Mapping conditional distributions for domain adaptation under generalized target shift

Matthieu Kirchmeyer[1,2], Alain Rakotomamonjy[2,3], Emmanuel de Bézenac[1], Patrick Gallinari[1,2]
[1]*Sorbonne Université, MLIA,* [2]*Criteo AI Lab,* [3]*Université de Rouen, LITIS* (Contact: `matthieu.kirchmeyer@isir.upmc.fr`)

Motivation

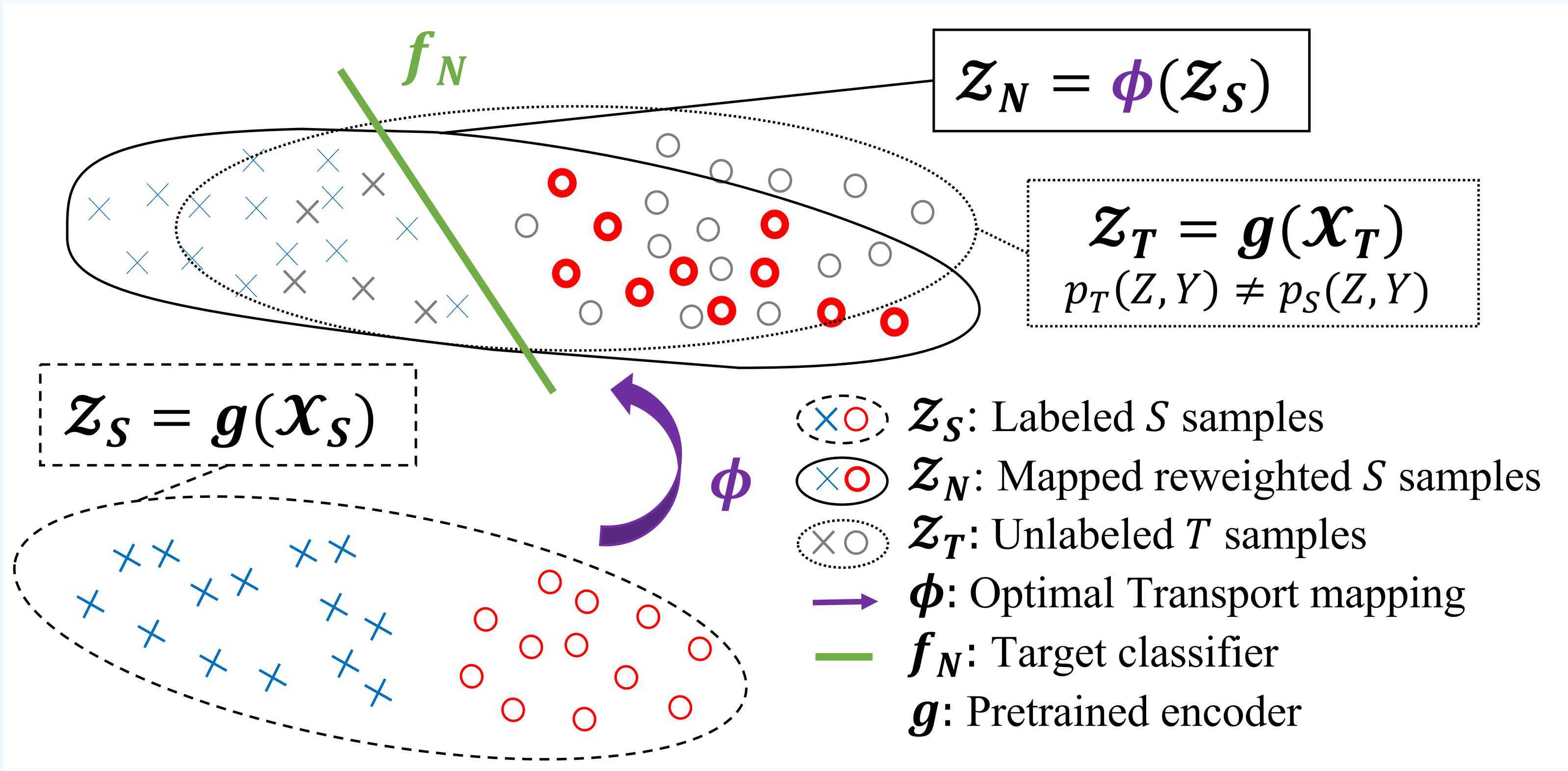
- Unsupervised Domain Adaptation with labelled domain S, **unlabelled** domain T s.t. $p_S(X,Y) \neq p_T(X,Y)$
- Generalized Target Shift (`GeTarS`) states conditional and label shift i.e. $\exists j, p_S(Z|Y=j) \neq p_T(Z|Y=j)$ and $\boldsymbol{p}_S^Y \neq \boldsymbol{p}_T^Y$.
- SOTA methods for `GeTarS` learn domain-invariant representations under class-reweighting [1, 2] but
  - Are prone to instabilities due to adversarial alignment.
  - May degrade target discriminativity [3].
  - Derive generalization guarantees under strong assumptions.

Contributions

- New approach, `OSTAR`, to align pretrained representations under `GeTarS` without constraining representation-invariance.
- Alignment with Optimal Transport implemented with a NN: native regularization, scalability and generalization beyond training samples.
- Strong theoretical guarantees under mild assumptions: 1) unicity of solution and 2) explicit control of the target risk with a new bound.
- Experimentally challenges the state-of-the-art for `GeTarS`.

OSTAR Framework

Components



1. Encode S samples ×○ and T samples ×○ with pretrained encoder g.
2. Define labelled domain N (samples ×○) by (i) mapping S samples with ϕ and (ii) reweighting by estimated class-ratios $\boldsymbol{p}_N^Y / \boldsymbol{p}_S^Y$.
3. Train a target classifier f_N on domain N for inference on T.

Optimal Transport alignment problem

$$\min_{\phi, \boldsymbol{p}_N^Y \in \Delta_K} \mathcal{C}(\phi) \triangleq \sum_{k=1 \cdots K} \int_{\mathbf{z} \in \mathcal{Z}} \|\phi(\mathbf{z}) - \mathbf{z}\|_2^2 \, p_S(\mathbf{z}|Y=k) d\mathbf{z}$$

$$\text{subject to } p_N^{\phi}(Z) \triangleq \sum_{k=1 \cdots K} \boldsymbol{p}_N^{Y=k} \phi_{\#}\Big(p_S(Z|Y=k)\Big) = p_T(Z) \qquad \text{(OT)}$$

Implementation

Joint Alignment and Classification objective

$$\min_{\phi, f_N} \mathcal{L}_{wd}^g(\phi, \boldsymbol{p}_N^Y) + \lambda_{OT} \, \mathcal{L}_{OT}^g(\phi) + \mathcal{L}_c^g(f_N, N)$$

$$\text{subject to } \boldsymbol{p}_N^Y = \operatorname*{arg\,min}_{\mathbf{p} \geq 0, \mathbf{p} \in \Delta_K} \frac{1}{2} \|\hat{\boldsymbol{p}}_T^Y - \hat{\mathbf{C}} \frac{\mathbf{p}}{\boldsymbol{p}_S^Y}\|_2^2 \ \text{(estimator in [4])} \qquad \text{(CAL)}$$

$$\mathcal{L}_{wd}^g(\phi, \boldsymbol{p}_N^Y) \triangleq \sup_{\|v\|_L \leq 1} \frac{1}{n} \sum_{i=1}^{n} \frac{\boldsymbol{p}_N^{y_S^{(i)}}}{\boldsymbol{p}_S^{y_S^{(i)}}} v \circ \phi(\mathbf{z}_\mathbf{S}^{(\mathbf{i})}) - \frac{1}{m} \sum_{j=1}^{m} v(\mathbf{z}_\mathbf{T}^{(\mathbf{j})})$$

$$\mathcal{L}_{OT}^g(\phi) \triangleq \sum_{k=1}^{K} \frac{1}{\#\{y_S^{(i)} = k\}_{i \in [\![1,n]\!]}} \sum_{y_S^{(i)} = k, i \in [\![1,n]\!]} \|\phi(\mathbf{z}_\mathbf{S}^{(\mathbf{i})}) - \mathbf{z}_\mathbf{S}^{(\mathbf{i})}\|_2^2$$

$$\mathcal{L}_c^g(f_N, N) \triangleq \frac{1}{n} \sum_{i=1}^{n} \frac{\boldsymbol{p}_N^{y_S^{(i)}}}{\boldsymbol{p}_S^{y_S^{(i)}}} \mathcal{L}_{ce}(f_N \circ \phi \circ g(\mathbf{x}_\mathbf{S}^{(\mathbf{i})}), y_S^{(i)})$$

Improve initial Target discriminativity

We update the encoder in a second stage. With δ softmax operation:

$$\mathcal{L}_{ent}^g(f_N, T) = \sum_{i=1}^{m} \sum_{k=1}^{K} \delta_k(f_N \circ g(\mathbf{x}_\mathbf{T}^{(\mathbf{i})})) \log(\delta_k(f_N \circ g(\mathbf{x}_\mathbf{T}^{(\mathbf{i})})))$$

$$\mathcal{L}_{div}^g(f_N, T) = D_{KL}(\hat{p}, \frac{\mathbf{1}_K}{K}) - \log K; \ \ \hat{p} = \mathbb{E}_{\mathbf{x_T} \in \mathcal{X}_T} \left[\delta\left(f_N \circ g(\mathbf{x_T})\right)\right]$$

$$\min_{f_N, g} \mathcal{L}_c^g(f_N, N) + \mathcal{L}_{ent}^g(f_N, T) + \mathcal{L}_{div}^g(f_N, T) + \mathcal{L}_c^g(f_S, S) \qquad \text{(SSg)}$$

Theoretical results

Theorem. *If* $\forall k, \boldsymbol{p}_N^{Y=k} > 0$, $\forall f_N \in \mathcal{H}$ M*-Lipschitz continuous over* $\mathcal{Z}$,

$$\epsilon_T^g(f_N) \leq \underbrace{\epsilon_N^g(f_N)}_{\text{Classification (C)}} + \frac{2M}{\min_{k=1}^K \boldsymbol{p}_N^{Y=k}} \underbrace{\mathcal{W}_1\Big(p_N(Z), p_T(Z)\Big)}_{\text{Alignment (A)}} +$$

$$2M\left(1 + \frac{1}{\min_{k=1}^K \boldsymbol{p}_N^{Y=k}}\right) \underbrace{\mathcal{W}_1\Big(\boldsymbol{p}_T(Z), \sum_{k=1}^{K} \boldsymbol{p}_N^{Y=k} p_T(Z|Y=k)\Big)}_{\text{Label (L)}}$$

Proposition. *If* $\mathcal{Z}$ *satisfies the following assumptions, there is a unique solution* $(\phi, \boldsymbol{p}_N^Y)$ *to (OT) and* $\phi_{\#}(p_S(Z|Y)) = p_T(Z|Y)$ *and* $\boldsymbol{p}_N^Y = \boldsymbol{p}_T^Y$.

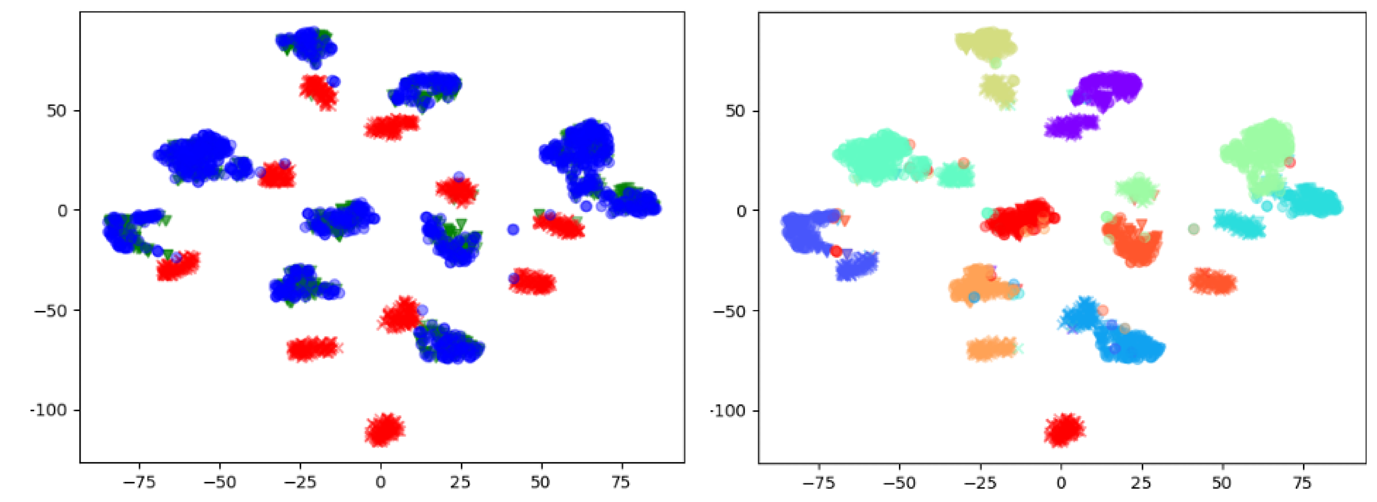
Assumption 1 (Clustering). $\forall k, \boldsymbol{p}_S^{Y=k} > 0$ *and there exists a partition* $\mathcal{Z}_S = \cup_{k=1}^K \mathcal{Z}_S^{(k)}$, *s.t.* $\forall k \, p_S(Z \in \mathcal{Z}_S^{(k)} | Y=k) = 1$.

Assumption 2 (Conditional matching). ϕ *solution to (OT) satisfies* $\forall k \, \exists j \, \phi_{\#}(p_S(Z|Y=k)) = p_T(Z|Y=j)$.

Assumption 3 (Cyclical monotonicity). $\forall \sigma$, $\sum_{k=1}^K \mathcal{W}_2(p_S(Z|Y=k), p_T(Z|Y=k)) \leq \sum_{k=1}^K \mathcal{W}_2(p_S(Z|Y=k), p_T(Z|Y=\sigma(k)))$.

Assumption 4 (Linear independence). $\{p_T(Z|Y=k)\}_{k=1}^K$ *linearly indep.*

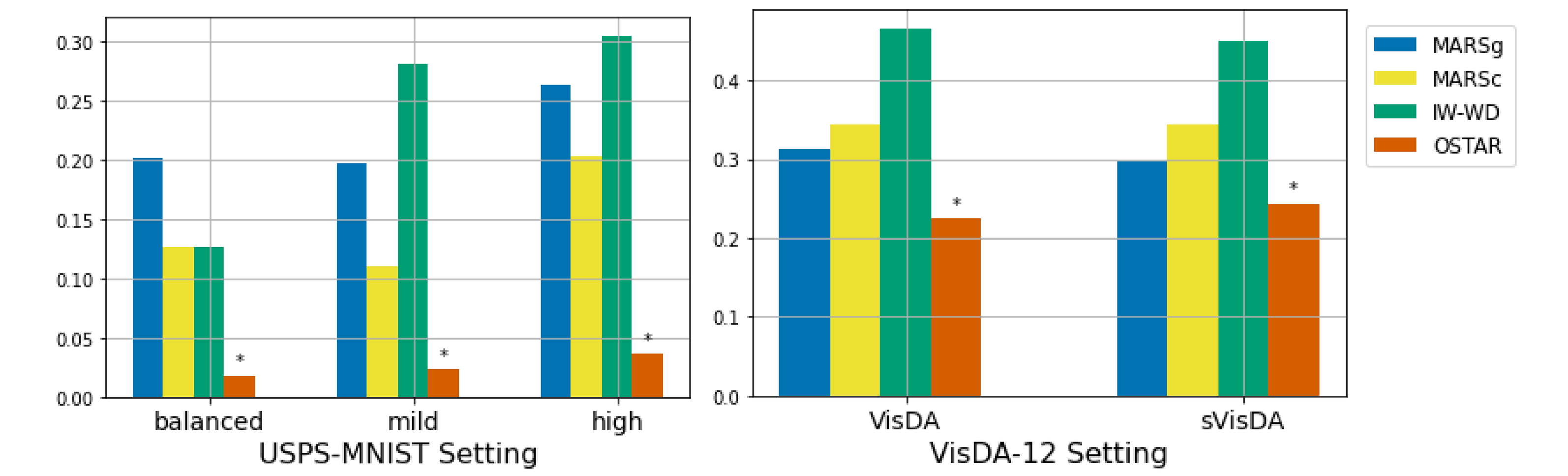
Results



On USPS→MNIST: S, T and N (left); classes in S and T (right).

Balanced Accuracy (↓) across visual UDA datasets with label shift

Setting	Source	DANN	$WD_{\beta=0}$	$WD_{\beta=1}$	MARSc	IW-WD	OSTAR+IM
Digits							
balanced	74.98 ± 3.8	90.81 ± 1.3	92.63 ± 1.0	82.80 ± 4.7	94.91 ± 1.4	95.89 ± 0.5	**97.51 ± 0.3**
subsampled	75.05 ± 3.1	89.91 ± 1.5	89.45 ± 1.0	81.56 ± 4.8	93.75 ± 1.4	93.22 ± 1.1	**96.69 ± 0.7**
VisDA12							
original	48.63 ± 1.0	53.72 ± 0.9	57.40 ± 1.1	47.56 ± 0.8	55.33 ± 0.8	51.88 ± 1.6	**59.24 ± 0.5**
subsampled	42.46 ± 1.4	47.57 ± 0.9	47.32 ± 1.4	41.48 ± 1.6	51.86 ± 2.0	50.65 ± 1.5	**58.84 ± 1.0**
Office31							
subsampled	74.50 ± 0.5	76.13 ± 0.3	76.24 ± 0.3	74.23 ± 0.5	80.00 ± 0.5	77.28 ± 0.4	**82.61 ± 0.4**
OfficeHome							
subsampled	50.56 ± 2.8	50.87 ± 1.05	53.47 ± 0.7	52.24 ± 1.1	56.22 ± 0.6	54.87 ± 0.4	**59.51 ± 0.4**



Estimation error $\|\boldsymbol{p}_N^Y - \boldsymbol{p}_T^Y\|_1$ (↓). "⋆": best model for balanced accuracy.

Domain-invariant baselines designed for:

- Covariate shift without reweighting (`DANN` [5], $WD_{\beta=0}$ [6]).
- `GeTarS` with reweighting ($WD_{\beta \in \{1,2\}}$ [7], `MARS` [2]; `IW-WD` [1]).

References

[1] Tachet des Combes et al. Domain adaptation with conditional distribution matching and generalized label shift. In *NeurIPS*, 2020.

[2] Rakotomamonjy et al. Optimal transport for conditional domain matching and label shift. *Machine Learning*, 2021.

[3] Liu et al. Transferable adversarial training: A general approach to adapting deep classifiers. In *ICML*, 2019.

[4] Lipton et al. Detecting and correcting for label shift with black box predictors. In *ICML*, 2018.

[5] Ganin et al. Domain-adversarial training of neural networks. *JMLR*, 2016.

[6] Shen et al. Wasserstein distance guided representation learning for domain adaptation. In *AAAI*, 2018.

[7] Wu et al. Domain adaptation with asymmetrically-relaxed distribution alignment. In *ICML*, 2019.