Joint distribution optimal transportation for domain adaptation

Changhuang Wan

Mechanical and Aerospace Engineering Department

The Ohio State University

March 8th , 2018

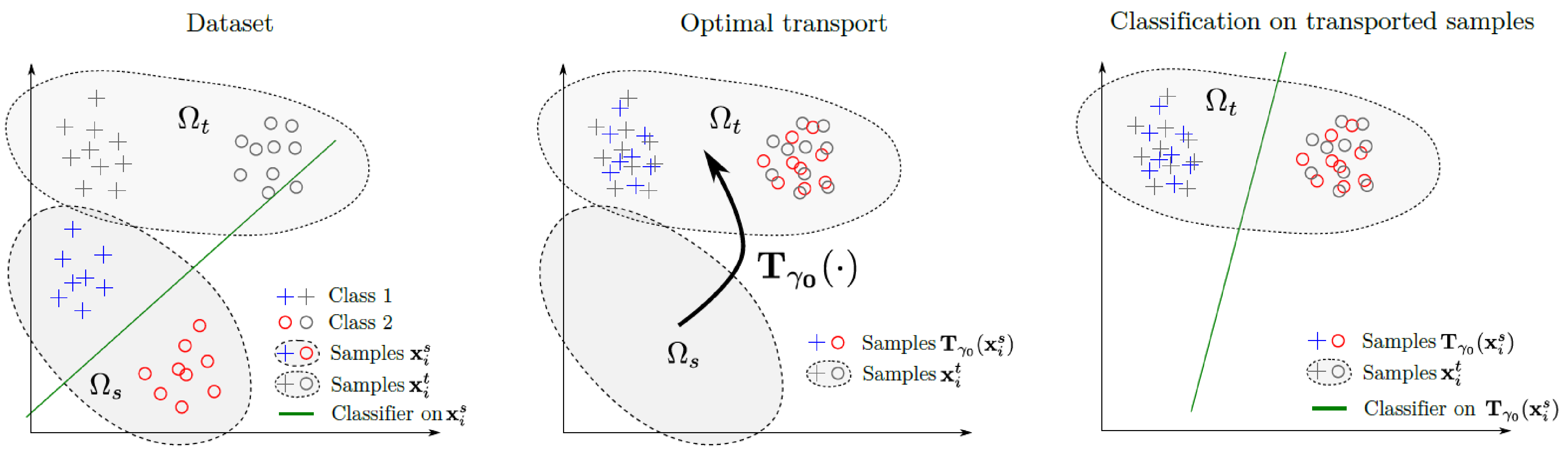
OUTLINE

- Problem Statement
- Assumption and Notations
- Joint Distribution Optimal Transport
- Bound on the Target Error
- Learning with Joint Distribution OT
- Examples

OUTLINE

- ❖ Problem Statement
- ❖ Assumption and Notations
- ❖ Joint Distribution Optimal Transport
- ❖ Bound on the Target Error
- ❖ Learning with Joint Distribution OT
- ❖ Examples

Problem Statement



In DA problem, we study two different (but related) distributions D_S and D_T on $X \times Y$. The DA task consists of the transfer of knowledge from the D_S to D_T. The objective is to learn f (from labeled or unlabeled samples of two domains) such that it commits as small error as possible on the target domain D_T.

OUTLINE

Assumption and Notations

Assumption: there exists a nonlinear transformation between the label space distributions of the two domain P_S and P_T that can be estimated with optimal transport.

Notations:

$X_S = \{x_i^s\}_{i=1}^{N_s}$	A set of data from sample domain
$X_T = \{x_i^t\}_{i=1}^{N_t}$	A set of data from target domain
$Y_S = \{y_i^s\}_{i=1}^{N_s}$	A set of class label information associated with *Xs*
$Y_T = \{y_i^t\}_{i=1}^{N_t}$	A set of class label information associated with X_T
$\Omega \in \mathbb{R}^d$	Compact input measureable space with dimension *d*
$C \in \mathbb{R}^1$	Label space
$P(\Omega)$	All probability over Ω
$P_s(X,Y)$	Joint probability distributions in D_S
$P_T(X,Y)$	Joint probability distributions in D_T

OUTLINE

- ❖ Problem Statement
- ❖ Assumption and Notations
- ❖ Joint Distribution Optimal Transport
- ❖ Bound on the Target Error
- ❖ Learning with Joint Distribution OT
- ❖ Examples

Joint Distribution Optimal Transport

Optimal transport in domain adaptation

Seek for a transport plan (or equivalently a joint probability distribution) $\gamma \in P(\Omega \times \Omega)$ such that:

$$\gamma_0 = \underset{\gamma \in \Pi(\mu_s, \mu_t)}{\operatorname{argmin}} \int_{\Omega \times \Omega} d(\mathbf{x}_1, \mathbf{x}_2) d\gamma(\mathbf{x}_1, \mathbf{x}_2),$$

where $\Pi(\mu_s, \mu_t) = \left\{ \gamma \in P(\Omega \times \Omega) \middle| p^+ \# \gamma = \mu_s, p^- \# \gamma = \mu_t \right\}$ and p^+ and p^-denotes the two marginal projections of $\Omega \times \Omega$ to Ω, and $p \# \gamma$ the image measure of γ by p.

Joint distribution optimal transport loss in DA

To handle a change in both marginal and conditional distributions.

$$\gamma_0 = \underset{\gamma \in \Pi(\mathcal{P}_s, \mathcal{P}_t)}{\operatorname{argmin}} \int_{(\Omega \times \mathcal{C})^2} \mathcal{D}(\mathbf{x}_1, y_1; \mathbf{x}_2, y_2) d\gamma(\mathbf{x}_1, y_1; \mathbf{x}_2, y_2),$$

where $D(\mathrm{x}_1, y_1; \mathrm{x}_2, y_2) = \alpha \, \mathrm{d}(\mathrm{x}_1, \mathrm{x}_2) + \mathcal{L}(y_1, y_2)$ is a joint cost measure combining both distance and a loss function measuring the discrepancy between y_1 and y_2

Joint Distribution Optimal Transport

Joint distribution optimal transport loss in DA

To handle a change in both marginal and conditional distributions.

$$\gamma_0 = \underset{\gamma \in \Pi(\mathcal{P}_s, \mathcal{P}_t)}{\operatorname{argmin}} \int_{(\Omega \times \mathcal{C})^2} \mathcal{D}(\mathbf{x}_1, y_1; \mathbf{x}_2, y_2) d\gamma(\mathbf{x}_1, y_1; \mathbf{x}_2, y_2),$$

In the **unsupervised DA** problem, one does **not** have access to labels in the target domain, and as such it is not possible to find the optimal coupling. Since our goal is to find a function on the target domain $f : \Omega \to \mathcal{C}$

Define the following joint distribution that uses a given function f as a proxy for y in target domain:

$$P_t^f = \left(\mathrm{x}, f\left(\mathrm{x}\right)\right)_{\mathrm{x} \sim \mu_t}$$

In practice we consider empirical versions of P_s and P_t^f , i.e.

$$\hat{P}_s = \frac{1}{N_s} \sum_{i=1}^{N_s} \delta_{x_i^s, y_i^s}, \hat{P}_t^f = \frac{1}{N_t} \sum_{i=1}^{N_s} \delta_{x_i^t, f\left(x_i^t\right)}$$

Joint Distribution Optimal Transport

Joint distribution optimal transport loss in DA

to handle a change in both marginal and conditional distributions.

$$\gamma_0 = \underset{\gamma \in \Pi(\mathcal{P}_s, \mathcal{P}_t)}{\operatorname{argmin}} \int_{(\Omega \times \mathcal{C})^2} \mathcal{D}(\mathbf{x}_1, y_1; \mathbf{x}_2, y_2) d\gamma(\mathbf{x}_1, y_1; \mathbf{x}_2, y_2),$$

$$f : \Omega \to \mathcal{C}$$

$$P_t^f = \left(\mathrm{x}, f(\mathrm{x})\right)_{\mathrm{x} \sim \mu_t}$$

$$\hat{P}_s = \frac{1}{N_s} \sum_{i=1}^{N_s} \delta_{x_i^s, y_i^s}, \hat{P}_t^f = \frac{1}{N_s} \sum_{i=1}^{N_s} \delta_{x_i^t, f(x_i^t)}$$

JDOT:

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

where W_1 is the 1-Wasserstein distance for the loss *D*.

Remark: The function *f* we retrieve is theoretically bound with respect to the target error.

OUTLINE

A Bound on the Target Error

Define the expected loss in the target domain $err_T(f)$

$$err_T(f) \triangleq \mathbf{E}_{(\mathrm{x},y)\sim P_t} \mathcal{L}(y, f(\mathrm{x}))$$

Similarly,

$$err_S(f) \triangleq \mathbf{E}_{(\mathrm{x},y)\sim P_s} \mathcal{L}(y, f(\mathrm{x}))$$

Assume the loss function $\mathcal{L}$ to be bounded, symmetric, *k*-Lipschitz and satisfying the triangle inequality.

Symmetric: $\mathcal{L}(y_1, y_2) = \mathcal{L}(y_2, y_1),\ y_1, y_2 \in \mathcal{C}$

k-Lipschitz: there exists k such that $\left|\mathcal{L}(y_1, y_2) - \mathcal{L}(y_1, y_3)\right| \le k\left|y_2 - y_3\right|,\ y_1, y_2, y_3 \in \mathcal{C}$

Triangle inequality $\mathcal{L}(y_1, y_3) \le \mathcal{L}(y_1, y_2) + \mathcal{L}(y_1, y_3),\ y_1, y_2, y_3 \in \mathcal{C}$

A Bound on the Target Error

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

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

Note: Given a deterministic labeling functions *f* and a coupling Π , it bounds the probability of finding pairs of source-target instances labelled differently in a $(1/\lambda)$-ball with respect to Π .

A Bound on the Target Error

Theorem 3.1 *Let f be any labeling function of $\in \mathcal{H}$. Let $\Pi^* = \operatorname{argmin}_{\Pi \in \Pi(\mathcal{P}_s, \mathcal{P}_t^f)} \int_{(\Omega \times \mathcal{C})^2} \alpha d(\mathbf{x}_s, \mathbf{x}_t) + \mathcal{L}(y_s, y_t) d\Pi(\mathbf{x}_s, y_s; \mathbf{x}_t, y_t)$ and $W_1(\hat{\mathcal{P}}_s, \hat{\mathcal{P}}_t^f)$ the associated 1-Wasserstein distance. Let $f^* \in \mathcal{H}$ be a Lipschitz labeling function that verifies the ϕ-probabilistic transfer Lipschitzness (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^*. We assume the input instances are bounded s.t. $|f^*(\mathbf{x}_1) - f^*(\mathbf{x}_2)| \leq M$ for all $\mathbf{x}_1, \mathbf{x}_2$. Let $\mathcal{L}$ be any symmetric loss function, k-Lipschitz and satisfying the triangle inequality. Consider a sample of N_s labeled source instances drawn from $\mathcal{P}_s$ and N_t unlabeled instances drawn from μ_t, and then for all $\lambda > 0$, with $\alpha = k\lambda$, we have with probability at least $1 - \delta$ that:*

$$err_T(f) \leq W_1(\hat{\mathcal{P}}_s, \hat{\mathcal{P}}_t^f) + \sqrt{\frac{2}{c'} \log(\frac{2}{\delta})} \left(\frac{1}{\sqrt{N_S}} + \frac{1}{\sqrt{N_T}} \right) + err_S(f^*) + err_T(f^*) + kM\phi(\lambda).$$

Correspond to the objective function

Correspond to the joint error minimizer illustrating that domain adaptation can work only if we can predict well in both domains

Assesses the probability under which the PTL does not hold

A Bound on the Target Error

Theorem 3.1 *Let f be any labeling function of $\in \mathcal{H}$. Let $\Pi^* = \operatorname{argmin}_{\Pi \in \Pi(\mathcal{P}_s, \mathcal{P}_t^f)} \int_{(\Omega \times \mathcal{C})^2} \alpha d(\mathbf{x}_s, \mathbf{x}_t) + \mathcal{L}(y_s, y_t) d\Pi(\mathbf{x}_s, y_s; \mathbf{x}_t, y_t)$ and $W_1(\hat{\mathcal{P}}_s, \hat{\mathcal{P}}_t^f)$ the associated 1-Wasserstein distance. Let $f^* \in \mathcal{H}$ be a Lipschitz labeling function that verifies the ϕ-probabilistic transfer Lipschitzness (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^*. We assume the input instances are bounded s.t. $|f^*(\mathbf{x}_1) - f^*(\mathbf{x}_2)| \leq M$ for all $\mathbf{x}_1, \mathbf{x}_2$. Let $\mathcal{L}$ be any symmetric loss function, k-Lipschitz and satisfying the triangle inequality. Consider a sample of N_s labeled source instances drawn from $\mathcal{P}_s$ and N_t unlabeled instances drawn from μ_t, and then for all $\lambda > 0$, with $\alpha = k\lambda$, we have with probability at least $1 - \delta$ that:*

$$\Pr_{(x_1, x_2) \sim \Pi(\mu_s, \mu_t)} \left[\left| f^*(x_1) - f^*(x_2) \right| > \lambda d(x_1, x_2) \right] \leq \phi(\lambda)$$

A Bound on the Target Error

$$err_T(f) \le W_1\left(\hat{P}_s, \hat{P}_t^f\right) + \sqrt{\frac{2}{c'}\log\left(\frac{2}{\delta}\right)}\left(\frac{1}{\sqrt{N_s}+\sqrt{N_t}}\right) + err_S(f^*) + err_T(f^*) + kM\phi(\lambda)$$

Proof:

$$\begin{aligned} \mathrm{err}_T(f) &= E_{(\mathbf{x},y)\sim P_t}\mathcal{L}(y, f(\mathbf{x})) \\ &\le E_{(\mathbf{x},y)\sim P_t}\mathcal{L}(y, f^*(\mathbf{x})) + \mathcal{L}(f^*(\mathbf{x}), f(\mathbf{x})) \\ &= E_{(\mathbf{x},y)\sim P_t}\mathcal{L}(f(\mathbf{x}), f^*(\mathbf{x})) + err_T(f^*) \end{aligned}$$

Triangle inequality

Definition $err_T(f)$, Symmetric

Since $E_{(\mathbf{x},y)\sim\mathcal{P}_t}L(f(\mathbf{x}), f^*(\mathbf{x})) = E_{(\mathbf{x},f(\mathbf{x}))\sim\mathcal{P}_t^f}L(f(\mathbf{x}), f^*(\mathbf{x})) \stackrel{\text{def}}{=} err_{T^f}(f^*(\mathbf{x}))$, then

$$\begin{aligned} \mathrm{err}_T(f) &= E_{(\mathbf{x},y)\sim P_t}\mathcal{L}(y, f(\mathbf{x})) \\ &= err_{T^f}(f^*) - err_S(f^*) + err_S(f^*) + err_T(f^*) \\ &\le \boxed{\left|err_{T^f}(f^*) - err_S(f^*)\right|} + err_S(f^*) + err_T(f^*) \end{aligned}$$

A Bound on the Target Error

$$err_T(f) \le W_1\left(\hat{P}_s, \hat{P}_t^f\right) + \sqrt{\frac{2}{c'}\log\left(\frac{2}{\delta}\right)}\left(\frac{1}{\sqrt{N_s}+\sqrt{N_t}}\right) + err_S(f^*) + err_T(f^*) + kM\phi(\lambda)$$

Proof:

$$\begin{aligned}
&|err_{T^f}(f^*) - err_S(f^*)| \\
&= \left|\int_{\Omega\times\mathcal{C}} \mathcal{L}(y, f^*(\mathbf{x}))(\mathcal{P}_t^f(\boldsymbol{X}=\mathbf{x}, Y=y) - \mathcal{P}_s(\boldsymbol{X}=\mathbf{x}, Y=y))d\mathbf{x}dy\right| \\
&= \left|\int_{\Omega\times\mathcal{C}} \mathcal{L}(y, f^*(\mathbf{x}))d(\mathcal{P}_t^f - \mathcal{P}_s)\right| \\
&\le \int_{(\Omega\times\mathcal{C})^2} \left|\mathcal{L}(y_t^f, f^*(\mathbf{x}_t)) - \mathcal{L}(y_s, f^*(\mathbf{x}_s))\right| \mathrm{d}\Pi^*((\mathbf{x}_s, y_s), (\mathbf{x}_t, y_t^f)) \\
&= \int_{(\Omega\times\mathcal{C})^2} \Big|\mathcal{L}(y_t^f, f^*(\mathbf{x}_t)) - \underline{\mathcal{L}(y_t^f, f^*(\mathbf{x}_s))} + \\
&\qquad \underline{\mathcal{L}(y_t^f, f^*(\mathbf{x}_s))} - \mathcal{L}(y_s, f^*(\mathbf{x}_s))\Big| \mathrm{d}\Pi^*((\mathbf{x}_s, y_s), (\mathbf{x}_t, y_t^f)) \\
&\le \int_{(\Omega\times\mathcal{C})^2} \left|\mathcal{L}(y_t^f, f^*(\mathbf{x}_t)) - \mathcal{L}(y_t^f, f^*(\mathbf{x}_s))\right| \\
&\qquad + \left|\mathcal{L}(y_t^f, f^*(\mathbf{x}_s)) - \mathcal{L}(y_s, f^*(\mathbf{x}_s))\right| \mathrm{d}\Pi^*((\mathbf{x}_s, y_s), (\mathbf{x}_t, y_t^f))
\end{aligned}$$

→ Conditional probability definition

Duality form of Kantorovitch-Rubinstein theorem

→ Triangle inequality

A Bound on the Target Error

$$err_T(f) \le W_1\left(\hat{P}_s, \hat{P}_t^f\right) + \sqrt{\frac{2}{c'}\log\left(\frac{2}{\delta}\right)}\left(\frac{1}{\sqrt{N_s}+\sqrt{N_t}}\right) + err_S(f^*) + err_T(f^*) + kM\phi(\lambda)$$

Proof:

$$\begin{aligned}
|err_{T^f}(f^*) - err_S(f^*)| &\le \int_{(\Omega\times\mathcal{C})^2} \left|\mathcal{L}(y_t^f, f^*(\mathbf{x}_t)) - \mathcal{L}(y_t^f, f^*(\mathbf{x}_s))\right| \\
&\quad + \left|\mathcal{L}(y_t^f, f^*(\mathbf{x}_s)) - \mathcal{L}(y_s, f^*(\mathbf{x}_s))\right| \mathrm{d}\Pi^*((\mathbf{x}_s, y_s), (\mathbf{x}_t, y_t^f)) \\
&\le \int_{(\Omega\times\mathcal{C})^2} k\,|f^*(\mathbf{x}_t) - f^*(\mathbf{x}_s)| + && \longrightarrow \text{k-Lipchitz inequality} \\
&\quad \left|\mathcal{L}(y_t^f, f^*(\mathbf{x}_s)) - \mathcal{L}(y_s, f^*(\mathbf{x}_s))\right| \mathrm{d}\Pi^*((\mathbf{x}_s, y_s), (\mathbf{x}_t, y_t^f)) \\
&\le k * M * \phi(\lambda) + \int_{(\Omega\times\mathcal{C})^2} k\lambda d(\mathbf{x}_t, \mathbf{x}_s) + && \longrightarrow \text{PTL} \\
&\quad \left|\mathcal{L}(y_t^f, f^*(\mathbf{x}_s)) - \mathcal{L}(y_s, f^*(\mathbf{x}_s))\right| \mathrm{d}\Pi^*((\mathbf{x}_s, y_s), (\mathbf{x}_t, y_t^f)) \\
&\le \int_{(\Omega\times\mathcal{C})^2} \alpha d(\mathbf{x}_s, \mathbf{x}_t) + \mathcal{L}(y_t^f, y_s)\mathrm{d}\Pi^*((\mathbf{x}_s, y_s), (\mathbf{x}_t, y_t^f)) + k * M * \phi(\lambda) && \longrightarrow \text{Triangle, } \alpha = k\lambda \\
&\le \int_{(\Omega\times\mathcal{C})^2} \alpha d(\mathbf{x}_s, \mathbf{x}_t) + \mathcal{L}(y_s, y_t^f)\mathrm{d}\Pi^*((\mathbf{x}_s, y_s), (\mathbf{x}_t, y_t^f)) + k * M * \phi(\lambda) \\
&= W_1(\mathcal{P}_s, \mathcal{P}_t^f) + k * M * \phi(\lambda).
\end{aligned}$$

A Bound on the Target Error

$$err_T(f) \leq W_1\left(\hat{P}_s, \hat{P}_t^f\right) + \sqrt{\frac{2}{c'}\log\left(\frac{2}{\delta}\right)}\left(\frac{1}{\sqrt{N_s}+\sqrt{N_t}}\right) + err_S(f^*) + err_T(f^*) + kM\phi(\lambda)$$

Proof:

$$\begin{aligned} |err_{T^f}(f^*) - err_S(f^*)| &\leq \int_{(\Omega\times\mathcal{C})^2} \alpha d(\mathbf{x}_s, \mathbf{x}_t) + \mathcal{L}(y_s, y_t^f)\mathbf{d}\Pi^*((\mathbf{x}_s, y_s), (\mathbf{x}_t, y_t^f)) + k * M * \phi(\lambda) \\ &= W_1(\mathcal{P}_s, \mathcal{P}_t^f) + k * M * \phi(\lambda). \end{aligned}$$

Using triangle inequality of W1 distance:

$$\begin{aligned} W_1(\mathcal{P}_s, \mathcal{P}_t^f) &\leq W_1(\mathcal{P}_s, \hat{\mathcal{P}}_s) + W_1(\hat{\mathcal{P}}_s, \hat{\mathcal{P}}_t^f) + W_1(\hat{\mathcal{P}}_t^f, \mathcal{P}_t^f) \\ &\leq W_1(\hat{\mathcal{P}}_s, \hat{\mathcal{P}}_t^f) + \sqrt{\frac{2}{c'}\log(\frac{2}{\delta})}\left(\frac{1}{\sqrt{N_s}} + \frac{1}{\sqrt{N_t}}\right). \end{aligned}$$

Using a result from Bolley's paper

A Bound on the Target Error

$$err_T(f) \leq W_1\left(\hat{P}_s, \hat{P}_t^f\right) + \sqrt{\frac{2}{c'}\log\left(\frac{2}{\delta}\right)}\left(\frac{1}{\sqrt{N_s} + \sqrt{N_t}}\right) + err_S(f^*) + err_T(f^*) + kM\phi(\lambda)$$

Proof:

$$\begin{aligned} W_1(\mathcal{P}_s, \mathcal{P}_t^f) &\leq W_1(\mathcal{P}_s, \hat{\mathcal{P}}_s) + W_1(\hat{\mathcal{P}}_s, \hat{\mathcal{P}_t^f}) + W_1(\hat{\mathcal{P}_t^f}, \mathcal{P}_t^f) \\ &\leq W_1(\hat{\mathcal{P}}_s, \hat{\mathcal{P}_t^f}) + \sqrt{\frac{2}{c'}\log(\frac{2}{\delta})}\left(\frac{1}{\sqrt{N_s}} + \frac{1}{\sqrt{N_t}}\right). \end{aligned}$$

Using a result from Bolley's paper

Theorem E.1 (from [35], Theorem 1.1.) *Let μ be a probability measure in Z so that for some $\alpha > 0$ we have for any $\mathbf{z}'$ $\int_{\mathbb{R}^d} e^{\alpha dist(\mathbf{z},\mathbf{z}')^2} d\mu < \infty$ and $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} \delta_{z_i}$ be the associated empirical measure defined on a sample of independent variables $\{\mathbf{z}_i\}_{i=1}^N$ drawn from μ. Then, for any $d' > dim(Z)$ and $c' < c$, there exists some constant N_0 depending on d' and some square exponential moments of μ such that for any $\epsilon > 0$ and $N \geq N_0 \max(\epsilon^{-(d'+2)}, 1)$,*

$$P[W_1(\mu, \hat{\mu}) > \epsilon] \leq \exp\left(-\frac{c'}{2}N\epsilon^2\right)$$

where c' can be calculated explicitly.

A Bound on the Target Error

$$err_T(f) \le W_1(\hat{P}_s, \hat{P}_t^f) + \sqrt{\frac{2}{c'}\log\left(\frac{2}{\delta}\right)}\left(\frac{1}{\sqrt{N_s}+\sqrt{N_t}}\right) + err_S(f^*) + err_T(f^*) + kM\phi(\lambda)$$

Proof:

$$\begin{aligned} W_1(\mathcal{P}_s, \mathcal{P}_t^f) &\le W_1(\mathcal{P}_s, \hat{\mathcal{P}}_s) + W_1(\hat{\mathcal{P}}_s, \hat{\mathcal{P}_t^f}) + W_1(\hat{\mathcal{P}_t^f}, \mathcal{P}_t^f) \\ &\le W_1(\hat{\mathcal{P}}_s, \hat{\mathcal{P}_t^f}) + \sqrt{\frac{2}{c'}\log(\frac{2}{\delta})}\left(\frac{1}{\sqrt{N_s}} + \frac{1}{\sqrt{N_t}}\right). \end{aligned}$$

$$P[W_1(\mu, \hat{\mu}) > \epsilon] \le \exp\left(-\frac{c'}{2}N\epsilon^2\right) \Rightarrow$$

$$\Pr\left[W_1\left(P_s, \hat{P}_s\right) > \varepsilon\right] \le \exp\left(-\frac{c'}{2}N_s\varepsilon^2\right) \triangleq \frac{\delta}{2},$$

$$\Pr\left[W_1\left(P_t^f, \hat{P}_f^f\right) > \varepsilon\right] \le \exp\left(-\frac{c'}{2}N_t\varepsilon^2\right) \triangleq \frac{\delta}{2},$$

$\Rightarrow$ $W_1\left(P_s, \hat{P}_s\right) + W_1\left(P_t^f, \hat{P}_f^f\right) \le \sqrt{\frac{2}{c'}\log\left(\frac{2}{\delta}\right)}\left(\frac{1}{\sqrt{N_s}} + \frac{1}{\sqrt{N_t}}\right)$ with at least1-δ probability. □

OUTLINE

Learning with Joint Distribution OT

Optimization using BCD

Assume that the function space $\mathcal{H}$ to which f belongs is either a RKHS or a function space parametrized by some parameters $\mathbf{w} \in \mathbb{R}^p$.

RKHS: Reproducing kernel Hilbert space

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

where the loss function $\mathcal{L}$ is continuous and differentiable with respects to its second variable. $\Omega(f)$ is the regularization term either a non-decreasing function of the squared-norm or a squared-norm on the vector parameter. $\Omega(f)$ is continuously differentiable.

The optimization problem with fixed leads to a new learning problem expressed as

$$\min_{f \in \mathcal{H}} \sum_{i,j} \gamma_{i,j} \mathcal{L}(y_i^s, f(\mathbf{x}_j^t)) + \lambda \Omega(f)$$

Learning with Joint Distribution OT

Optimization using BCD

Algorithm 1 Optimization with Block Coordinate Descent

Initialize function f^0 and set $k = 1$
Set α and λ
while not converged **do**
  $\gamma^k \leftarrow$ Solve OT problem (in paper) with fixed f^{k-1}
  $f^k \leftarrow$ Solve learning problem (in paper) with fixed γ^k
  $k \leftarrow k + 1$
end while

$$\gamma_0 = \underset{\gamma \in \Pi(\mathcal{P}_s, \mathcal{P}_t)}{\operatorname{argmin}} \int_{(\Omega \times \mathcal{C})^2} \mathcal{D}(\mathbf{x}_1, y_1; \mathbf{x}_2, y_2) d\gamma(\mathbf{x}_1, y_1; \mathbf{x}_2, y_2),$$

$$y_2 = f^{k-1}(x_2)$$

$$\min_{f \in \mathcal{H}} \sum_{i,j} \gamma_{i,j} \mathcal{L}(y_i^s, f(\mathbf{x}_j^t)) + \lambda \Omega(f)$$

OUTLINE

Examples

3-class toy example

Source domain samples: drawn from three different 2D Gaussian distributions with different centers and standard deviations. (+)

Target domain: obtained by rotating the source distribution by $\pi/4$ radian.(°)

Two types of kernel are considered: linear and RBF

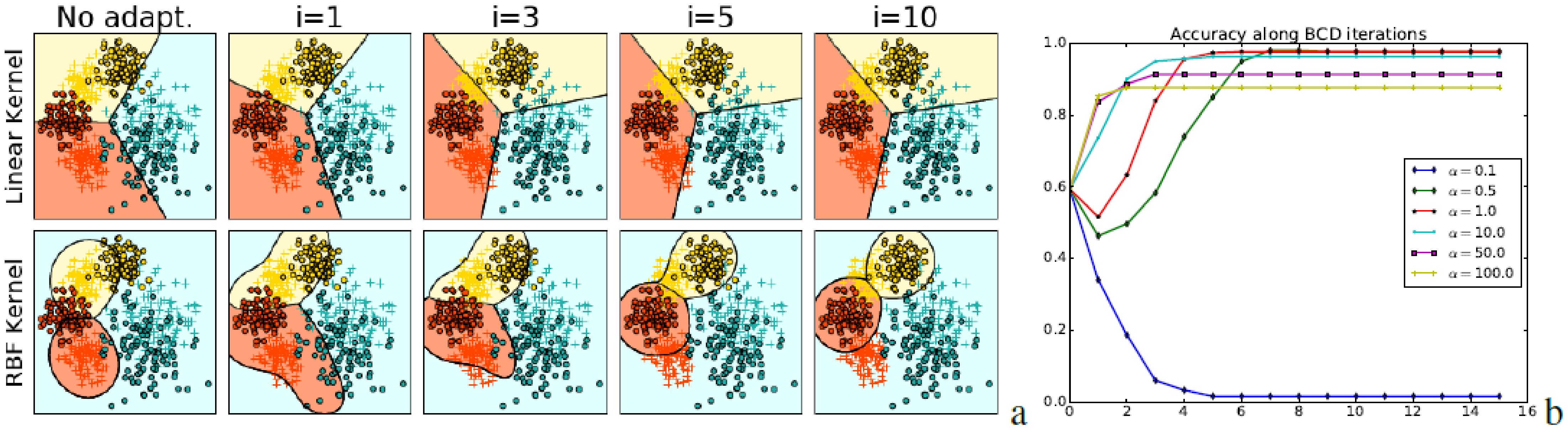
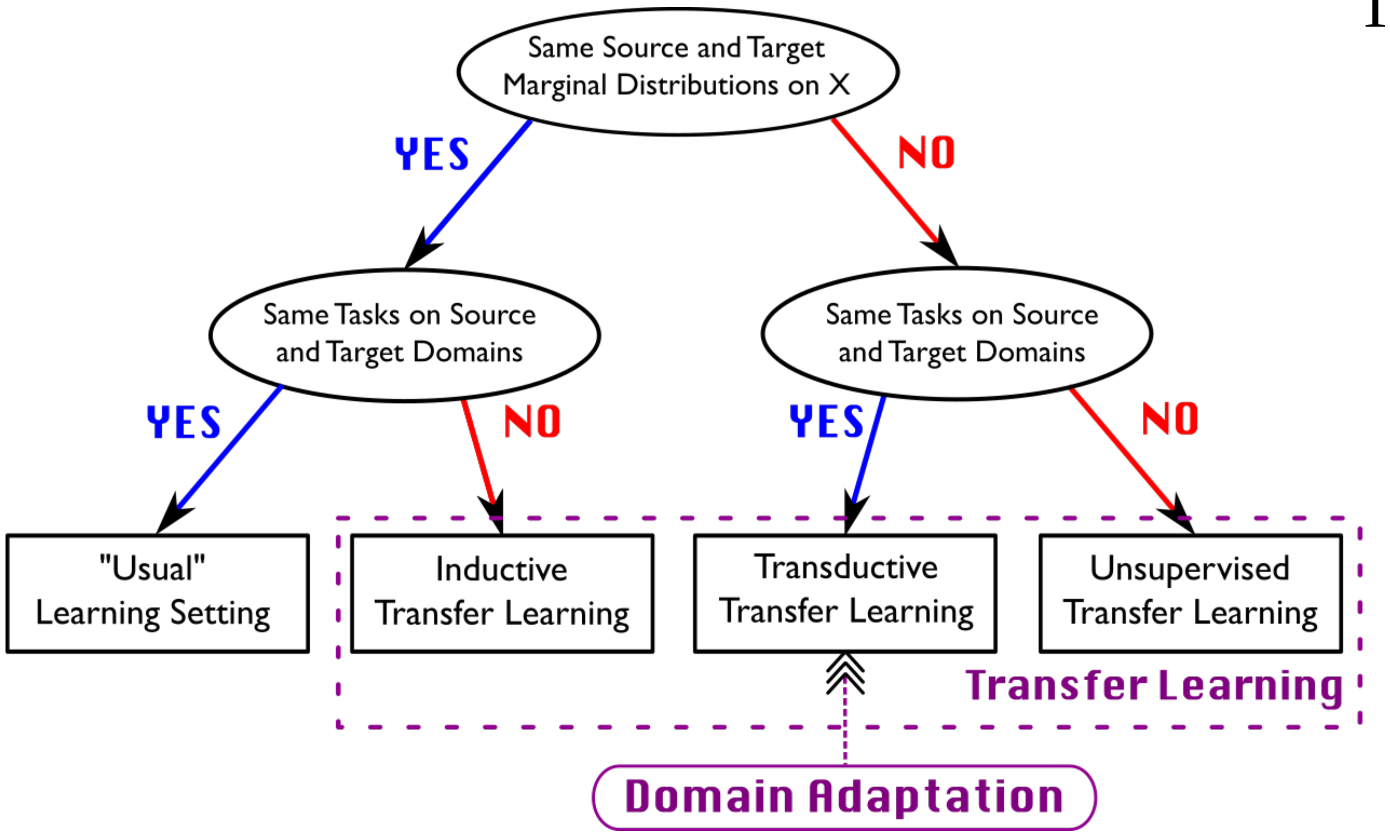


Figure 2: **Illustration on a toy example.** (a): Decision boundaries for linear and RBF kernels on selected iterations. The source domain is depicted with crosses, while the target domain samples are class-colored circles. (b): Evolution of the accuracy along 15 iterations of the method for different values of the α parameter;

QUESTION

- Thank you!

Problem Statement



Distinction between usual machine learning setting and transfer learning, and positioning of domain adaptation.

The different types of domain adaptation:

Unsupervised domain adaptation:

the learning sample contains a set of labeled source examples, a set of unlabeled source examples and an unlabeled set of target examples.

Semi-supervised domain adaptation:

consider a "small" set of labeled target examples.

Supervised domain adaptation:

all the examples considered are supposed to be labeled.