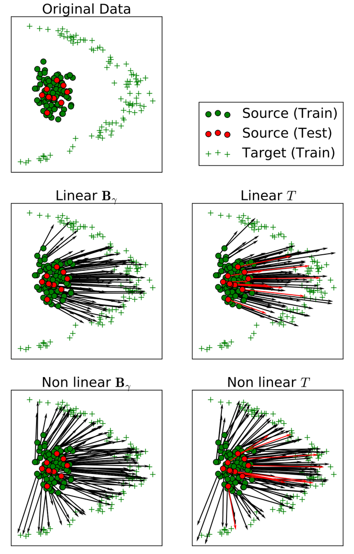


Mapping Estimation for Discrete Optimal Transport

Michaël Perrot[1], Nicolas Courty[2], Rémi Flamary[3], and Amaury Habrard[1]

[1]Univ Lyon, UJM-Saint-Etienne, CNRS, Lab. Hubert Curien UMR 5516, F-42023
{michael.perrot,amaury.habrard}@univ-st-etienne.fr

[2]Université de Bretagne Sud, IRISA, UMR 6074, CNRS,
courty@univ-ubs.fr

[3]Université Côte d'Azur, Lagrange, UMR 7293 , CNRS, OCA
remi.flamary@unice.fr

Objective: Jointly learn the coupling of a discrete optimal transport problem and a transformation approximating the corresponding transport map.

Optimal Transport

Setting

- Metric spaces: $\Omega_{\mathcal{S}} \in \mathbb{R}^{d_s}$ and $\Omega_{\mathcal{T}} \in \mathbb{R}^{d_t}$
- Probability measures: $\mu_{\mathcal{S}}$ on $\Omega_{\mathcal{S}}$ and $\mu_{\mathcal{T}}$ on $\Omega_{\mathcal{T}}$
- Training sets: $\mathbf{X}_s = \{\mathbf{x}_i^s\}_{i=1}^{n_s}$ with $\mathbf{x}^s \sim \mu_{\mathcal{S}}$ and $\mathbf{X}_t = \{\mathbf{x}_i^t\}_{i=1}^{n_t}$ with $\mathbf{x}^t \sim \mu_{\mathcal{T}}$
- Empirical distributions: $\hat{\mu}_{\mathcal{S}} = \sum_{i=1}^{n_s} p_i^s \delta_{\mathbf{x}_i^s}$ and $\hat{\mu}_{\mathcal{T}} = \sum_{i=1}^{n_t} p_i^t \delta_{\mathbf{x}_i^t}$
- Cost function: $c : \Omega_{\mathcal{S}} \times \Omega_{\mathcal{T}} \to [0, \infty[$, here $c(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_2^2$

Monge's problem Find a transport map such that:

$$T^* = \operatorname*{arg\,inf}_{T:\Omega_{\mathcal{S}} \to \Omega_{\mathcal{T}}} \left\{ \int_{\Omega_{\mathcal{S}}} c(\mathbf{x}, T(\mathbf{x})) d\mu_{\mathcal{S}}(\mathbf{x}), \ \ T\#\mu_{\mathcal{S}} = \mu_{\mathcal{T}} \right\}.$$

Kantorovich's relaxation Find a probabilistic coupling such that:

$$\gamma_0 = \operatorname*{arg\,min}_{\gamma \in \Pi} \int_{\Omega_{\mathcal{S}} \times \Omega_{\mathcal{T}}} c(\mathbf{x}^s, \mathbf{x}^t) d\gamma$$

where $\Pi = \{\gamma \,|\, \gamma_{\Omega_{\mathcal{S}}} = \mu_{\mathcal{S}}, \gamma_{\Omega_{\mathcal{T}}} = \mu_{\mathcal{T}}\}$. In the discrete case $\mathbf{C}$ is a cost matrix and $\hat{\Pi} = \left\{\gamma \in (\mathbb{R}^+)^{n_s \times n_t} \,|\, \gamma \mathbf{1}_{n_t} = \hat{\mu}_{\mathcal{S}}, \gamma^T \mathbf{1}_{n_s} = \hat{\mu}_{\mathcal{T}}\right\}$ such that:

$$\gamma = \operatorname*{arg\,min}_{\gamma \in \hat{\Pi}} \langle \gamma, \mathbf{C} \rangle_{\mathcal{F}} = \operatorname*{arg\,min}_{\gamma \in \hat{\Pi}} \operatorname{Tr}(\gamma^T \mathbf{C}).$$

Barycentric mapping Given a probabilistic coupling γ, we need to map the examples from $\Omega_{\mathcal{S}}$ to $\Omega_{\mathcal{T}}$:

$$\begin{aligned} \widehat{\mathbf{X}_s} = \mathbf{B}_\gamma(\mathbf{X}_s) &= \operatorname{diag}(\gamma \mathbf{1}_{n_t})^{-1} \gamma \mathbf{X}_t \\ &= n_s \gamma \mathbf{X}_t \text{ when the examples are i.i.d..} \end{aligned}$$

Drawbacks

- No out-of-sample projections, no explicit transport map: the coupling matrix has to be computed again for each new example.
- No control over the nature of the coupling: inducing specific properties on the transport map (*i.e.* regularity, divergence free, etc.) is hard.

Our solution We propose to jointly learn the coupling and an approximation of the corresponding transport map:

- Explicit transformation: there is no need to compute the transport map again to project new examples.
- Limited set of transformations: the nature of the transformations considered regularizes the transport.

Mapping Estimation

$$\operatorname*{arg\,min}_{T \in \mathcal{H}, \gamma \in \hat{\Pi}} f(\gamma, T) = \underbrace{\frac{1}{n_s d_t} \|T(\mathbf{X}_s) - \mathbf{B}_\gamma(\mathbf{X}_s)\|_{\mathcal{F}}^2}_{\text{Difference between Transformation and Coupling}} + \underbrace{\frac{\lambda_\gamma}{\max(\mathbf{C})} \langle \gamma, \mathbf{C} \rangle_{\mathcal{F}}}_{\text{Cost of the Coupling}} + \underbrace{\frac{\lambda_T}{d_s d_t} R(T)}_{\text{Regularization of the Transformation}}$$

Linear Transformation

$$\mathcal{H} = \left\{ T : \exists\, \mathbf{L} \in \mathbb{R}^{d^s \times d^t}, \forall \mathbf{x}^s \in \Omega_{\mathcal{S}}, T(\mathbf{x}^s) = \mathbf{x}^{sT} \mathbf{L} \right\}$$

Non-linear Transformation

$$\mathcal{H} = \left\{ T : \exists\, \mathbf{L} \in \mathbb{R}^{n^s \times d^t} \forall \mathbf{x}^s \in \Omega_{\mathcal{S}}, T(\mathbf{x}^s) = k_{\mathbf{X}_s}(\mathbf{x}^{sT}) \mathbf{L} \right\}$$

Optimisation

Algorithm 1: Joint Learning of $\mathbf{L}$ and γ.

input : $\mathbf{X}_s, \mathbf{X}_t$ source and target examples and $\lambda_\gamma, \lambda_T$ hyper parameters (tuned by circular validation).
output: $\mathbf{L}, \gamma$.

begin
  Initialize $k = 0$, $\gamma^0 \in \hat{\Pi}$ and $\mathbf{L}^0 = \mathbf{I}$
  repeat
    Learn γ^{k+1} with fixed $\mathbf{L}^k$ using a **Frank-Wolfe** approach.
    Learn $\mathbf{L}^{k+1}$ with fixed γ^{k+1} using a **closed form** approach.
    Set $k = k + 1$.
  until *convergence*

Theoretical Discussion

- Learned transformation: $T(\mathbf{x}^s)$
- True transport map: $T^*(\mathbf{x}^s)$
- Empirical barycentric mapping: $\mathbf{B}_\gamma(\mathbf{x}^s)$
- Theoretical barycentric mapping: $\mathbf{B}_{\gamma_0}(\mathbf{x}^s)$

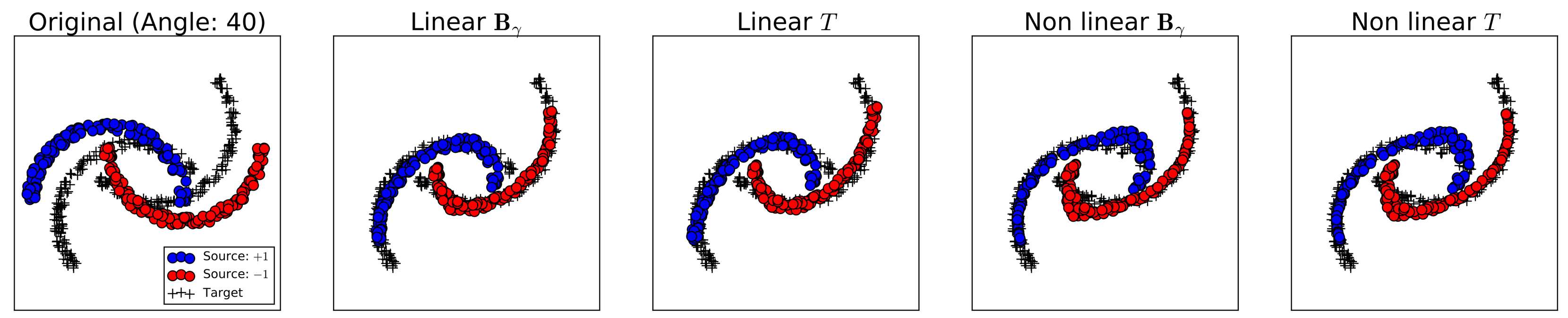
$$\begin{aligned} \mathop{\mathbb{E}}_{\mathbf{x}^s \sim \Omega_{\mathcal{S}}} \|T(\mathbf{x}^s) - T^*(\mathbf{x}^s)\|_{\mathcal{F}}^2 \le\ & 4 \sum_{\mathbf{x}^s \in \mathbf{X}_s} \|T(\mathbf{x}^s) - \mathbf{B}_\gamma(\mathbf{x}^s)\|_{\mathcal{F}}^2 + \mathcal{O}\left(\frac{1}{\sqrt{n_s}}\right) \\ & + 4 \sum_{\mathbf{x}^s \in \mathbf{X}_s} \|\mathbf{B}_\gamma(\mathbf{x}^s) - \mathbf{B}_{\gamma_0}(\mathbf{x}^s)\|_{\mathcal{F}}^2 \\ & + 2 \mathop{\mathbb{E}}_{\mathbf{x}^s \sim \Omega_{\mathcal{S}}} \|\mathbf{B}_{\gamma_0}(\mathbf{x}^s) - T^*(\mathbf{x}^s)\|_{\mathcal{F}}^2 . \end{aligned}$$

Links

POT: Python Optimal Transport: https://github.com/rflamary/POT
Poisson Blending with Adapted Gradients: https://github.com/ncourty/PoissonGradient

Experiments

Domain Adaptation: Moons Dataset



Domain Adaptation: Office-Caltech Dataset

Task	1NN	GFK	SA	L1L2	OTE	OTLin T	OTLin $\mathbf{B}_\gamma$	OTKer T	OTKer $\mathbf{B}_\gamma$
$D \to W$	89.47	93.31	95.56	95.7	95.7	97.28	97.28	98.41	**98.48**
$D \to A$	62.52	77.23	88.5	74.9	74.85	85.73	85.73	**89.92**	89.9
$D \to C$	51.81	69.73	**78.99**	67.85	68.03	77.15	77.15	69.1	69.17
$W \to D$	99.25	**99.75**	99.63	94.38	94.38	99.38	99.38	97.25	97.25
$W \to A$	62.5	72.38	79.25	71.33	71.35	**81.46**	**81.46**	78.5	78.35
$W \to C$	59.5	63.74	55.02	67.78	67.78	**75.87**	**75.87**	72.71	72.7
$A \to D$	65.25	75.88	**83.75**	70.13	70.5	80.63	80.63	65.63	65.5
$A \to W$	56.75	68.01	74.57	67.15	67.28	**74.64**	**74.64**	66.36	64.77
$A \to C$	70.09	75.71	79.2	74.06	74.31	81.81	81.81	84.38	**84.43**
$C \to D$	75.88	79.5	85.	69.75	70.25	**87.13**	**87.13**	70.13	70.
$C \to W$	65.17	70.66	74.44	63.77	63.77	78.28	78.28	80.	**80.4**
$C \to A$	85.79	87.13	89.33	76.63	76.67	**89.94**	**89.94**	82.38	82.15
Mean	70.33	77.75	81.94	74.45	74.57	**84.11**	**84.11**	79.56	79.43

Seamless Copy in Images: Poisson Blending [Perez 03] with Gradient Adaptation

Acknowledgments French ANR project LIVES ANR-15-CE23-0026-03.