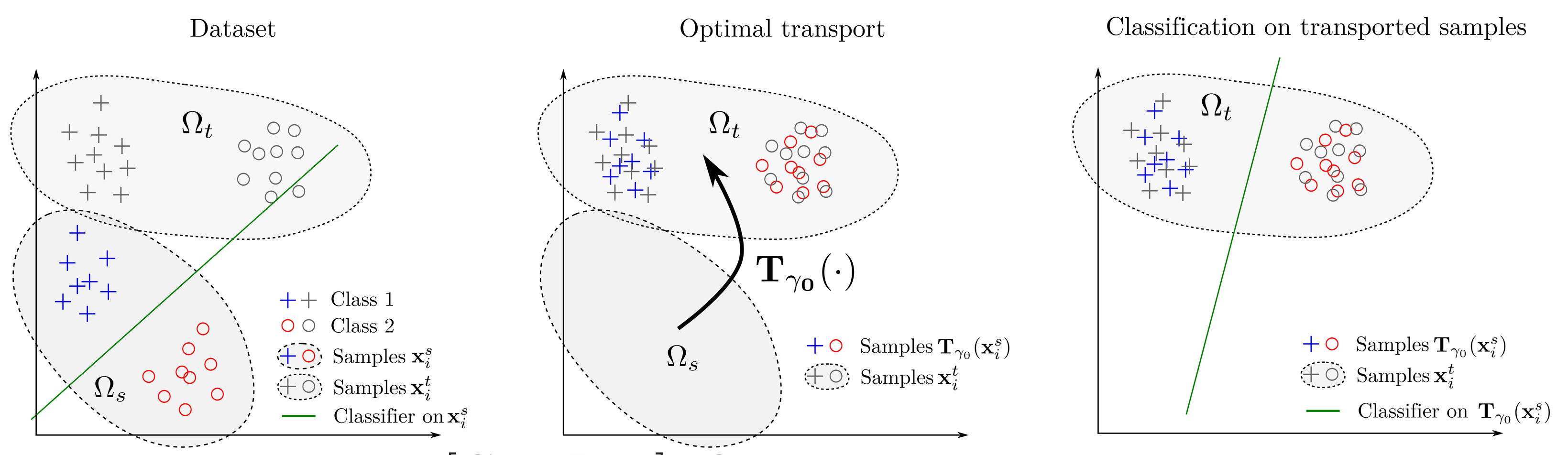
Joint Distribution Optimal Transportation for Domain Adaptation

Nicolas Courty[1,*], Rémi Flamary[2,*], Amaury Habrard[3] and Alain Rakotomamonjy[4]

[1]Université de Bretagne Sud, IRISA, UMR 6074, CNRS, courty@univ-ubs.fr
[2]Université Côte d'Azur Lagrange, UMR 7293 , CNRS, OCA remi.flamary@unice.fr
[3]Univ Lyon, UJM-Saint-Etienne, CNRS, Lab. Hubert Curien UMR 5516, F-42023 amaury.habrard@univ-st-etienne.fr
[4]Normandie Université, LITIS EA 4108 alain.rakoto@insa-rouen.fr

Objective: Solve unsupervised domain adaptation (DA) problems by aligning Joint Distributions as a discrete Optimal Transport (**JDOT**) problem.

OT and DA



In a previous work [CFTR16], OT was used to align one source and one target domain distributions under the following assumptions: *i)* There exist a transport in the feature space $\mathbf{T}$ between the two domains, *ii)*The transport preserves the conditional distributions

$$P_s(y|\mathbf{x}_s) = P_t(y|\mathbf{T}(\mathbf{x}_s)).$$

A 3-step strategy was then employed to solve for the problem:

1. Estimate optimal transport between distributions.
2. Transport the training samples with barycentric mapping .
3. Learn a classifier on the transported training samples.

Works very well in practice for a large class of transformation **but**:

- Model transformation only in the feature space
- Requires the same class proportion between domains
- We only search for a classifier $f : \mathbb{R}^d \to \mathbb{R}$, finding the mapping $\mathbf{T}$ as a first step is more complex and unnecessary.

Our solution We propose to transport Joint distributions (measures in the product space between input and label space)

Notations: Let $\Omega \in \mathbb{R}^d$ be a compact input measurable space of dimension d and $\mathcal{C}$ the set of labels.

- Let $\mathcal{P}_s(X,Y) \in \mathcal{P}(\Omega \times \mathcal{C})$ and $\mathcal{P}_t(X,Y) \in \mathcal{P}(\Omega \times \mathcal{C})$ the source and target joint distribution.
- We have access to an empirical sampling $\hat{\mathcal{P}}_s = \frac{1}{N_s}\sum_{i=1}^{N_s} \delta_{\mathbf{x}_i^s,\mathbf{y}_i^s}$ of the source distribution defined by $\mathbf{X}_s = \{\mathbf{x}_i^s\}_{i=1}^{N_s}$ and label information $\mathbf{Y}_s = \{\mathbf{y}_i^s\}_{i=1}^{N_s}$.
- yet the target domain is defined through its empirical distribution in the feature space with samples $\mathbf{X}_t = \{\mathbf{x}_i^t\}_{i=1}^{N_t}$ **without labels.**

{*}: equal contributions

Acknowledgments French ANR project OATMIL ANR-17-CE23-0012.

Work presented at NIPS'2017, December 4-9 2017, Long Beach, California.

Proxy distribution

Let f be a $\Omega \to \mathcal{C}$ function from a given class of hypothesis $\mathcal{H}$. We define the following joint distribution that uses f as a proxy of y

$$\mathcal{P}_t^f = (\mathbf{x}, f(\mathbf{x}))_{\mathbf{x}\sim\mu_t}$$

and its empirical counterpart $\hat{\mathcal{P}}_t^f = \frac{1}{N_t}\sum_{i=1}^{N_t} \delta_{\mathbf{x}_i^t, f(\mathbf{x}_i^t)}$.

Learning with JDOT

We propose to learn the predictor f that minimizes :

$$\min_f \left\{ W_1(\hat{\mathcal{P}}_s, \hat{\mathcal{P}}_t^f) = \inf_{\gamma\in\Delta} \sum_{ij} \mathcal{D}(\mathbf{x}_i^s, \mathbf{y}_i^s; \mathbf{x}_j^t, f(\mathbf{x}_j^t))\gamma_{ij} \right\}$$

- Δ is the transport polytope.
- $\mathcal{D}(\mathbf{x}_i^s, \mathbf{y}_i^s; \mathbf{x}_j^t, f(\mathbf{x}_j^t)) = \alpha\|\mathbf{x}_i^s - \mathbf{x}_j^t\|^2 + \mathcal{L}(\mathbf{y}_i^s, f(\mathbf{x}_j^t))$ with $\alpha > 0$.
- We search for the predictor f that best aligns the joint distributions.

Learning bound

We showcase a generalization bound for the unsupervised DA problem. It is based on a notion of probabilistic transfer Lipschitzness, which extends the original notion of Probabilistic Lipschitzness [BDSSU12].

Probabilistic Transfer Lipschitzness (PTL) Let μ_s and μ_t be respectively the source and target distributions. Let $\phi : \mathbb{R} \to [0,1]$. A labeling function $f : \Omega \to \mathbb{R}$ and a joint distribution $\Pi(\mu_s, \mu_t)$ over μ_s and μ_t are ϕ-Lipschitz transferable if for all $\lambda > 0$:

$$Pr_{(\mathbf{x}_1,\mathbf{x}_2)\sim\Pi(\mu_s,\mu_t)}\left[|f(\mathbf{x}_1) - f(\mathbf{x}_2)| > \lambda d(\mathbf{x}_1, \mathbf{x}_2)\right] \le \phi(\lambda).$$

Theorem Let $f^* \in \mathcal{H}$ be a Lipschitz labeling function that verifies the ϕ-PTL assumption w.r.t. Π^* and that minimizes the joint error $err_S(f^*) + err_T(f^*)$ w.r.t all PTL functions compatible with Π^*. For all $\lambda > 0$, with $\alpha = k\lambda$, we have with probability at least $1-\delta$ that:

$$\mathrm{err}_T(f) \le W_1(\hat{\mathcal{P}}_s, \hat{\mathcal{P}_t^f}) + \sqrt{\frac{2}{c'}\log(\frac{2}{\delta})}\left(\frac{\sqrt{N_S}+\sqrt{N_T}}{\sqrt{N_S N_T}}\right) + err_S(f^*) + err_T(f^*) + kM\phi(\lambda).$$

References

[BDSSU12] S. Ben-David, S. Shalev-Shwartz, and R. Urner. Domain adaptation–can quantity compensate for quality? In *Proc of ISAIM*, 2012.

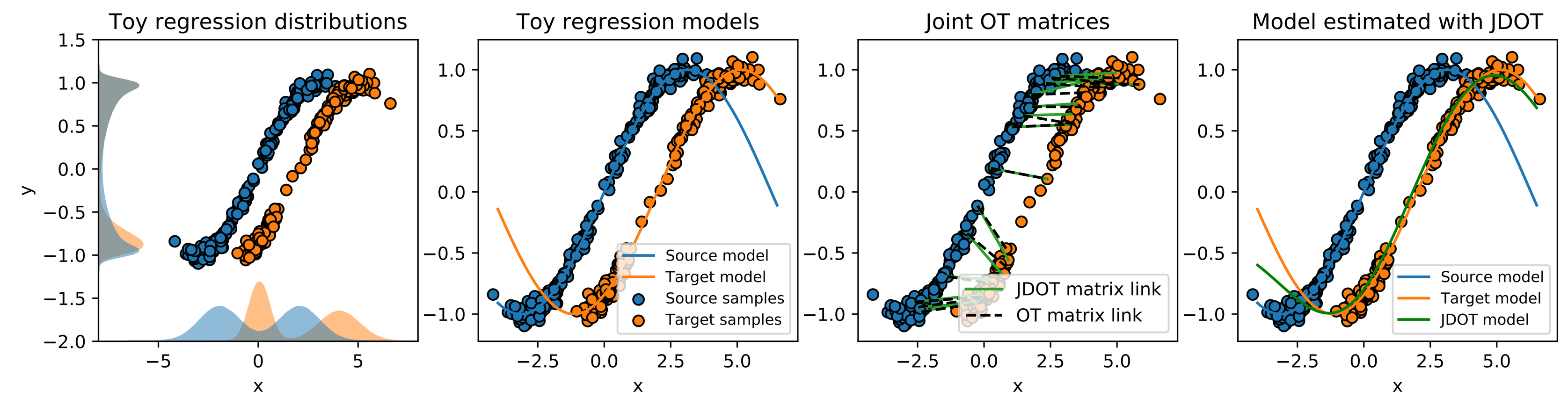
[CFTR16] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2016.

Optimization problem

$$\min_{f\in\mathcal{H},\gamma\in\Delta} \sum_{i,j} \gamma_{i,j}\left(\alpha d(\mathbf{x}_i^s, \mathbf{x}_j^t) + \mathcal{L}(y_i^s, f(\mathbf{x}_j^t))\right) + \lambda\Omega(f)$$

- $\Omega(f)$ is a regularization for the predictor f
- We propose to use block coordinate descent (BCD)/Gauss Seidel: solve alternatively for γ and f.
- Provably converges to a stationary point of the problem.

Regression with JDOT

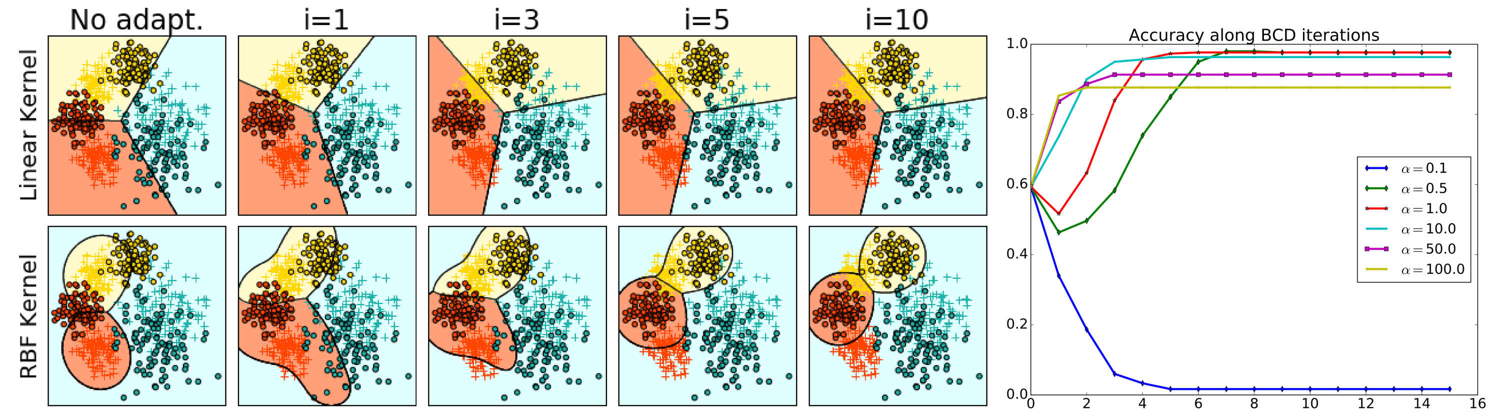


Least square regression with quadratic regularization For a fixed γ the optimization problem is equivalent to

$$\min_{f\in\mathcal{H}} \sum_j \frac{1}{n_t}\|\hat{y}_j - f(\mathbf{x}_j^t)\|^2 + \lambda\|f\|^2$$

where $\hat{y}_j = n_t\sum_j \gamma_{i,j} y_i^s$ is a weighted average of the source target values.

Classification with JDOT



Multiclass classification with Hinge loss For a fixed γ the optimization problem is equivalent to

$$\min_{f_k\in\mathcal{H}} \sum_{j,k} \hat{P}_{j,k}\mathcal{L}(1, f_k(\mathbf{x}_j^t)) + (1-\hat{P}_{j,k})\mathcal{L}(-1, f_k(\mathbf{x}_j^t)) + \lambda\sum_k \|f_k\|^2$$

where $\hat{\mathbf{P}}$ is the class proportion matrix $\hat{\mathbf{P}} = \frac{1}{N_t}\gamma^\top\mathbf{P}^s$. and $\mathbf{P}^s$ and $\mathbf{Y}^s$ are defined from the source data with One-vs-All strategy

Conclusion

- Powerful and versatile method, with applications to learn deep neural networks
- State-of-the art on several benchmarks, see paper for numerical results