87_{\pm 0.77}$	$63.62_{\pm 1.28}$	$59.86_{\pm 2.43}$	$77.29_{\pm 1.04}$
OTK	$50.96_{\pm 1.11}$	$63.68_{\pm 1.59}$	$\mathbf{61.66_{\pm 2.39}}$	$79.40_{\pm 1.08}$
SWE	$51.05_{\pm 2.15}$	$63.21_{\pm 2.02}$	$61.37_{\pm 3.13}$	$80.64_{\pm 1.86}$
WEGL	$\mathbf{51.67_{\pm 0.85}}$	$\mathbf{63.79_{\pm 2.54}}$	$61.36_{\pm 2.30}$	$81.98_{\pm 0.77}$
ROTP_{S}	$\mathbf{51.96_{\pm 0.71}}$	$62.91_{\pm 1.13}$	$59.40_{\pm 0.90}$	$79.75_{\pm 0.71}$
$\text{ROTP}_{\text{B-E}}$	$51.26_{\pm 0.84}$	$\mathbf{63.86_{\pm 2.41}}$	$\mathbf{62.57_{\pm 1.34}}$	$\mathbf{82.55_{\pm 0.42}}$
$\text{ROTP}_{\text{B-Q}}$	$\mathbf{52.72_{\pm 0.66}}$	$63.15_{\pm 1.27}$	$60.88_{\pm 1.65}$	$81.43_{\pm 1.12}$

* The top-3 results are bolded and the best result is in red.

A Quasi-Wasserstein Loss for Learning Graph Neural Networks

Motivation: Eliminate Inconsistence of GNNs



Loss function: $\min_{\theta} \sum_{v \in \mathcal{V}_L} \psi(g_v(\boldsymbol{X}, \boldsymbol{A}; \theta),\ \boldsymbol{y}_v).$ (5)

- ψ is often implemented as cross-entropy loss, KL-divergence, Euclidean distance, and so on.
- The learning paradigm in (4) treats each node independently and evenly, inconsistent with the **non-i.i.d.** nature of graph-structured data.

A Quasi-Wasserstein Loss for Learning Graph Neural Networks

- Wasserstein distance:

$$QW(\widehat{\boldsymbol{Y}}_{\mathcal{V}_L}, \boldsymbol{Y}_{\mathcal{V}_L}) = \sum_{c=1}^{C} W_1^{(P)}(\hat{\boldsymbol{y}}_{\mathcal{V}_L}^{(c)}, \boldsymbol{y}_{\mathcal{V}_L}^{(c)}) \tag{6}$$

- Wasserstein distance are parameterized distance

$$W_1(\boldsymbol{\mu}, \boldsymbol{\gamma}) := \min_{\boldsymbol{T}\in\Pi(\boldsymbol{\mu},\boldsymbol{\gamma})} \langle \boldsymbol{D},\ \boldsymbol{T} \rangle = \min_{\boldsymbol{T}\in\Pi(\boldsymbol{\mu},\boldsymbol{\gamma})} \sum_{v,v'\in\mathcal{V}\times\mathcal{V}} t_{vv'} d_{vv'}, \tag{7}$$

where $\Pi(\boldsymbol{\mu}, \boldsymbol{\gamma}) = \{\boldsymbol{T} \geq \boldsymbol{0} | \boldsymbol{T}\boldsymbol{1}_{|\mathcal{V}|} = \boldsymbol{\mu}, \boldsymbol{T}^{\top}\boldsymbol{1}_{|\mathcal{V}|} = \boldsymbol{\gamma}\}$

- Optimal Transport on Graphs– minimum-cost flow problem[1]

$$W_1(\boldsymbol{\mu}, \boldsymbol{\gamma}) = \min_{\boldsymbol{f}\in\Omega^{|\mathcal{E}|}} \|\mathrm{diag}(\boldsymbol{w})\boldsymbol{f}\|_1 \qquad s.t.\ \boldsymbol{S}_{\mathcal{V}}\boldsymbol{f} = \boldsymbol{\gamma} - \boldsymbol{\mu}, \tag{8}$$

where

$$s_{ve} = \begin{cases} 1 & \text{if } v \text{ is "head" of edge } e \\ -1 & \text{if } v \text{ is "tail" of edge } e \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Montacer Essid and Justin Solomon. Quadratically regularized optimal transport on graphs.391 SIAM Journal on Scientific Computing, 40(4):A1961–A1986, 2018.
Enrico Facca and Michele Benzi. Fast iterative solution of the optimal transport problem on393 graphs. SIAM Journal on Scientific Computing, 43(3):A2295–A2319, 2021

A Quasi-Wasserstein Loss for Learning Graph Neural Networks

- Minimum-cost flow problem:

$$\begin{aligned} QW(\widehat{\boldsymbol{Y}}_{\mathcal{V}_L}, \boldsymbol{Y}_{\mathcal{V}_L}) &= \sum\nolimits_{c=1}^{C} W_1^{(P)}(\hat{\boldsymbol{y}}_{\mathcal{V}_L}^{(c)}, \boldsymbol{y}_{\mathcal{V}_L}^{(c)}) \\ &= \sum\nolimits_{c=1}^{C} \min\nolimits_{\boldsymbol{f}^{(c)}\in\Omega} \|\mathrm{diag}(\boldsymbol{w})\boldsymbol{f}^{(c)}\|_1 \quad s.t.\, \boldsymbol{S}_{\mathcal{V}_L}\boldsymbol{f}^{(c)} = \boldsymbol{y}_{\mathcal{V}_L}^{(c)} - \hat{\boldsymbol{y}}_{\mathcal{V}_L}^{(c)} \\ &= \min\nolimits_{\boldsymbol{F}\in\Omega_C} \|\mathrm{diag}(\boldsymbol{w})\boldsymbol{F}\|_1 \quad s.t.\, \boldsymbol{S}_{\mathcal{V}_L}\boldsymbol{F} = \boldsymbol{Y}_{\mathcal{V}_L} - g_{\mathcal{V}_L}(\boldsymbol{A}, \boldsymbol{X}; \theta), \end{aligned} \tag{10}$$

- Learning Paradigm:

$$\min\nolimits_{\theta} \underbrace{\min\nolimits_{\boldsymbol{F}\in\Omega} \|\mathrm{diag}(\boldsymbol{w})\boldsymbol{F}\|_1 \quad s.t.\, \boldsymbol{S}_{\mathcal{V}_L}\boldsymbol{F} = \boldsymbol{Y}_{\mathcal{V}_L} - g_{\mathcal{V}_L}(\boldsymbol{X}, \boldsymbol{A}; \theta)}_{QW(\widehat{\boldsymbol{Y}}_{\mathcal{V}_L}, \boldsymbol{Y}_{\mathcal{V}_L})}. \tag{11}$$

$$\Downarrow$$

$$\min\nolimits_{\theta, \boldsymbol{F}\in\Omega} \|\mathrm{diag}(\boldsymbol{w})\boldsymbol{F}\|_1 + \lambda \underbrace{B_{\phi}(g_{\mathcal{V}_L}(\boldsymbol{X}, \boldsymbol{A}; \theta) + \boldsymbol{S}_{\mathcal{V}_L}\boldsymbol{F},\ \boldsymbol{Y}_{\mathcal{V}_L})}_{\sum_{v\in\mathcal{V}_L} \psi(g_v(\boldsymbol{X}, \boldsymbol{A}; \theta) + \boldsymbol{S}_v\boldsymbol{F},\ \boldsymbol{y}_v)}. \tag{12}$$

- A New Transductive Prediction Paradigm:

$$\tilde{\boldsymbol{y}}_v := g_v(\boldsymbol{X}, \boldsymbol{A}(\boldsymbol{F}^*; \xi^*); \theta^*) + \boldsymbol{S}_v\boldsymbol{F}^*, \tag{13}$$

A Quasi-Wasserstein Loss for Learning Graph Neural Networks

Experiments:

Table 2: Comparisons on node classification accuracy (%) on homophilic graphs.

Model	Method	Cora	Citeseer	Pubmed	Computers	Photo	Improve
GCN	(1)	87.14$_{\pm1.01}$	79.86$_{\pm0.67}$	86.74$_{\pm0.27}$	83.32$_{\pm0.33}$	88.26$_{\pm0.73}$	
	(1)+LPA	86.34$_{\pm1.45}$	78.51$_{\pm1.22}$	84.72$_{\pm0.70}$	82.48$_{\pm0.69}$	88.10$_{\pm1.31}$	-1.07
	QW	86.95$_{\pm1.12}$	81.30$_{\pm0.37}$	87.89$_{\pm0.44}$	88.39$_{\pm0.55}$	93.80$_{\pm0.37}$	+2.60
APPNP	(1)	88.14$_{\pm0.73}$	80.47$_{\pm0.74}$	88.12$_{\pm0.31}$	85.32$_{\pm0.37}$	88.51$_{\pm0.31}$	
	QW	88.65$_{\pm1.00}$	80.94$_{\pm0.61}$	89.39$_{\pm0.31}$	84.69$_{\pm0.46}$	93.25$_{\pm0.38}$	+0.94
BernNet	(1)	88.28$_{\pm1.00}$	79.81$_{\pm0.79}$	88.87$_{\pm0.38}$	87.61$_{\pm0.46}$	93.68$_{\pm0.28}$	
	QW	**89.03$_{\pm0.76}$**	**81.35$_{\pm0.71}$**	89.03$_{\pm0.38}$	89.14$_{\pm0.39}$	94.28$_{\pm0.44}$	+0.92
ChebNetII	(1)	88.26$_{\pm0.89}$	80.00$_{\pm0.74}$	88.57$_{\pm0.36}$	86.58$_{\pm0.71}$	93.50$_{\pm0.34}$	
	QW	87.98$_{\pm0.80}$	79.47$_{\pm0.70}$	**89.15$_{\pm0.44}$**	**89.52$_{\pm0.54}$**	**94.84$_{\pm0.37}$**	+0.81

Table 3: Comparisons on node classification accuracy (%) on heterophilic graphs.

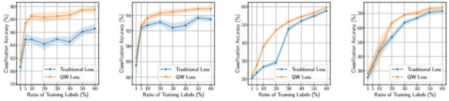
Model	Method	Squirrel	Chameleon	Actor	Texas	Cornell	Improve
GCN	(1)	46.55$_{\pm1.15}$	63.57$_{\pm1.16}$	34.00$_{\pm1.28}$	77.21$_{\pm3.28}$	61.91$_{\pm5.11}$	
	(1)+LPA	44.81$_{\pm1.81}$	60.90$_{\pm1.63}$	32.43$_{\pm1.59}$	78.69$_{\pm6.47}$	68.72$_{\pm5.95}$	+0.46
	QW	51.04$_{\pm0.51}$	67.77$_{\pm0.92}$	38.09$_{\pm0.50}$	84.10$_{\pm2.95}$	84.26$_{\pm2.98}$	+8.40
APPNP	(1)	36.15$_{\pm0.75}$	52.93$_{\pm1.71}$	40.46$_{\pm0.64}$	91.31$_{\pm1.97}$	87.66$_{\pm2.13}$	
	QW	37.11$_{\pm0.60}$	53.76$_{\pm1.25}$	40.78$_{\pm0.74}$	91.48$_{\pm2.30}$	87.87$_{\pm2.34}$	+0.50
BernNet	(1)	51.15$_{\pm1.09}$	67.96$_{\pm1.05}$	40.72$_{\pm0.80}$	93.28$_{\pm1.48}$	90.21$_{\pm2.35}$	
	QW	53.29$_{\pm0.65}$	70.96$_{\pm1.31}$	40.91$_{\pm0.71}$	93.44$_{\pm1.80}$	**90.85$_{\pm2.34}$**	+1.23
ChebNetII	(1)	57.78$_{\pm0.84}$	71.71$_{\pm1.40}$	40.70$_{\pm0.77}$	92.79$_{\pm1.48}$	88.94$_{\pm2.78}$	
	QW	**59.55$_{\pm0.86}$**	**74.05$_{\pm1.12}$**	**41.37$_{\pm0.67}$**	**93.93$_{\pm0.98}$**	87.23$_{\pm3.62}$	+0.84

Table 8: Comparisons on node classification accuracy (%) on homophilic graphs.

Model	Method	Cora	Citeseer	Pubmed	Computers	Photo	Improve
MLP	(1)	77.16$_{\pm1.10}$	76.71$_{\pm0.86}$	86.14$_{\pm0.32}$	84.32$_{\pm0.53}$	89.42$_{\pm0.39}$	
	QW	76.77$_{\pm0.99}$	77.74$_{\pm0.60}$	86.56$_{\pm0.41}$	82.39$_{\pm0.40}$	89.51$_{\pm0.34}$	-0.216
GAT	(1)	89.20$_{\pm0.79}$	80.75$_{\pm0.78}$	87.42$_{\pm0.33}$	90.08$_{\pm0.36}$	94.38$_{\pm0.25}$	
	QW	88.56$_{\pm0.92}$	80.19$_{\pm0.64}$	88.38$_{\pm0.23}$	90.41$_{\pm0.28}$	94.65$_{\pm0.24}$	+0.05

Table 9: Comparisons on node classification accuracy (%) on heterophilic graphs.

Model	Method	Squirrel	Chameleon	Actor	Texas	Cornell	Improve
MLP	(1)	32.19$_{\pm0.77}$	48.21$_{\pm1.47}$	40.61$_{\pm0.60}$	91.80$_{\pm1.31}$	88.30$_{\pm2.55}$	
	QW	34.41$_{\pm0.47}$	49.61$_{\pm1.31}$	40.63$_{\pm0.51}$	91.15$_{\pm2.30}$	88.72$_{\pm2.77}$	+0.68
GAT	(1)	48.20$_{\pm1.67}$	64.31$_{\pm2.01}$	35.68$_{\pm0.60}$	80.00$_{\pm3.11}$	68.09$_{\pm2.13}$	
	QW	55.03$_{\pm1.35}$	67.35$_{\pm1.42}$	33.83$_{\pm1.07}$	80.33$_{\pm1.97}$	70.21$_{\pm2.13}$	+2.09



(a) Computers (b) Photo (c) Squirrel (d) Chameleon

Figure 2: Illustrations of the learning methods' performance given different amounts of labeled nodes.