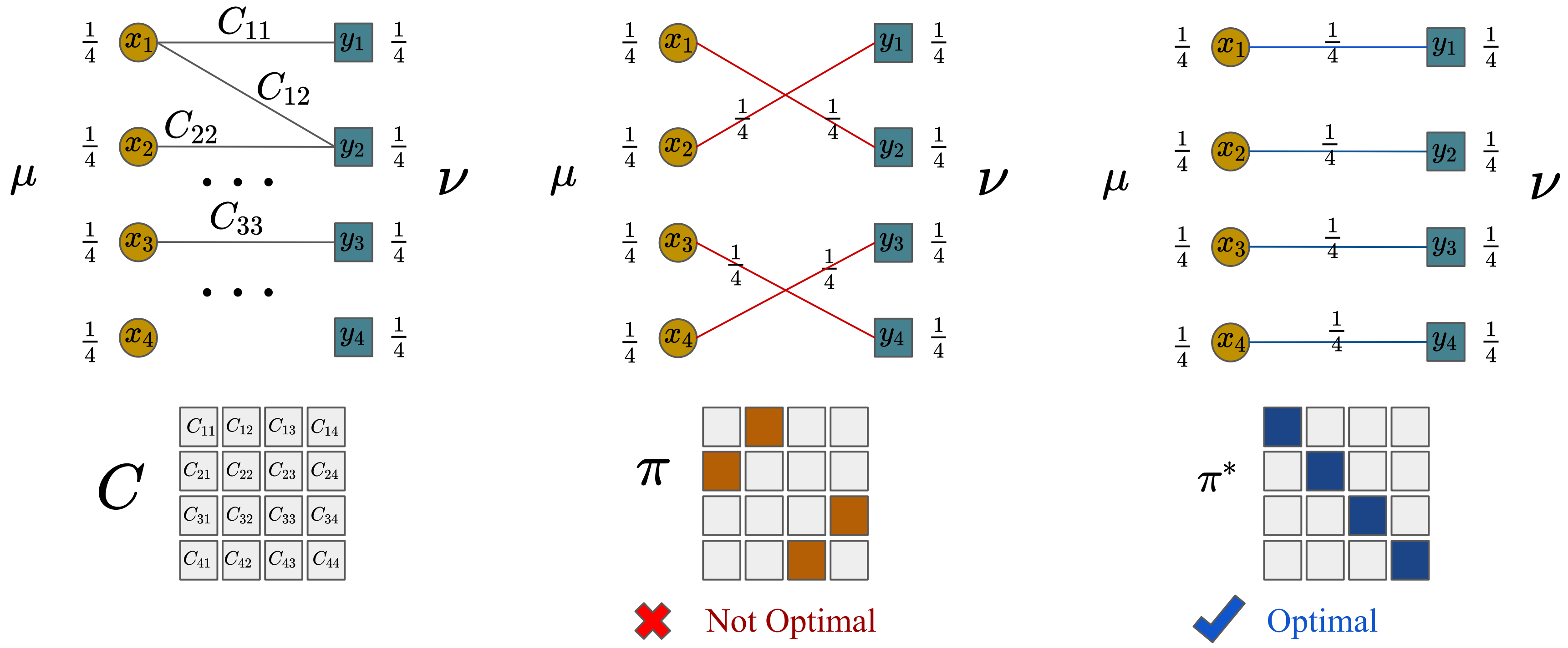
Improving Mini-batch Optimal Transport via Partial Transportation

Khai Nguyen[1*], Dang Nguyen[2*], The-Anh Vu-Le[2], Tung Pham[2], Nhat Ho[1]

[1]Department of Statistics and Data Sciences, University of Texas at Austin
[2]VinAI Research

Optimal Transport



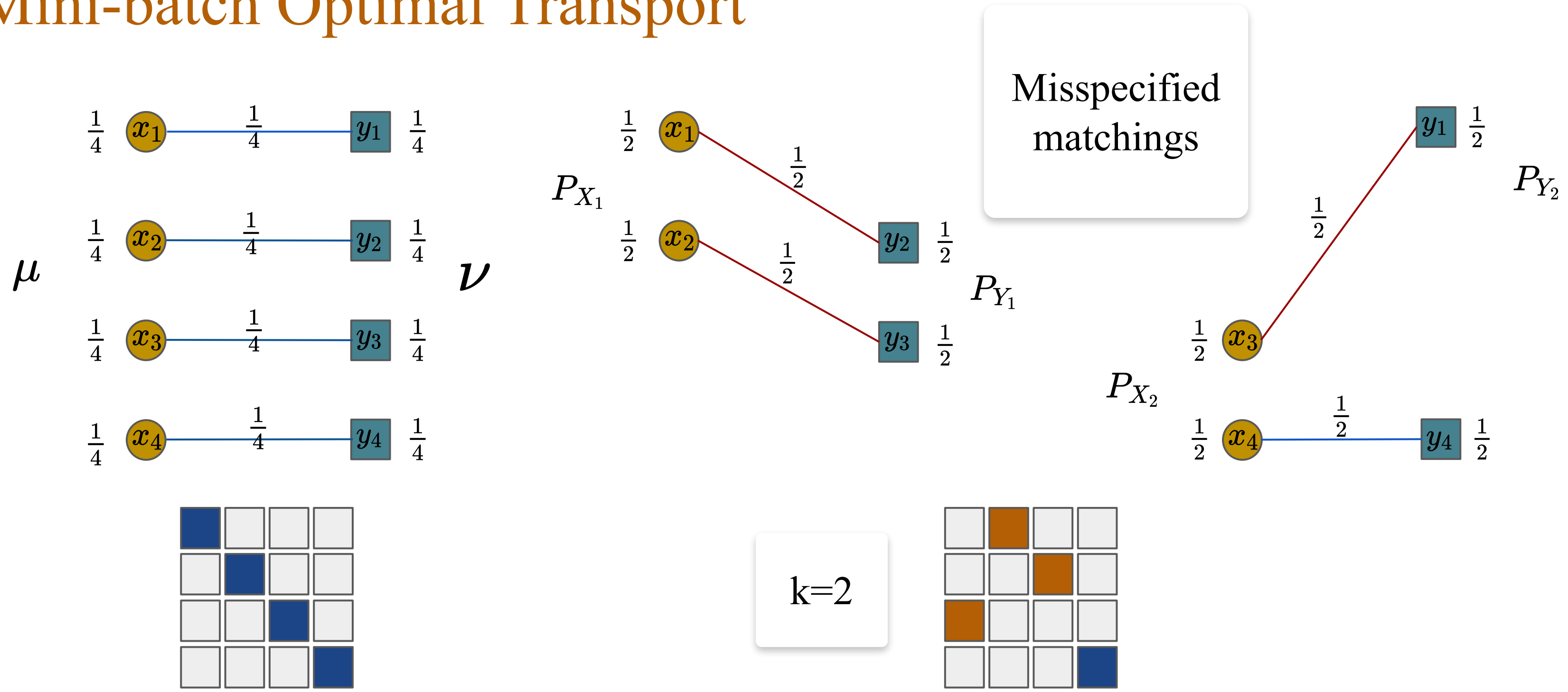
Mini-batch Optimal Transport

The number of supports is large?
e.g., millions

Repeated computation?
e.g., deep learning

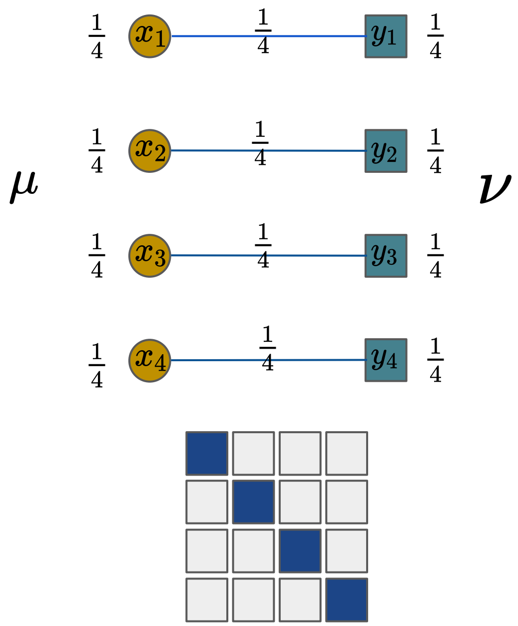
- ❑ Impossible to store the cost matrix C in the computational graph
- ❑ Slow computation of OT losses which leads to slow training

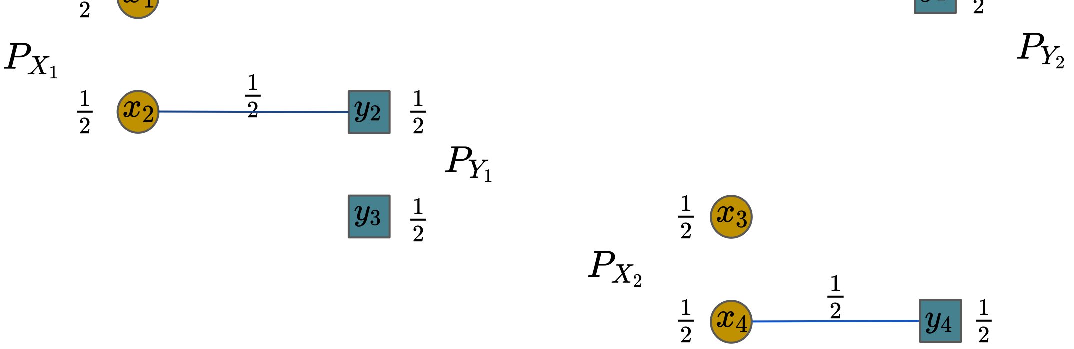
Mini-batch Optimal Transport



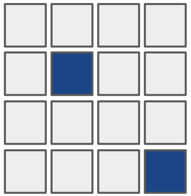
Mini-batch Partial Optimal Transport

$$s = \frac{1}{2}$$



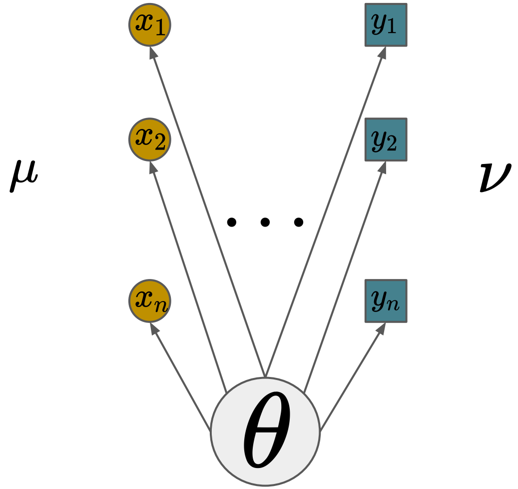


k=2



Alleviate misspecified matchings

Training deep networks with m-POT loss



Supports are functions of parameters of neural networks

Set $\nabla_\theta^k = 0$

For $i = 1$ to k

Compute $X_i(\theta), Y_i(\theta)$

Compute $\nabla_\theta^k = \nabla_\theta^k + \frac{1}{k}\nabla_\theta POT^s(P_{X_i(\theta)}, P_{Y_i(\theta)})$

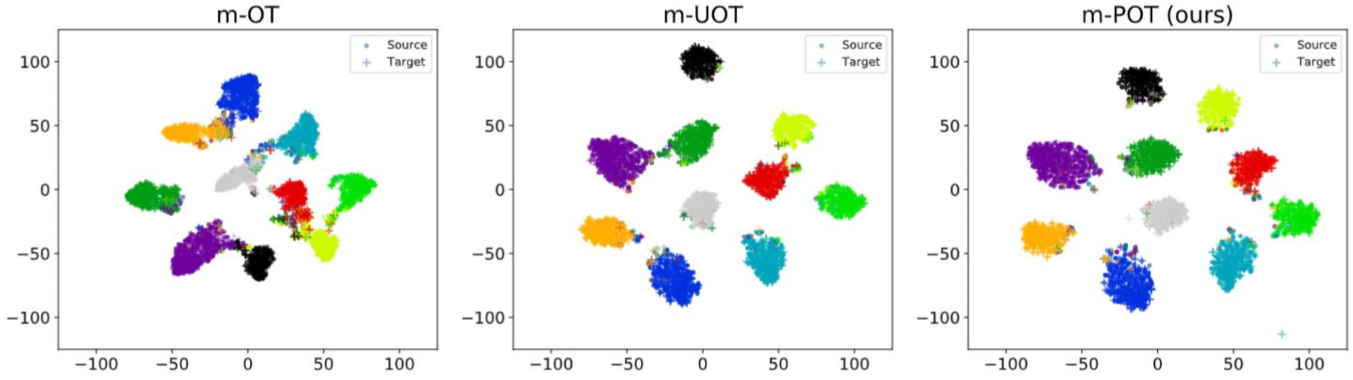
Update θ based on the stochastic gradient ∇_θ^k

- ❑ Only one OT problem in memory at a time
- ❑ Parallel training

Experiments on Deep Domain Adaptation

Adapting classification on digits datasets

Method	SVHN to MNIST	USPS to MNIST	MNIST to USPS	Avg
DANN	95.80 ± 0.29	94.71 ± 0.12	91.63 ± 0.53	94.05
ALDA	98.81 ± 0.08	98.29 ± 0.07	95.29 ± 0.16	97.46
m-OT	94.18 ± 0.32	96.71 ± 0.24	86.93 ± 1.16	92.60
m-UOT	98.89 ± 0.13	98.54 ± 0.20	95.83 ± 0.05	97.75
m-POT (Ours)	$\mathbf{98.98} \pm \mathbf{0.08}$	$\mathbf{98.63} \pm \mathbf{0.13}$	$\mathbf{96.04} \pm \mathbf{0.02}$	**97.88**



Experiments on Deep Domain Adaptation

Adapting classification on Office-Home datasets

Method	A2C	A2P	A2R	C2A	C2P	C2R	P2A	P2C	P2R	R2A	R2C	R2P	Avg
RESNET-50 (*)	34.90	50.00	58.00	37.40	41.90	46.20	38.50	31.20	60.40	53.90	41.20	59.90	46.1
DANN	47.92	67.08	74.85	53.80	63.47	66.42	52.99	44.35	74.43	65.53	52.96	79.41	61.93
CDAN-E (*)	52.50	71.40	76.10	59.70	69.90	71.50	58.70	50.30	77.50	70.50	57.90	83.50	66.60
ALDA	54.04	74.89	77.14	61.37	70.62	72.75	60.32	51.03	76.66	67.90	55.94	81.87	67.04
ROT (*)	47.20	71.80	76.40	58.60	68.10	70.20	56.50	45.00	75.80	69.40	52.10	80.60	64.30
m-OT	51.75	70.01	75.79	59.60	66.46	70.07	57.60	47.88	75.29	66.82	55.71	78.11	64.59
m-UOT	54.99	74.45	80.78	65.66	**74.93**	74.91	64.70	53.42	80.01	74.58	59.88	83.73	70.17
m-POT (Ours)	55.65	73.80	80.76	66.34	74.88	76.16	64.46	53.38	**80.60**	74.55	59.71	83.81	70.34
TS-OT (Ours)	53.89	71.01	77.13	59.82	69.20	71.95	59.18	51.17	76.54	66.46	56.97	80.19	66.13
TS-UOT (Ours)	56.35	73.56	80.16	65.02	73.12	76.50	63.66	54.49	79.97	71.24	60.11	82.92	69.76
TS-POT (Ours)	**57.06**	**76.13**	**81.53**	**68.44**	72.82	**76.53**	**66.21**	**54.87**	80.39	**75.57**	**60.50**	**84.31**	**71.20**

Method	Accuracy
DANN	67.63 ± 0.34
ALDA	71.22 ± 0.12
m-OT	62.42 ± 0.12
m-UOT	72.34 ± 0.32
m-POT (Ours)	73.59 ± 0.15
TS-OT (Ours)	69.14 ± 0.72
TS-UOT (Ours)	70.91 ± 0.11
TS-POT (Ours)	$\mathbf{75.96 \pm 0.44}$

Adapting classification on VISDA dataset

Conclusion

- Using partial optimal transport (POT) could alleviate misspecified matchings in mini-batch optimal transport:
  - Replacing OT by POT in mini-batch losses could improve the performance.

- Two stage training is better than the conventional training when having two computational memories e.g., RAM and GPUs' memory.

- Future works
  - Develop algorithms to choose the fraction of masses $\boldsymbol{s}$.

Thank you for listening!

Khai Nguyen: khainb@utexas.edu @KhaiBaNguyen