Incorporating Importance Weighting in Optimal Transport Based Domain Alignment

Okan Koç[1], Alexander Soen[1,2], Shanglin Li[3] and Masashi Sugiyama[1,4]

[1]RIKEN AIP, [2]KTH, [3]BIFOLD, [4]The University of Tokyo

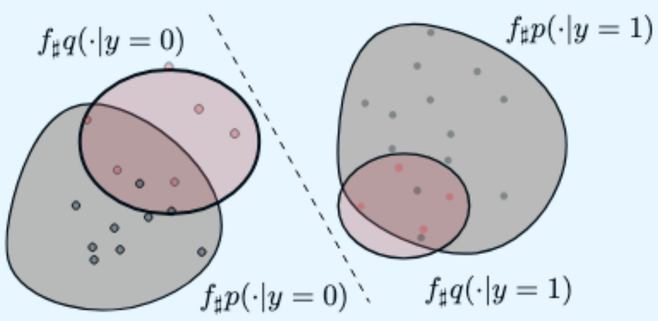
GRADUATE SCHOOL OF FRONTIER SCIENCES
THE UNIVERSITY OF TOKYO

RIKEN AIP

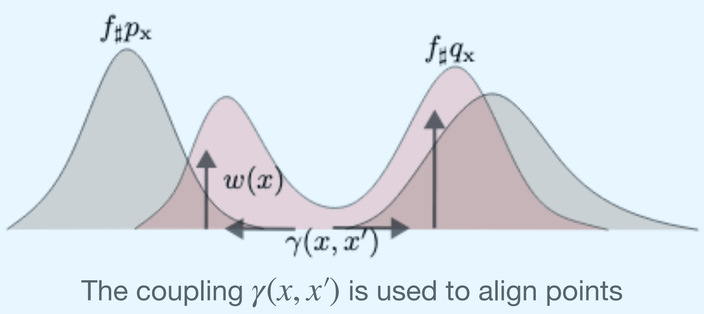
Paper

Our contribution in a nutshell

TL; DR: We introduce importance weighting to optimal transport based domain adaptation bounds and show the connection to unbalanced optimal transport.



A case where there is significant overlap between the source and target output distributions



The coupling $\gamma(x,x')$ is used to align points *horizontally*, while weighting the source $f_\sharp p_x$ can be seen as a *vertical* alignment.

- During UDA, if the source and target feature densities (where the alignment takes place) share some support, then the UDA bounds can be tightened by looking for the best weighted source distribution to compare against.
- This implies that the model capacity need not be unnecessarily constrained to match the two domains exactly: only the out-of-support part of the target distribution need to be aligned with the closest weighted source distribution.
- We use this idea to propose a more relaxed OT based alignment strategy and highlight the link to unbalanced OT (UOT).
- We also explore assumptions under which such alignment is guaranteed to produce low target risk classifiers.

Background and Related Work

Definitions and concepts needed: W-distance and Unbalanced OT (UOT)

- The α-Wasserstein distance w.r.t. a metric c

$$W_{\alpha,c}(\mu,\nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)^\alpha \, d\gamma(x,y) \right)^{1/\alpha},$$
$$\Gamma(\mu,\nu) = \{\gamma \in \triangle(\mathcal{X}\times\mathcal{Y}) : \pi_1\#\gamma = \mu, \pi_2\#\gamma = \nu\}$$

- $$\mathrm{UOT}(\mu,\nu) := \inf_{\gamma\in\triangle(\mathcal{X}\times\mathcal{Y})} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\gamma(x,y) + \tau_1 \mathrm{KL}(\pi_{1\sharp}\gamma \parallel \mu) + \tau_2 \mathrm{KL}(\pi_{2\sharp}\gamma \parallel \nu)$$
- In the limit as $\tau_1 \to 0$ and $\tau_2 \to \infty$ (or vice versa), we get a semi-relaxed OT, where one of the OT constraints drops entirely.

OT based UDA bounds and Entanglement

- Suppose the loss function $\ell : \triangle(\mathcal{Y}) \times \triangle(\mathcal{Y}) \to \mathbb{R}_+$ is a metric and that the classifier $f : \mathcal{X} \to \triangle(\mathcal{Y})$ is surjective, then

$$\epsilon_q(f) \le \epsilon_p(f) + W_{1,l}(f\#p_x, f\#q_x) + \mathscr{E}_y(f)$$

→ Useful for analyzing UDA optimization

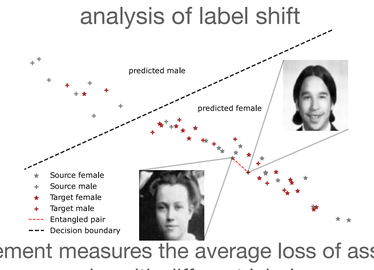
or

$$\epsilon_q(f) \le \epsilon_p(f) + W_{1,l}(p_y, q_y) + \mathscr{E}_{\hat{y}}(f)$$

→ Useful for further analysis, e.g., for analysis of label shift

- Label entanglement quantifies the assignment of wrong labels during optimal transport

$$\mathscr{E}_y(f) := \int_{\mathcal{Y}\times\mathcal{Y}} W_{1,l}(p_{y|f}(\cdot \mid \hat{y}), q_{y|f}(\cdot \mid \hat{y}')d\gamma_f^\star(\hat{y},\hat{y}')$$



Entanglement measures the average loss of associating pairs with different labels.

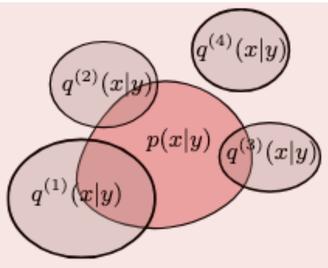
- UDA optimizer can optimize only the first two terms: Wasserstein-regularized risk (WRR) $R_\lambda(f) := \epsilon_p(f) + \lambda W_{1,\ell}(f_\sharp p_x, f_\sharp q_x)$
- Label entanglement depends on both (f,γ) optimized by WRR!
- **Lower bound.** Minimizing WRR can actually inflate the target error!
$$W_{1,\ell}(p_y,q_y) - \mathscr{R}_2(f) \le \mathscr{E}(f) - \mathscr{R}_1(f) \le \epsilon_q(f) \le \mathscr{E}(f) + \mathscr{R}_1(f).$$
- Entanglement can be high even when WRR is low! How can we guarantee that the entanglement is low as we minimize WRR?

Assumptions

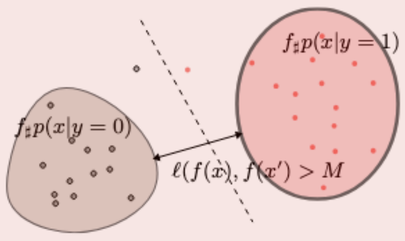
Required to tame entanglement

- The GS assumption $GS(a,b,s,\varepsilon)$ holds for joint densities p and q and a classifier f if there exists $q^{(i)}_{x|y}(\cdot \mid y) \in \triangle(\mathcal{X})$ and $a \le r_i \le b$ such that for all $y \in \mathcal{Y}$, $q_{x|y}(x \mid y) = \sum_{i=1}^{s} r_i q^{(i)}_{x|y}(x \mid y)$, where $(i-1)\varepsilon \le \ell(\hat{y},\hat{y}') \le i\varepsilon$ for $\hat{y} \sim (f_\sharp p_{x|y})(\cdot \mid y), \hat{y}' \sim (f_\sharp q^{(i)}_{x|y})(\cdot \mid y)$ almost surely.
- The Probabilistic Margin assumption $PM(M,\eta)$ holds for a joint distribution p and a classifier f if for all $y \ne y'$,

$$Pr_{\hat{y}\sim(f_\sharp p_{x|y})(\cdot|y)}\left(\inf_{\hat{y}'\in supp_f(p_{x|y'})} \ell(\hat{y},\hat{y}') \ge M \right) \ge 1-\eta.$$



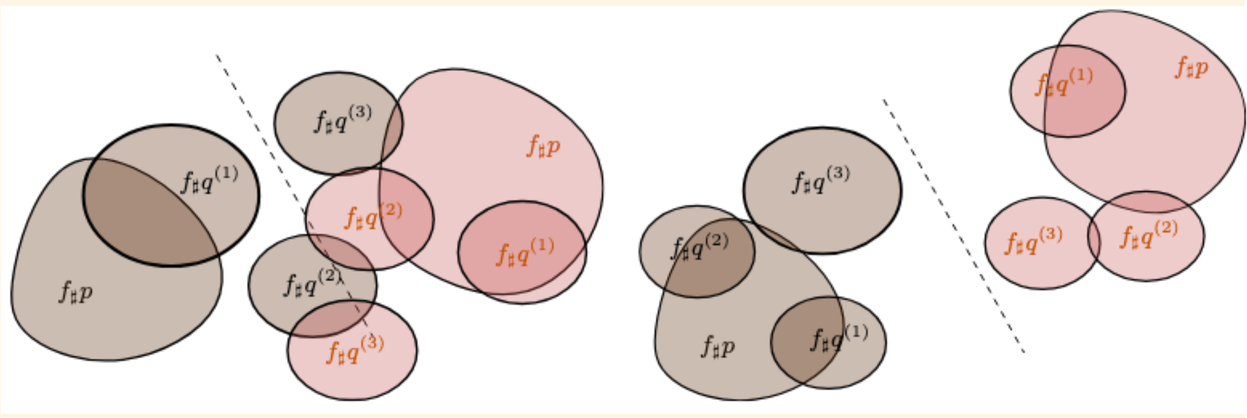
An example illustrating gradual shift in the target distribution q.



An example illustrating the PM assumption: If $f_\sharp p_x$ satisfies the $PM(M,\eta)$ with prob. $1-\eta$, then most samples will be separated with margin at least M.

Theoretical analysis

- Given the probabilistic margin $PM(M,\eta)$ of the classifier and gradual shift $GS(a,b,s,\varepsilon)$ in the output space, we show that low WRR implies low target risk. (GS and PM assumptions help to explain the success of UDA strategies that are effective at optimizing their unsupervised objectives without requiring target labels.)
- Thm 3.1. Assume $GS(a,b,s,\varepsilon)$ and $PM(M,\eta)$ hold for densities p and q and a surjective classifier f; and loss ℓ is a metric. Then if $1 - |\mathcal{Y}|\eta > 0$ and there exists a $d \in \mathbb{Z}_+$ such that $\varepsilon \le \frac{M}{2d}$, then $\epsilon_q(f) \le \rho \mathscr{R}_\lambda(f)$, where $\lambda = \frac{s(s+1)}{d(d-1)} \frac{b}{\rho a(1-|\mathcal{Y}|\eta)}$ and the maximum label ratio is $\rho := \max_{y\in\mathcal{Y}} q_y(y)/p_y(y)$.



Given enough gradual shift, the accuracy of a good source classifier even with large margin can drop significantly.

If the UDA performance is good (i.e., WRR value is low) then necessarily the conditionals will be aligned (i.e., entanglement will be low) and the target accuracy will be high.

- **Proof sketch: the margin assumption forces any shifting conditionals with significant mass to inflate the marginal distance for small enough** $\varepsilon < M/2$. Note that the shifting conditionals do not need to be *connected* to each other: the assumption is satisfied whenever the distances to the source conditional are ordered by ε.

Relaxing the alignment: unbalanced OT

Equivalently, tightening the WRR upper bound

- The weighted WRR minimization is defined as $\min_{f,w} \mathscr{L}_\lambda(f,w) := \epsilon_p^w(f) + \lambda W_{1,\ell}(f_\sharp p_x^w, f_\sharp q_x)$,
$\epsilon_p^w(f) := \mathbb{E}_{(x,y)\sim p}[w(x)\ell(f(x),y)]$
- Cor 4.1. Assume $GS(a,b,s,\varepsilon^w)$ and $PM(M^w,\eta^w)$ hold for distributions p^w, q, a surjective classifier f and normalized importance weights w; loss ℓ is a metric. If $\varepsilon^w \le M^w/(2d)$ for some $d \in \mathbb{Z}_+$, we have that $W_{1,\ell}(p_y^w, q_y) - \mathscr{L}_2(f,w) \le \epsilon_q(f) \le \rho^w \mathscr{L}_{\lambda^w}(f,w)$, for $\lambda^w = \frac{s(s+1)}{d(d-1)} \frac{b}{a(1-|\mathcal{Y}|\eta^w)\rho^w}$
(Notice: GS and PM now also depend on the IW w!)
- We show that the inner optimization of WRR w.r.t. w corresponds to a semi-relaxed OT problem. Expanding $W_{1,\ell}(f_\sharp p_x^w, f_\sharp q_x)$ and using the marginal constraint $\pi_{1\sharp}\gamma_f^w = f_\sharp p_x^w$, WRR in the discrete case can be written as
- $\min_{\Gamma\in\mathbb{R}_+^{m\times n}} \min_{w\in\mathbb{R}_+^n} \ell^{\mathrm{T}} w + \lambda \mathrm{tr}(C\Gamma^{\mathrm{T}})$ such that $\Gamma 1_n = w$ and $\Gamma^{\mathrm{T}} 1_m = \frac{1}{n} 1_n$
* Loss vector $\ell_i := \ell(f(x_i), y_i)$ and cost matrix elements $C_{ij} = \ell(f(x_i), f(x_j'))$
- Plugging in $\Gamma 1_n = w$ to the objective, we get an LP without w
- $\mathscr{L}_\lambda(f) = \min_{\Gamma\in\mathbb{R}_+^{m\times n}} \mathrm{tr}(\tilde{C}^\lambda \Gamma^{\mathrm{T}})$ such that $\Gamma^{\mathrm{T}} 1_m = \frac{1}{n} 1_n$: A semi-relaxed OT with costs $\tilde{C}^\lambda = \ell 1_n^{\mathrm{T}} + \lambda C$!
- Has an explicit solution, using only the nearest neighbors during alignment!

$$\mathscr{L}_\lambda^\star(f) := \frac{1}{n} \sum_{j=1}^{n} \min_{i\in[m]} \{\ell(f(x_i),y_i) + \lambda \ell(f(x_i),f(x_j'))\}$$

(The alignment chooses shortest dist. pairings $\ell(f(x_i),f(x_j'))$ and ignores the large majority of other source-target pairings.)

Experiments

Comparison with UDA methods

- Tested over three different models: MLP, ConvNet and ResNet18
- Same hyperparameters used throughout
- Using pre-training for five epochs
- Using margin loss to improve margin during pretraining
- W^2R^2 improves over WRR consistently

Scenario	Model	Oracles LJE	Oracles CC	Baseline ERM	UDA Algorithms DANN	rev-KL	WRR	W^2R^2
MNIST → USPS	MLP	0.95 ± 0.0020	0.89 ± 0.0040	0.33 ± 0.0085	0.55 ± 0.0540	0.62 ± 0.0110	0.77 ± 0.0010	0.86 ± 0.0053
	ConvNet	0.97 ± 0.0000	0.97 ± 0.0027	0.85 ± 0.0017	0.89 ± 0.0053	0.91 ± 0.0015	0.95 ± 0.0017	0.94 ± 0.0033
	ResNet	0.97 ± 0.0026	0.90 ± 0.0330	0.88 ± 0.0062	0.88 ± 0.0037	0.88 ± 0.0150	0.88 ± 0.0370	0.95 ± 0.0015
USPS → MNIST	MLP	0.95 ± 0.0013	0.91 ± 0.0013	0.30 ± 0.0018	0.42 ± 0.0110	0.51 ± 0.0120	0.57 ± 0.0072	0.62 ± 0.0140
	ConvNet	0.98 ± 0.0014	0.97 ± 0.0031	0.72 ± 0.0150	0.84 ± 0.0088	0.85 ± 0.0019	0.91 ± 0.0052	0.95 ± 0.0064
	ResNet	0.98 ± 0.0000	0.95 ± 0.0025	0.73 ± 0.0077	0.74 ± 0.0310	0.67 ± 0.1100	0.90 ± 0.0066	0.93 ± 0.0070
MNIST → MNISTM	ConvNet	0.93 ± 0.0024	0.90 ± 0.0014	0.48 ± 0.0140	0.68 ± 0.0081	0.53 ± 0.0064	0.61 ± 0.0071	0.49 ± 0.0130
	ResNet	0.89 ± 0.0017	0.84 ± 0.0015	0.32 ± 0.0140	0.42 ± 0.0640	0.10 ± 0.0070	0.50 ± 0.0200	0.66 ± 0.0100
SVHN → MNIST	ResNet	0.98 ± 0.0005	0.96 ± 0.0046	0.76 ± 0.0120	0.55 ± 0.1200	0.57 ± 0.1100	0.78 ± 0.0160	0.73 ± 0.0073

Reducing the variance of semi-relaxed OT, we get the unbalanced OT as a regularized version of the relaxed alignment. In practice we use

$$\frac{1}{m}\ell^{\mathrm{T}} 1_m + \min_{\Gamma\in\mathbb{R}_+^{m\times n}} \mathrm{tr}\left(\tilde{C}^1\Gamma^{\mathrm{T}}\right) + \mathrm{KL}(\Gamma 1_n \parallel \tfrac{1_m}{m}) + 100\mathrm{KL}(\Gamma^{\mathrm{T}} 1_m \parallel \tfrac{1_n}{n})$$