Unbalanced minibatch Optimal Transport

Applications to Domain Adaptation

Kilian Fatras, Thibault Séjourné, Nicolas Courty, Rémi Flamary

ICML 2021

UBS, INRIA, CNRS, IRISA

Unbalanced Optimal Transport Introduction

Unbalanced Optimal Transport

Definition

Unbalanced Optimal Transport measures the distance between probablity distributions, but with relaxed marginals.

$$\mathrm{UOT}^{\tau,\varepsilon}(\alpha,\beta,c) = \min_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int c d\pi + \varepsilon \mathtt{KL}(\pi|\alpha\otimes\beta) + \tau(\mathtt{KL}(\pi_1\|\alpha) + \mathtt{KL}(\pi_2\|\beta)),$$

where π is the transport plan, π_1 and π_2 the plan's marginals, $\tau \geq 0$ is the marginal penalization and $\varepsilon \geq 0$ is the regularization coefficient.

Robustness of UOT

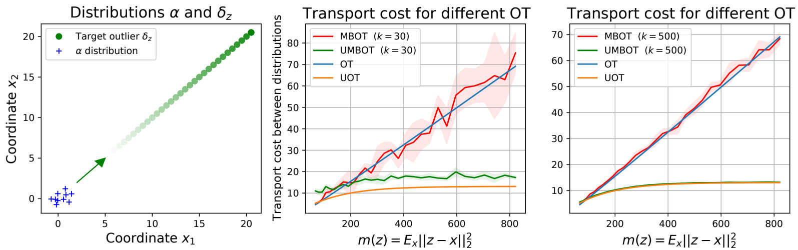
Lemma

Take (α, β) two probability distributions. For $\zeta \in [0,1]$, write $\tilde{\alpha} = \zeta\alpha + (1-\zeta)\delta_{\boldsymbol{z}}$. Write $m(\boldsymbol{z}) = \int C(\boldsymbol{z}, \boldsymbol{y}) d\beta(\boldsymbol{y})$.

$$\mathrm{UOT}^{\tau,0}(\tilde{\alpha}, \beta, C) \lesssim \zeta\, \mathrm{UOT}^{\tau,0}(\alpha, \beta, C) + 2\tau(1-\zeta)(1 - e^{-m(\boldsymbol{z})/2\tau})$$

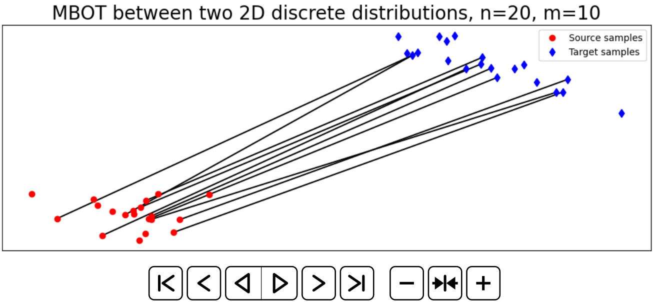
Let (f, g) be the optimal dual potentials of $\mathrm{OT}(\alpha, \beta)$, and y^* in β's support.

$$\mathrm{OT}(\tilde{\alpha}, \beta) \geq \zeta\, \mathrm{OT}(\alpha, \beta) + (1-\zeta)\Big(C(\boldsymbol{z}, y^*) - g(y^*) + \int g d\beta\Big)$$



Minibatch Optimal Transport

Minibatch Optimal Transport definition



Idea : Compute OT between the minibatches from domains

Expectation of minibatches

$$E_h(\alpha, \beta, C) := \mathbb{E}_{(X,Y)\sim\alpha^{\otimes m}\otimes\beta^{\otimes m}}[h(\mu_m, \mu_m, C(X, Y))]$$

- Can be defined for OT variants h
- Studied in [Fatras et al., 2020, Fatras et al., 2021]

Estimate minibatch OT distance

Definition (Estimators)

$$\overline{h}^m(X,Y) := \binom{n}{m}^{-2} \sum_{I,J \in \mathcal{P}_m} h(\mu_m, \mu_m, C_{I,J})$$

where $\mathcal{P}_m$ is the set of all m-tuples without replacement and ordered. Pick an integer $k > 0$ and let D_k be a set of cardinality k whose elements are minibatches drawn uniformly at random. Then,

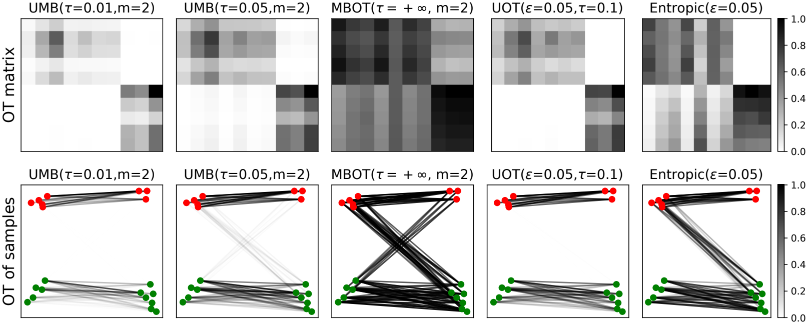
$$\widetilde{h}_k(X,Y) := k^{-1} \sum_{(I,J) \in D_k} h(\mu_m, \mu_m, C_{I,J})$$

Proposition

We have the following properties:

- $\widetilde{\mathrm{UOT}}_k, \overline{\mathrm{UOT}}^m$ *are unbiased estimators of* E_{UOT}
- *Strictly positive losses:* $\widetilde{\mathrm{UOT}}_k(X,Y) > 0$, $\overline{\mathrm{UOT}}^m(X,Y) > 0$

Unbalanced minibatch OT plan



Limits of unbalanced UOT

- Find the correct τ
- Lazy gradients for too small τ

Statistical and optimization properties

Theorem (Maximal deviation bound)

With probability at least $1-\delta$ on the draw of X, Y and D_k we have:

$$|\widetilde{\mathrm{UOT}^{\tau,\varepsilon}}_k^m(X,Y) - E_{\mathrm{UOT}}| \leq \mathcal{O}\left(\sqrt{\frac{\log(\frac{2}{\delta})}{2\lfloor\frac{n}{m}\rfloor}} + \sqrt{\frac{2\log(\frac{2}{\delta})}{k}}\right),$$

SGD converges [Majewski et al., 2018, Davis et al., 2020] if:

- $\overline{UOT}^m$ is an unbiased estimator of E_{UOT}
- Exchange Clarke gradients and expectations

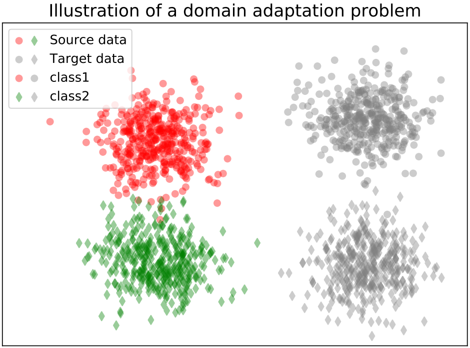
Theorem

Let $\hat{X}, \{\hat{Y}_\theta\}_{\theta\in\Theta}$ be two m-tuples of random vectors compactly supported and C^m a $\mathbf{C}^1$ cost. We have:

$$\partial_\theta\, \mathbb{E}[\mathrm{UOT}^{\tau,\varepsilon}(\mu_m, \mu_m, C^m(\hat{X}, \hat{Y}_\theta))] = \mathbb{E}[\partial_\theta\, \mathrm{UOT}^{\tau,\varepsilon}(\mu_m, \mu_m, C^m(\hat{X}, \hat{Y}_\theta))],$$

Experiments

Domain adaptation



Domain adaptation (DA) setting

- Two domains, only one with labels
- Share the same label distribution
- Goal: Classify unlabelled target data with source labelled data

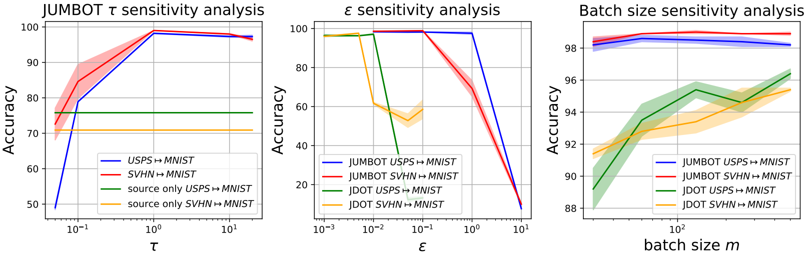
Office Home experiments

We replace minibatch OT by unbalanced minibatch OT in the state of the art DEEPJDOT algorithm [Damodaran et al., 2018]. This allows to reduce the weight of non optimal connections between samples. Our method is called JUMBOT.

	Method	A-C	A-P	A-R	C-A	C-P	C-R	P-A	P-C	P-R	R-A	R-C	R-P	avg
DA	RESNET-50	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
	DANN (*)	44.3	59.8	69.8	48.0	58.3	63.0	49.7	42.7	70.6	64.0	51.7	78.3	58.3
	CDAN-E(*)	52.5	71.4	76.1	59.7	69.9	71.5	58.7	50.3	77.5	70.5	57.9	**83.5**	66.6
	DEEPJDOT (*)	50.7	68.6	74.4	59.9	65.8	68.1	55.2	46.3	73.8	66.0	54.9	78.3	63.5
	ALDA (*)	52.2	69.3	76.4	58.7	68.2	71.1	57.4	49.6	76.8	70.6	57.3	82.5	65.8
	ROT (*)	47.2	71.8	76.4	58.6	68.1	70.2	56.5	45.0	75.8	69.4	52.1	80.6	64.3
	JUMBOT	**55.2**	**75.5**	**80.8**	**65.5**	**74.4**	**74.9**	**65.2**	**52.7**	**79.2**	**73.0**	**59.9**	83.4	**70.0**
PDA	RESNET-50	46.3	67.5	75.9	59.1	59.9	62.7	58.2	41.8	74.9	67.4	48.2	74.2	61.4
	DEEPJDOT(*)	48.2	66.2	76.6	56.1	57.8	64.5	58.3	42.7	73.5	65.7	48.2	73.7	60.9
	PADA	51.9	67.0	78.7	52.2	53.8	59.0	52.6	43.2	78.8	73.7	56.6	77.1	62.1
	ETN	59.2	77.0	79.5	62.9	65.7	75.0	68.3	55.4	84.4	75.7	57.7	**84.5**	70.4
	BA3US(*)	56.7	76.0	**84.8**	73.9	67.8	**83.7**	72.7	56.5	84.9	77.8	64.5	83.8	73.6
	JUMBOT	**62.7**	**77.5**	84.4	**76.0**	**73.3**	80.5	**74.7**	**60.8**	**85.1**	**80.2**	**66.5**	83.9	**75.5**

Analysis: Ablation and sensitivity

Methods	U → M	S → M
DEEPJDOT	96.4 ± 0.3	95.4 ± 0.1
ENTROPIC DEEPJDOT	97.1 ± 0.3	97.6 ± 0.1
JUMBOT	$\mathbf{98.2 \pm 0.1}$	$\mathbf{98.9 \pm 0.1}$



Paper

Full details in the paper !

Check it out : https://arxiv.org/abs/2103.03606

References i

[Damodaran et al., 2018] Damodaran, B. B., Kellenberger, B., Flamary, R., Tuia, D., and Courty, N. (2018).
DeepJDOT: Deep Joint Distribution Optimal Transport for Unsupervised Domain Adaptation.
In *ECCV 2018 - 15th European Conference on Computer Vision*. Springer.

[Davis et al., 2020] Davis, D., Drusvyatskiy, D., Kakade, S., and Lee, J. D. (2020).
Stochastic subgradient method converges on tame functions.
Foundations of computational mathematics, 20(1):119–154.

[Fatras et al., 2020] Fatras, K., Zine, Y., Flamary, R., Gribonval, R., and Courty, N. (2020).
Learning with minibatch wasserstein: asymptotic and gradient properties.
In *AISTATS*.

[Fatras et al., 2021] Fatras, K., Zine, Y., Majewski, S., Flamary, R., Gribonval, R., and Courty, N. (2021).
Minibatch optimal transport distances; analysis and applications.

[Majewski et al., 2018] Majewski, S., Miasojedow, B., and Moulines, E. (2018).
Analysis of nonsmooth stochastic approximation: the differential inclusion approach.
arXiv preprint arXiv:1805.01916.