

変数の拡張に対する最適輸送 を用いたドメイン適応

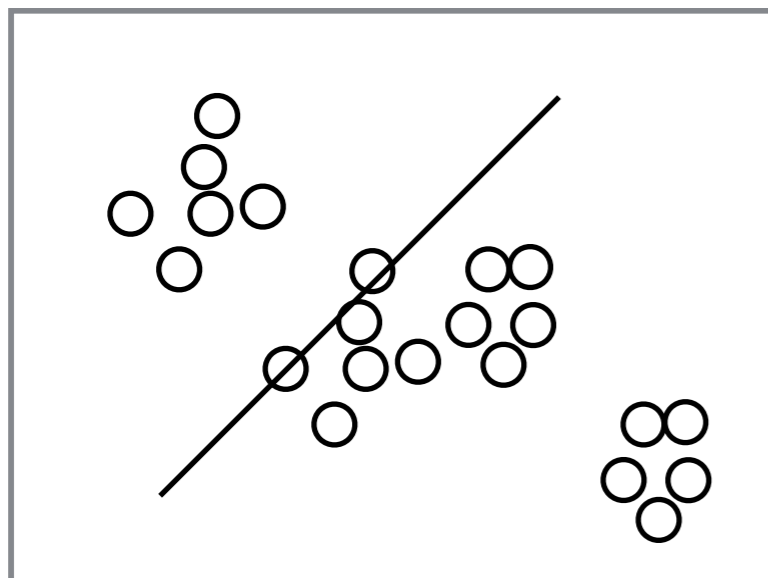
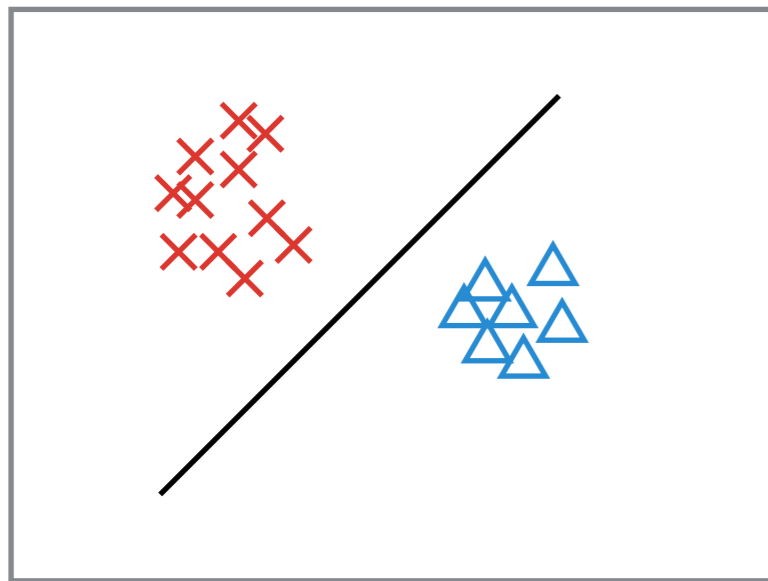
有竹 俊光¹, 日野 英逸^{1,2}

¹統計数理研究所, ²理化学研究所

情報論的学習理論と機械学習研究会 (IBISML)

2022/01/18

教師付き学習



- 教師付き学習

- 教師データ : $\mathcal{D}_S = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$

- テストデータ : $\mathcal{D}_T = \{x_j^t\}_{j=1}^{N_t}$

- 教師, テストデータの分布が異なるると予測精度が低下

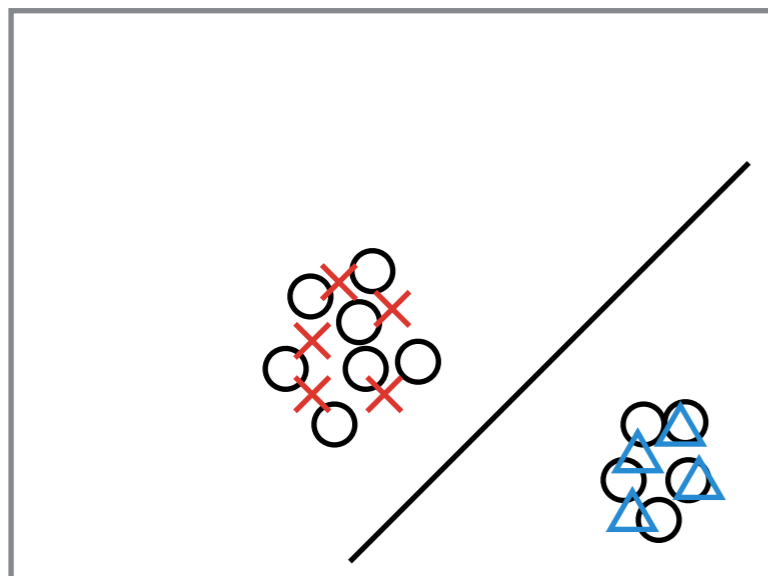
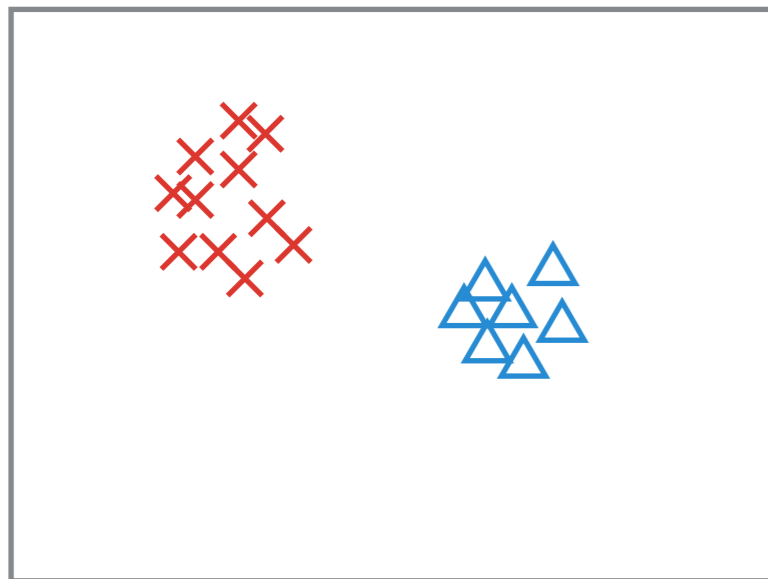
- ソースドメイン

教師データが得られる領域 + 分布

- ターゲットドメイン

テストデータが得られる領域 + 分布

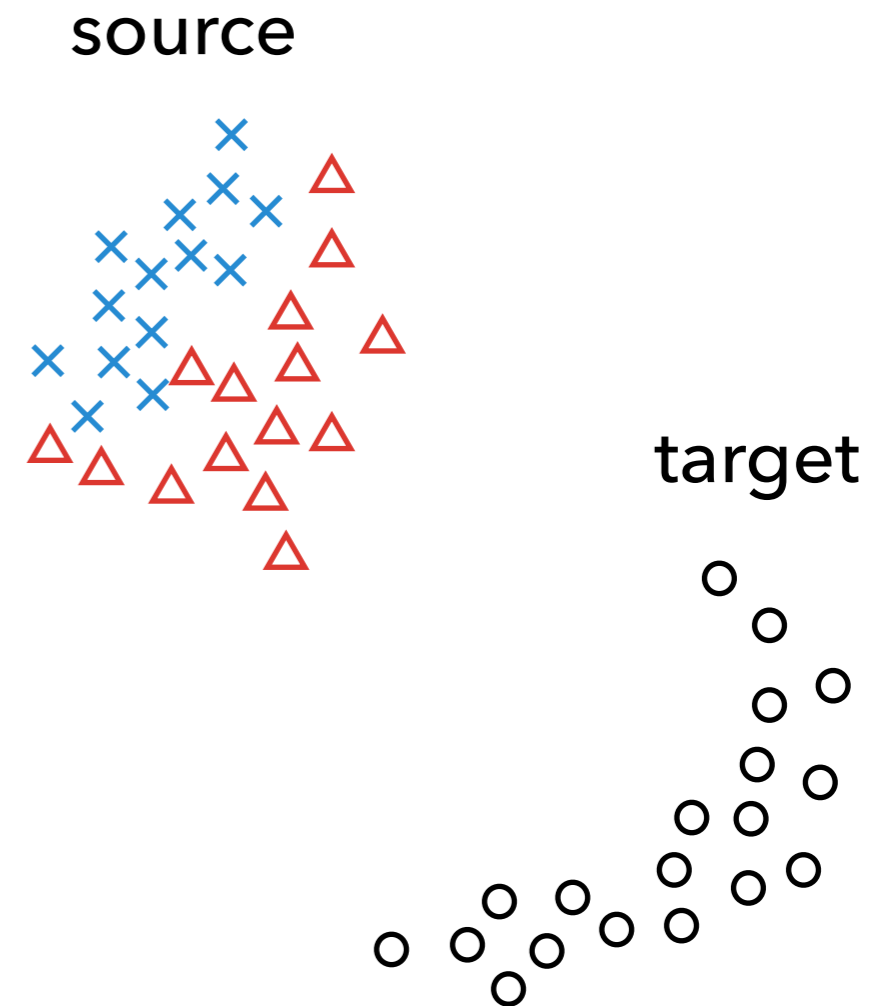
ドメイン適応



- 教師, テストデータの同時分布を近づける
 $p_s(x', y') \approx p_t(x', y')$
 - ドメイン不変な特徴の抽出
 - 密度比による標本の重み付け
 - 最適輸送による教師データの輸送
- テストデータのラベルは未知のもとで適応を行う
(教師なしドメイン適応)

背景：最適輸送を用いた ドメイン適応

- 教師データ $\{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ を
ターゲットドメインへ輸送
- 輸送先で判別器を学習
- テストデータのラベルは未知
 - $\mathcal{P}_S(y|\mathbf{x}) = \mathcal{P}_T(y|\mathcal{T}(\mathbf{x}))$ を仮定
 - 最適輸送前後での各点での
ラベル分布が一致



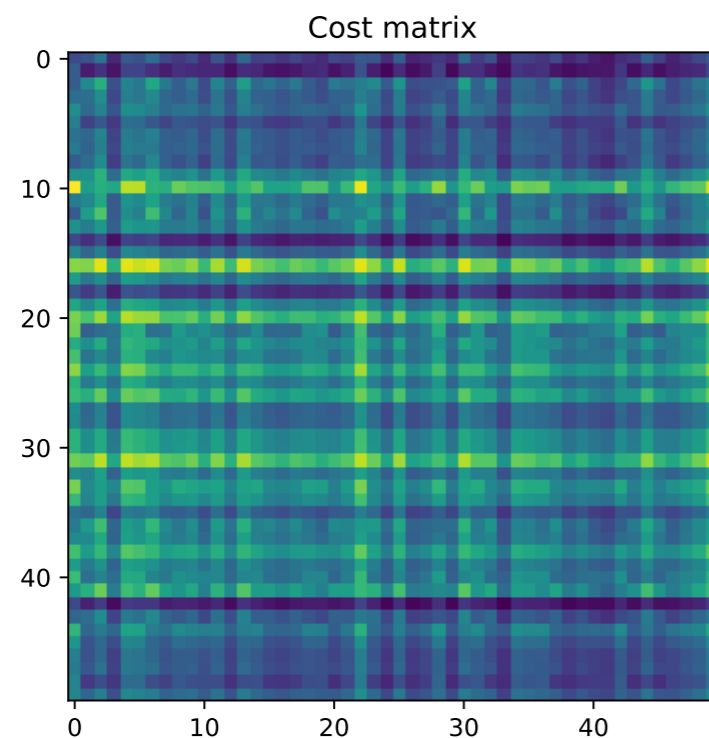
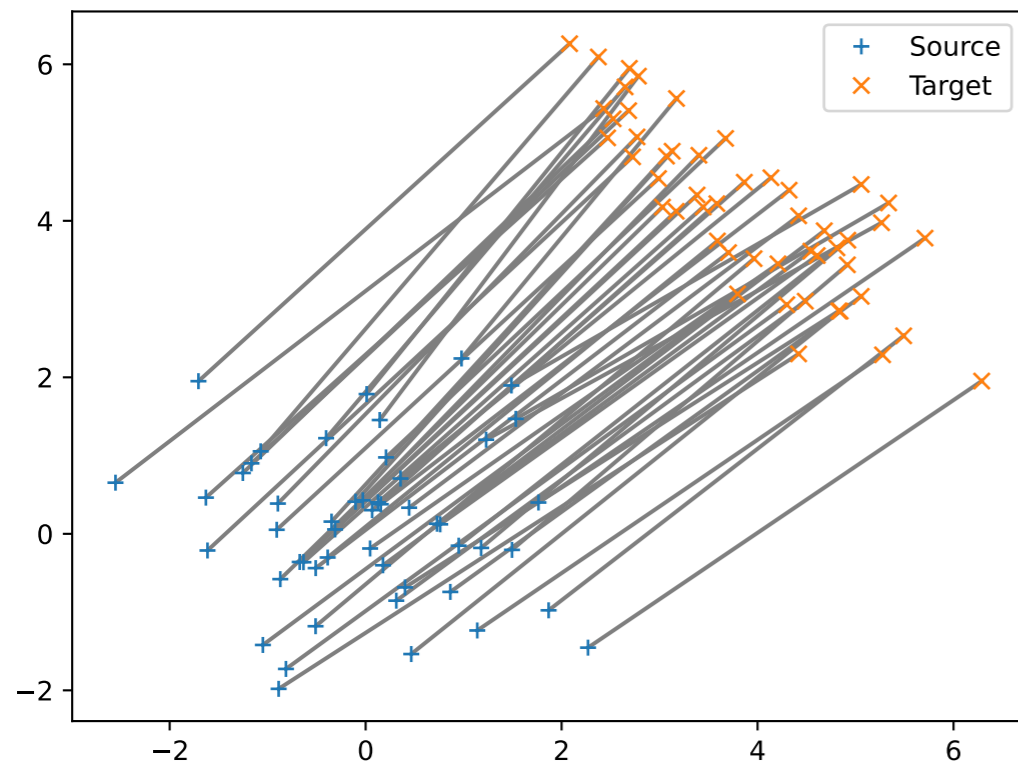
背景：最適輸送問題

$$\hat{\pi}^* = \arg \min_{\hat{\pi} \in \hat{\Pi}(\mathcal{D}_S, \mathcal{D}_T)} \sum_{i=1}^{N_s} \sum_{j=1}^{N_t} \hat{\pi}_{ij} \mathcal{E}(\mathbf{x}_{si}^c; \mathbf{x}_{tj}^c)$$

$$\hat{\Pi}(\mathcal{D}_S, \mathcal{D}_T) \equiv \left\{ \pi \in \mathbb{R}_+^{N_s \times N_t} \mid \sum_{i=1}^{N_s} \pi_{ij} = \frac{1}{N_t}, \sum_{j=1}^{N_t} \pi_{ij} = \frac{1}{N_s} \right\}$$

- 輸送計画により定まる輸送量と輸送コストの積の総和を最小化する問題
 - 輸送計画 $\hat{\pi} \in \Pi(\mathcal{D}_S, \mathcal{D}_T)$
輸送元, 輸送先の対応関係を決める
 - 輸送コストは $\mathcal{E}(\mathbf{x}_s, \mathbf{x}_t) = \|\mathbf{x}_s - \mathbf{x}_t\|_2^2$ などを利用
- 線形計画法によって解くことが可能

最適輸送の例



- 輸送コストとして

$$C_{i,j} = c(\mathbf{x}_i, \mathbf{x}'_j) = \|\mathbf{x}_i - \mathbf{x}'_j\|_2^2$$
 を利用
- 最適輸送問題を解いて
 ソース, ターゲットで
 対応する標本を推定
- $N_s \neq N_t$ のときは一對多の
 対応関係が生じる

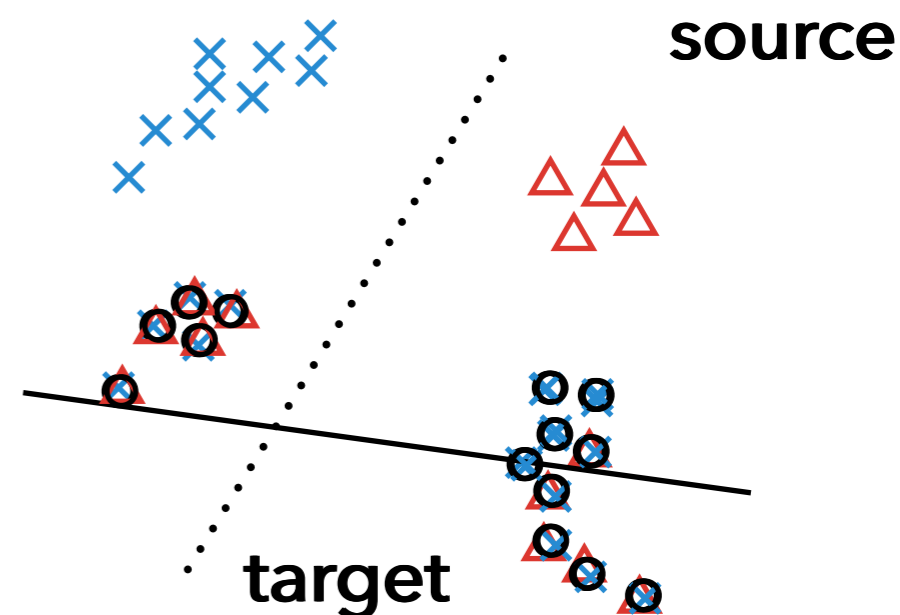
背景：Joint Distribution Optimal

Transport (JDOT)

$$\arg \min_{\hat{\pi} \in \hat{\Pi}(\mathcal{D}_S, \mathcal{D}_T^f), f \in \mathcal{F}} \sum_{i=1}^{N_s} \sum_{j=1}^{N_t} \hat{\pi}_{ij} \mathcal{E}_\alpha(\mathbf{x}_{si}^c, y_{si}; \mathbf{x}_{tj}^c, \hat{y}_{tj})$$

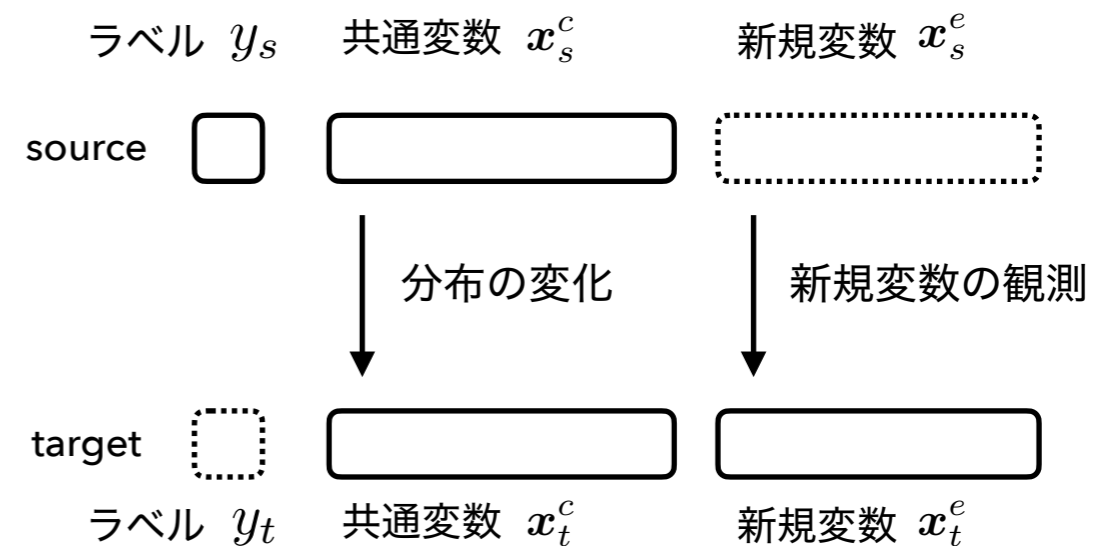
$$\mathcal{E}_\alpha(\mathbf{x}_{si}^c, y_{si}; \mathbf{x}_{tj}^c, \hat{y}_{tj}) = \alpha \|\mathbf{x}_{si}^c - \mathbf{x}_{tj}^c\|_2^2 + \mathcal{L}(y_{si}, \hat{y}_{tj})$$

- 教艦は $\mathcal{D}_S \times \mathcal{Y} \rightarrow \mathcal{D}_T$ (適応 $g(x)$)
- テストデータのラベルは不明
- 輸送コストにラベルの
- 類似致度を加える $f(x)$ をテスト
- データの同時分布 $p(x, y)$ を利用
- まづな輸送向期待学習



変数の拡張に対するドメイン適応

- ターゲットドメインでのみ新しい変数を観測
- センサー, 調査項目の追加



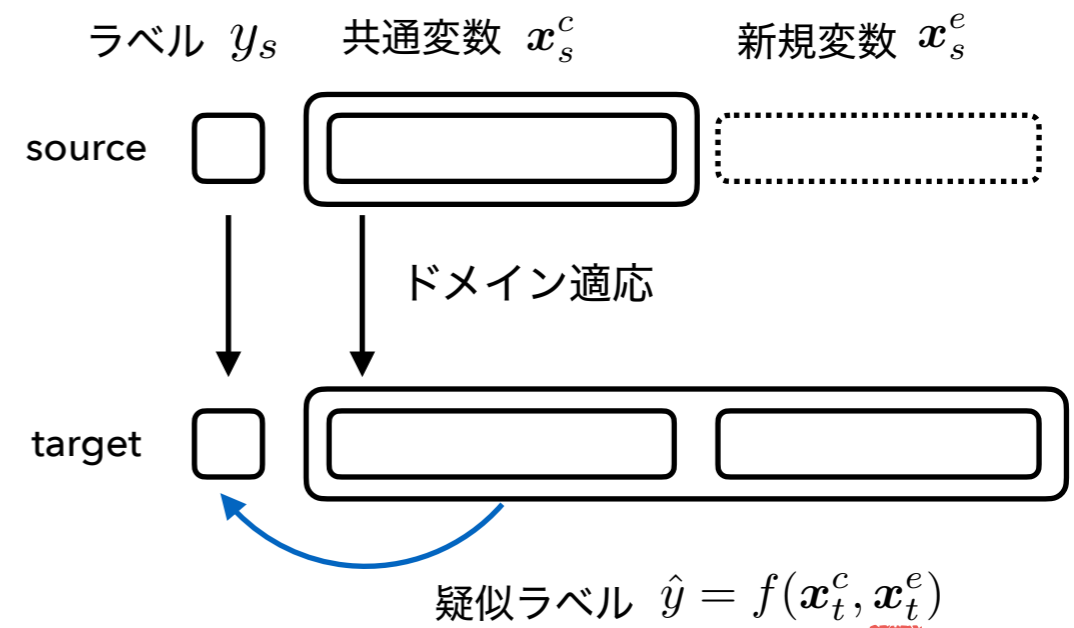
- 既存の変数はソース, ターゲットドメインで共通
- 同一のセンサ等により同じ情報を観測している
- 計測値の分布は変化する (計測対象の状態の違いなど)

本研究の貢献

- 共通変数の分布の変化, 新規変数の観測に対する最適輸送を用いたドメイン適応法の提案
 - ソース, ターゲット間双方向の最適輸送を利用
 - 実際には1方向の輸送により実現可能
- ターゲットドメインでの予測モデルのテスト誤差の上界の導出

変数の拡張に対するドメイン適応

- 次元の異なる変数間の輸送コストの定義が必要
 - ベクトル埋め込み等では変数の共通性を考慮できない



- ソースドメインの新規変数を推定してドメイン適応を行う
 - 新規変数の予測はターゲットドメインからソースドメインへの最適輸送により行う
 - JDOTと同様に疑似ラベルを利用した最適輸送を考える

双方向の輸送によるドメイン適応

- ソース, ターゲット間
双方向の輸送による
ドメイン適応

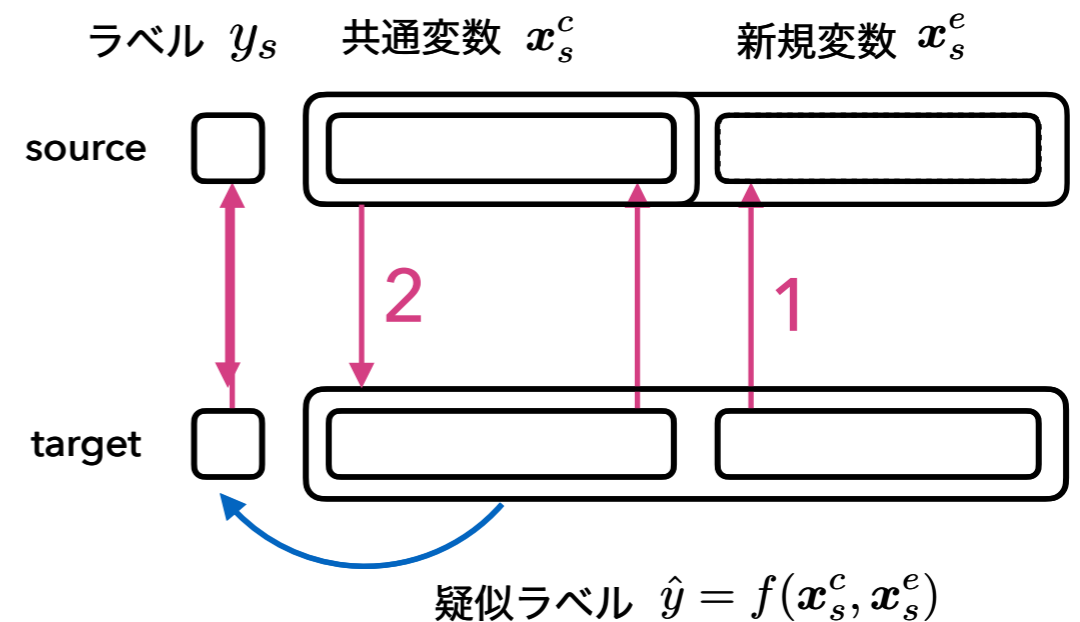
1. ターゲットからソースへ
新規変数を輸送

$$\mathcal{E}_\alpha(\mathbf{x}_{si}^c, y_{si}; \mathbf{x}_{tj}^c, \hat{y}_{tj}) = \alpha \|\mathbf{x}_{si}^c - \mathbf{x}_{tj}^c\|_2^2 + \mathcal{L}(y_{si}, \hat{y}_{tj})$$

2. ソースからターゲットへラベルを輸送

$$\mathcal{E}_\alpha(\mathbf{x}_{si}^c, \hat{\mathbf{x}}_{si}^e, y_{si}; \mathbf{x}_{tj}^c, \mathbf{x}_{tj}^e, \hat{y}_{tj}) = \alpha (\|\mathbf{x}_{si}^c - \mathbf{x}_{tj}^c\|_2^2 + \|\hat{\mathbf{x}}_{si}^e - \mathbf{x}_{tj}^e\|_2^2) + \mathcal{L}(y_{si}, \hat{y}_{tj})$$

3. 輸送したラベルによりモデル f を学習

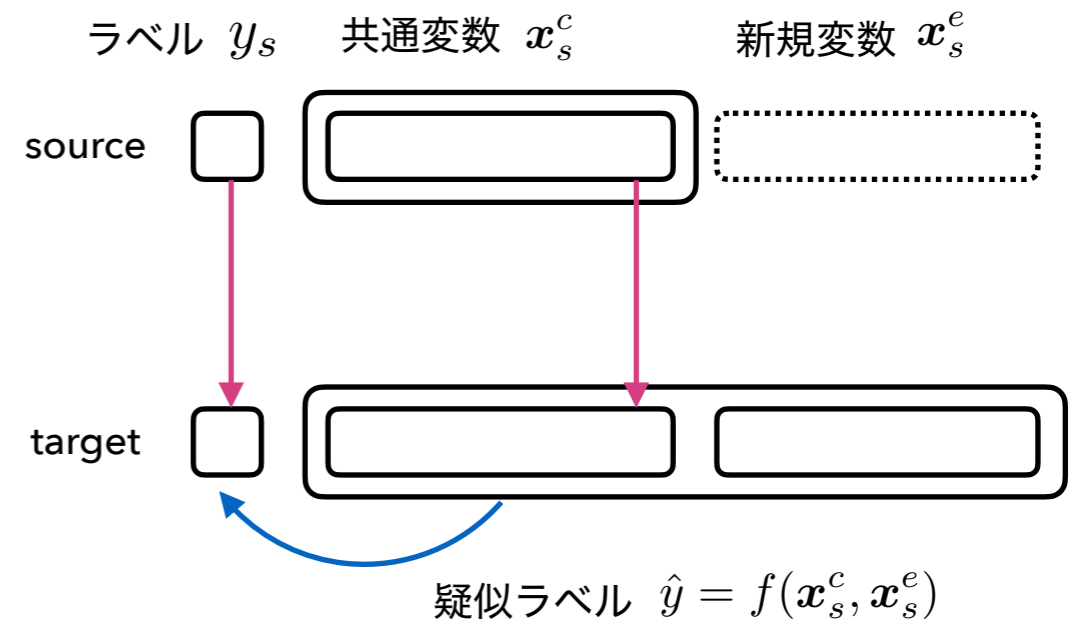


双方向の輸送によるドメイン適応

- 仮定

$$\mathcal{P}_T(x^e | x^c, y) = \mathcal{P}_S(x^e | \mathcal{T}(x^c, y))$$

- (x^c, y) の輸送先の x^e の分布が同一
- (x^c, y) の輸送先に輸送元の x^e と同じ x^e をもつ点が存在
- x^e の輸送コストが0となり (x^c, y) を用いた輸送と (x^c, x^e, y) を用いた輸送が等価



双方向の輸送によるドメイン適応

- 双方向の輸送はソースからターゲットへの1方向の輸送により実現可能

1. ソースからターゲットへラベルを輸送

$$\mathcal{E}_\alpha(\mathbf{x}_{si}^c, y_{si}; \mathbf{x}_{tj}^c, \hat{y}_{tj}) = \alpha \|\mathbf{x}_{si}^c - \mathbf{x}_{tj}^c\|_2^2 + \mathcal{L}(y_{si}, \hat{y}_{tj})$$

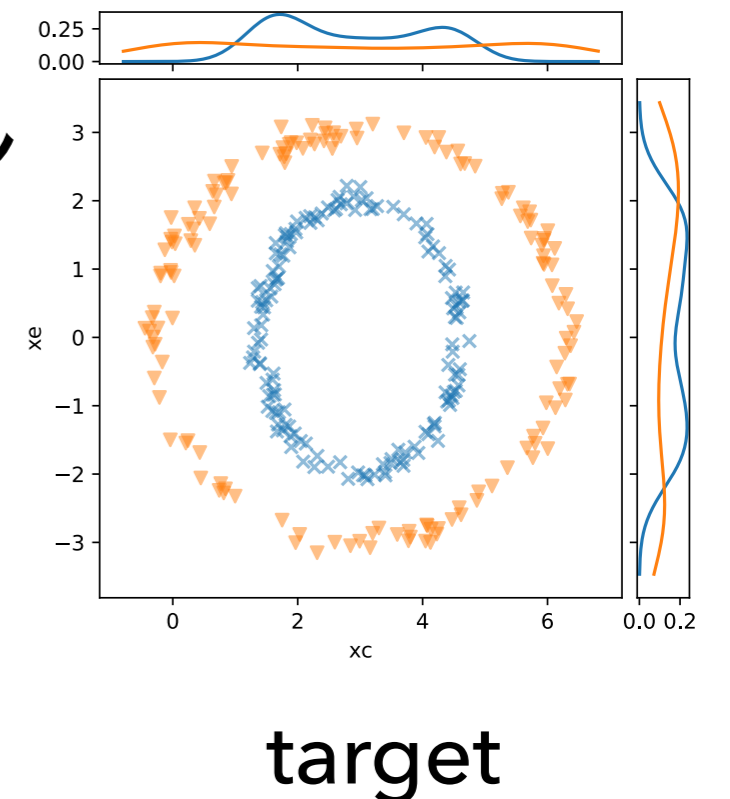
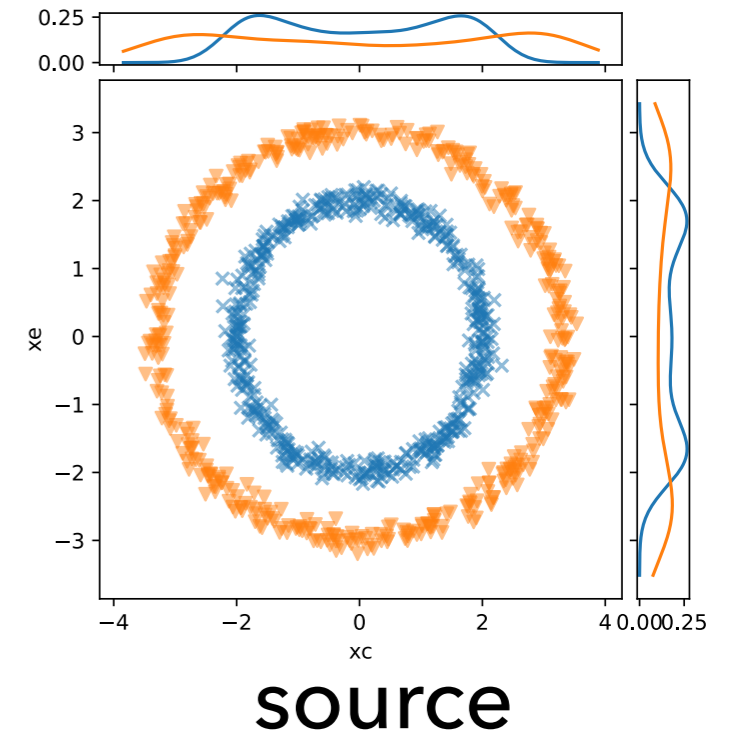
2. 輸送したラベルによりモデル f を学習

$$\operatorname{argmin}_{\hat{\pi} \in \hat{\Pi}(\mathcal{D}_S, \mathcal{D}_T^f), f \in \mathcal{F}} \sum_{i=1}^{N_s} \sum_{j=1}^{N_t} \hat{\pi}_{ij} \mathcal{E}_\alpha(\mathbf{x}_{si}^c, y_{si}; \mathbf{x}_{tj}^c, \hat{y}_{tj})$$

$$\mathcal{E}_\alpha(\mathbf{x}_{si}^c, y_{si}; \mathbf{x}_{tj}^c, \hat{y}_{tj}) = \alpha \|\mathbf{x}_{si}^c - \mathbf{x}_{tj}^c\|_2^2 + \mathcal{L}(y_{si}, \hat{y}_{tj} = f(\mathbf{x}_{tj}^c, \mathbf{x}_{tj}^e))$$

数値実験（人工データ）

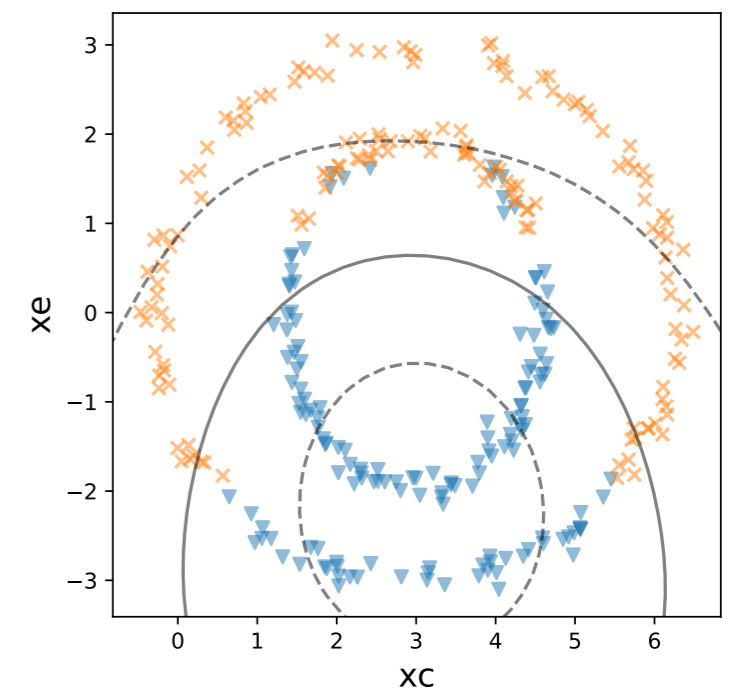
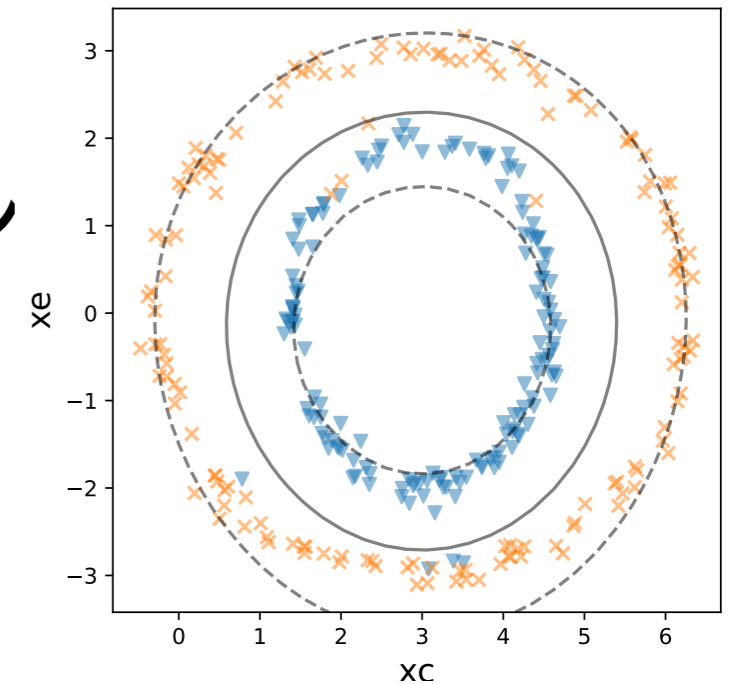
- 2次元の2つの円が異なるクラスとなるデータ
 - 共通変数, 新規変数がそれぞれ1次元
 - ソースの新規変数 x^e とターゲットのラベル y は輸送, 推定には利用しない
- モデル集合 \mathcal{F} としてRBFカーネルを用いたSVMを利用して推定



数値実験（人工データ）

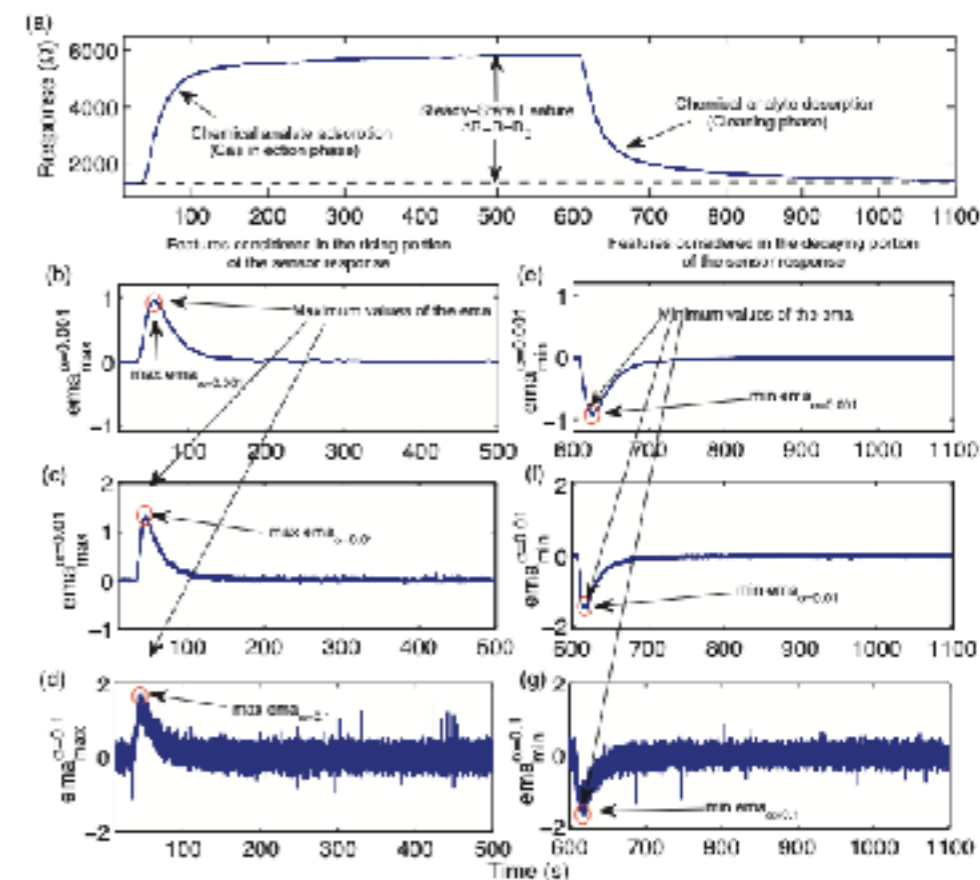
- 輸送したラベルの一部が不正確でも学習後のモデルは真のラベルを正しく予測可能
- 初期値等によっては誤った判別境界が得られる場合もある
- 新規変数を用いないJDOTよりも精度のよい推定が可能

	Proposed	JDOT (no extra)
transfer	93.8 (3.82)	86.6 (2.92)
model	98.9 (2.55)	88.1 (2.62)



数値実験（実データ）

- 2種類のガスの判別問題
 - 元データは時系列データ
 - 時系列データから1つのセンサにつき6個の特徴量を抽出しガスの種類を判別
- 計測時期の違いによりセンサの劣化度合いが異なる
- 16個のセンサのうち8個を共通のセンサ
残りの8個をテスト時のみ追加されるとして実験
- RBFカーネルを用いたSVMをモデルとして利用



Vergara et al., Sensor and Actuators B: Chemical (2012)

数値実験（実データ）

- 結果

- Baseline
ドメイン適応を行わない
- JDOT
新規変数を用いずに
適応を行った場合
- Proposed
提案法により新規変数を
考慮した場合

domains	Baseline	JDOT (no extra)	Proposed
1→2	83.13	77.71	78.31
1→3	52.51	93.45	96.02
1→4	64.49	60.75	87.85
2→1	52.13	79.26	84.04
2→3	57.31	89.36	89.47
2→4	63.55	69.16	71.96
3→1	52.13	92.02	94.15
3→2	72.09	81.92	88.76
3→4	94.39	77.57	81.31
4→1	49.47	52.66	53.72
4→2	42.97	71.08	74.30
4→3	91.11	82.81	82.57

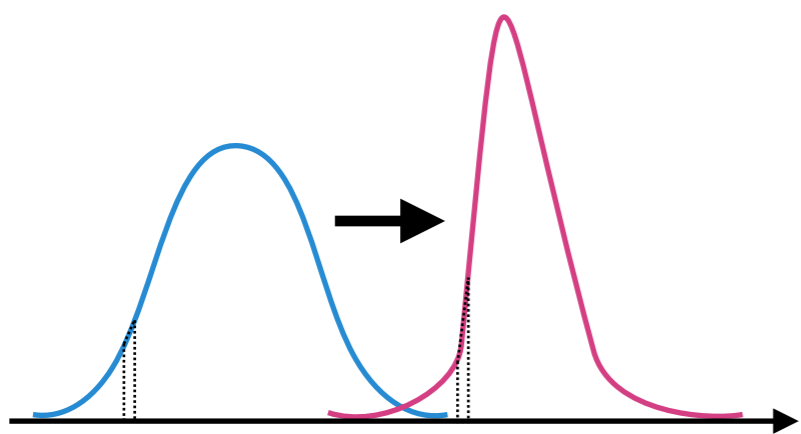
- 一部の適応が不要な場合を除いて新規変数を考慮することで予測精度が向上することを確認

まとめ

- ターゲットドメインでの新規変数の観測に対する最適輸送を用いたドメイン適応法を提案
 - ソース, ターゲットで共通となる変数の均質性を考慮
 - 双方向の輸送問題によりソースの新規変数とターゲットのラベルを推定してドメイン適応を実行
 - $\mathcal{P}_T(x^e|x^c, y) = \mathcal{P}_S(x^e|\mathcal{T}(x^c, y))$ という仮定のもとでは双方向の輸送はソースからターゲットへの輸送で実現可能
 - 実験的に新規変数を考慮しないJDOTと比較して精度のよい推定が可能なことを確認
- ターゲットドメインでの予測モデルの汎化誤差の上界の導出

Appendix

最適輸送問題



- ある質量の山の土を別の山に移す問題
- 土の移動コスト
 - 移す元と移す先の距離
 - 移す量
- 最小コストを達成する山の移し方は？

Mongeの問題

$$\text{下限} \quad \inf \left\{ \int_X c(x, T(x)) d\mu(x) \mid T_{\#}\mu = \nu \right\}$$

を達成する輸送写像 $T : X \rightarrow Y$ を求めよ

- μ : 距離空間 X 上の確率測度
- ν : 距離空間 Y 上の確率測度
- $c : X \times Y \rightarrow [0, \infty]$: 輸送コストを表す可測関数
- $T_{\#}\mu$: T による μ の像測度 (押し出し)

Mongeの問題

$$\text{下限 } \inf \left\{ \int_X c(x, T(x)) d\mu(x) \mid T_{\#}\mu = \nu \right\}$$

を達成する輸送写像 $T : X \rightarrow Y$ を求めよ

- 輸送の過程は考慮しない
- 解が存在しない場合がある
→ μ がDirac測度, ν が非Dirac測度
- 輸送射像を求める代わりに輸送計画を
求める問題を解く (Kantrovichの定式化)

Kantorovichの問題

$$\text{下限 } \inf \left\{ \int_{X \times Y} c(x, y) d\gamma(x, y) \mid \gamma \in \Gamma(\mu, \nu) \right\}$$

を達成するカップリング γ を求めよ

- $\Gamma(\mu, \nu)$: $X \times Y$ 上の確率測度かつ $\int_X \gamma(x, y) dx = \nu$
 $\int_Y \gamma(x, y) dy = \mu$ をみたす γ の集合
- $\gamma(x, y)$: $x \in X$ から $y \in Y$ への輸送量を表す

Kantorovichの問題

$$\text{下限 } \inf \left\{ \int_{X \times Y} c(x, y) d\gamma(x, y) \mid \gamma \in \Gamma(\mu, \nu) \right\}$$

を達成するカップリング γ を求めよ

- 実用上は $X = Y$, かつ $c(x, y)$ を
 適当な距離 $d(x, y)$ とすることが多い
 → Wasserstein距離の計算と等価

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu, \nu)} \int_{M \times M} d(x, y)^p d\gamma(x, y) \right)^{1/p}$$

離散化した場合

$$\text{minimize}_{\gamma \in \mathbb{R}^{n \times m}} \sum_{i=1}^n \sum_{j=1}^m c_{i,j} \gamma_{i,j}$$

$$\text{subject to } \sum_{i=1}^n \gamma_{i,j} = \nu_j, \sum_{j=1}^m \gamma_{i,j} = \mu_i$$

- 離散分布から離散分布へ輸送

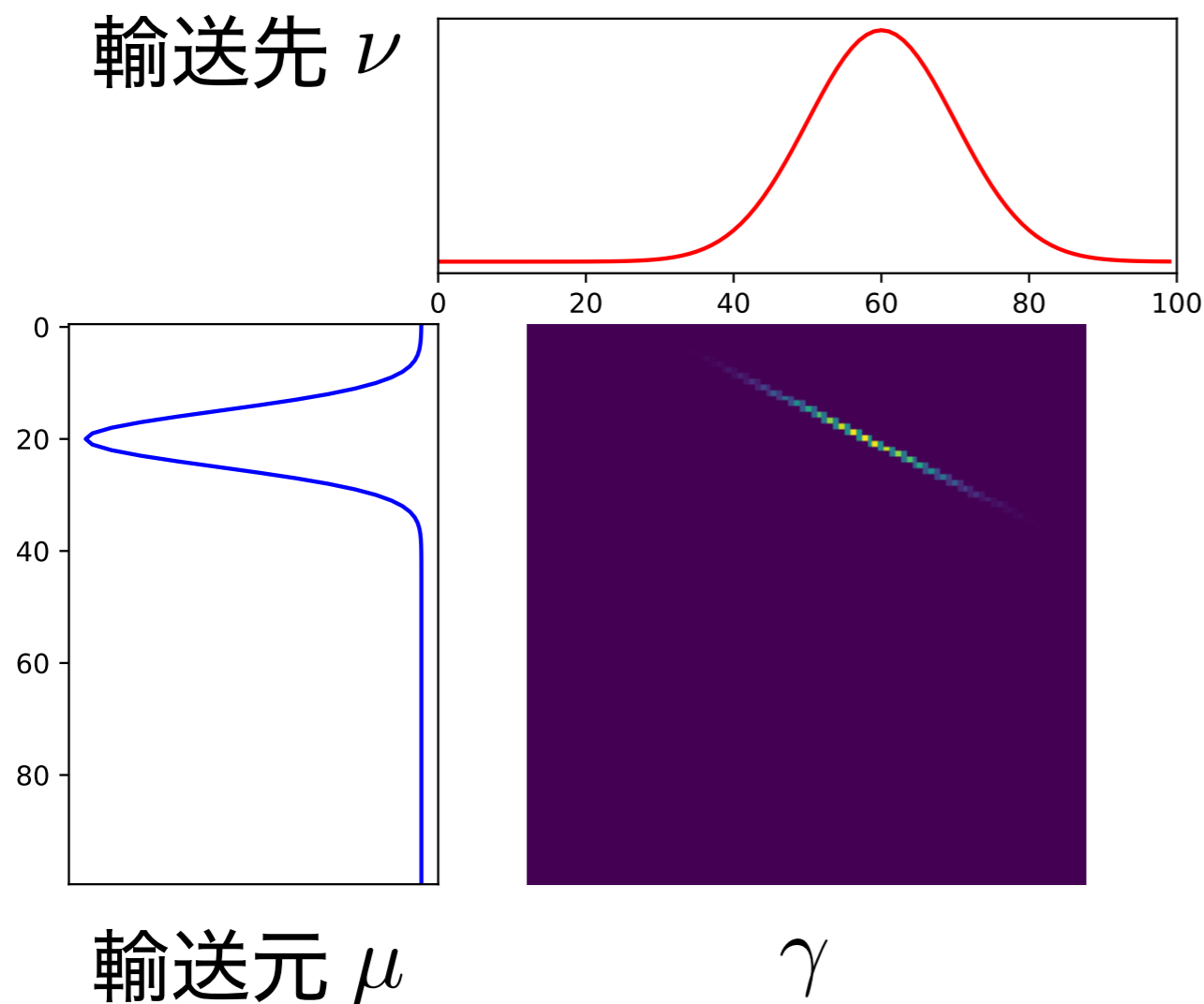
$$\mu \in W^n, \nu \in W^m, W^n = \{w \mid w \in \mathbb{R}_+^n, \sum_{i=1}^n w_i = 1\}$$

- カップリング γ およびコスト c は行列

$$\gamma_{i,j} = \gamma(x_i, y_j), c_{i,j} = c(x_i, y_j)$$

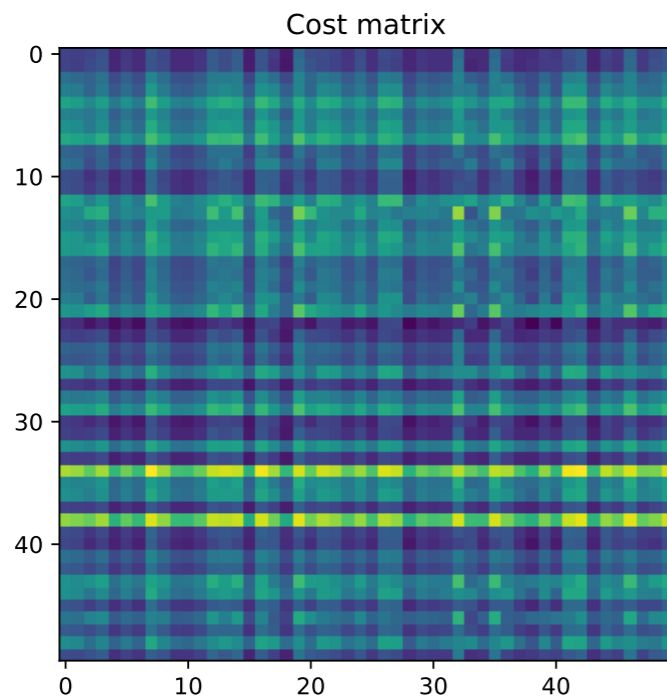
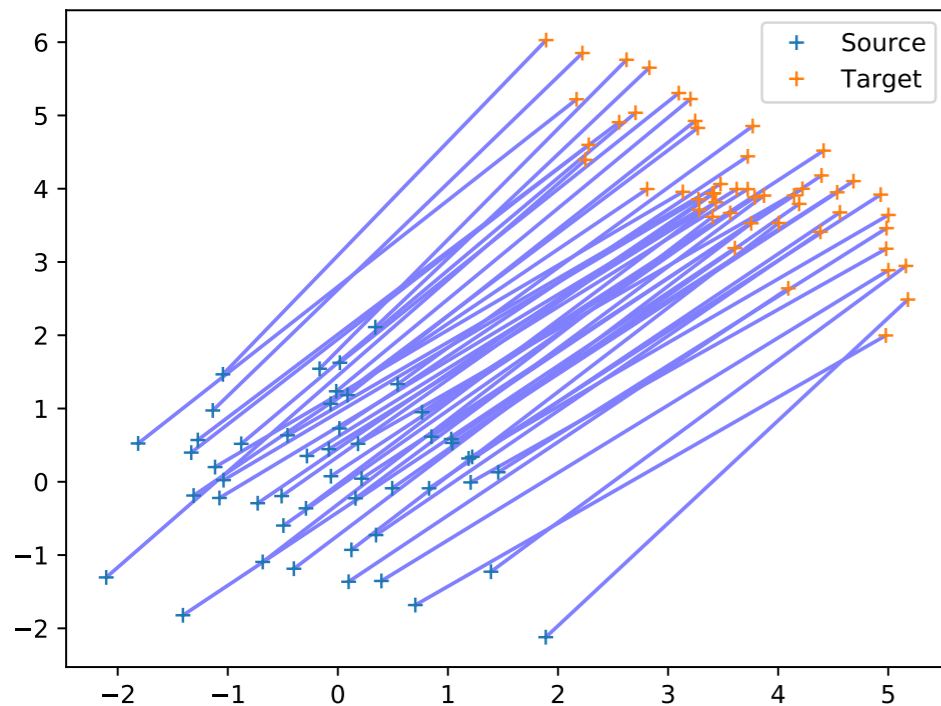
- 線形計画法を解くことで解が求まる

例：1次元の分布の最適輸送



- コストとして $c(x, y) = (x - y)^2$ を利用
- 1次元のヒストグラムを輸送元から輸送先へ最小コストで動かす
- γ によってどこからどこへ動かしたかわかる

例：経験分布の最適輸送



- 輸送コストとして

$$c_{i,j} = c(\mathbf{x}_i, \mathbf{x}'_j) = \|\mathbf{x}_i - \mathbf{x}'_j\|_2^2$$

を利用

- 各標本点上の一様分布を輸送する

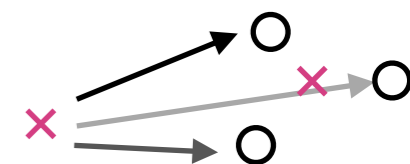
$$\mu = \mathbf{1}/n \in W^n, \nu = \mathbf{1}/n' \in W^{n'}$$

- $n \neq n'$ のときは一對多の対応関係が生じる

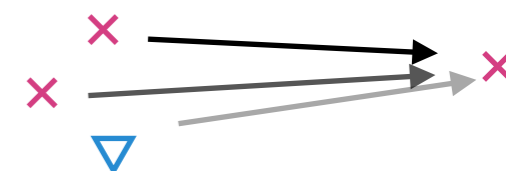
カップリングを用いた 標本点の輸送

- ソースの標本 : $\{(x_i^s, z_i^s)\}_{i=1}^{N^s}$, $z^s = (z_1^s, \dots, z_{N^s}^s)^\top$
- ターゲットの標本 : $\{x_j^t\}_{j=1}^{N^t}$, $X^t = [x_1^t \ \dots \ x_{N^t}^t]^\top$
- カップリング : $\gamma \in \Gamma(\mu, \nu)$
- 重心射像 (barycentric mapping)

1. $\{x_j^t\}_{j=1}^{N^t}$ の重心にソース標本に付随する
変数を輸送 $((\text{diag}(\gamma \mathbf{1}))^{-1} \gamma X^t, z^s)$

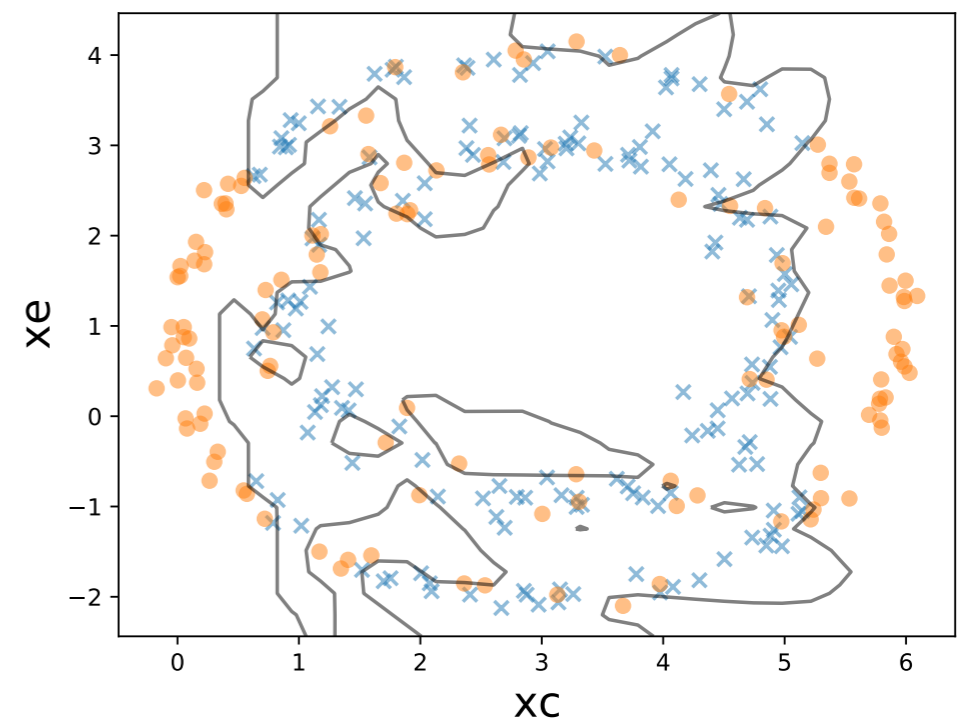
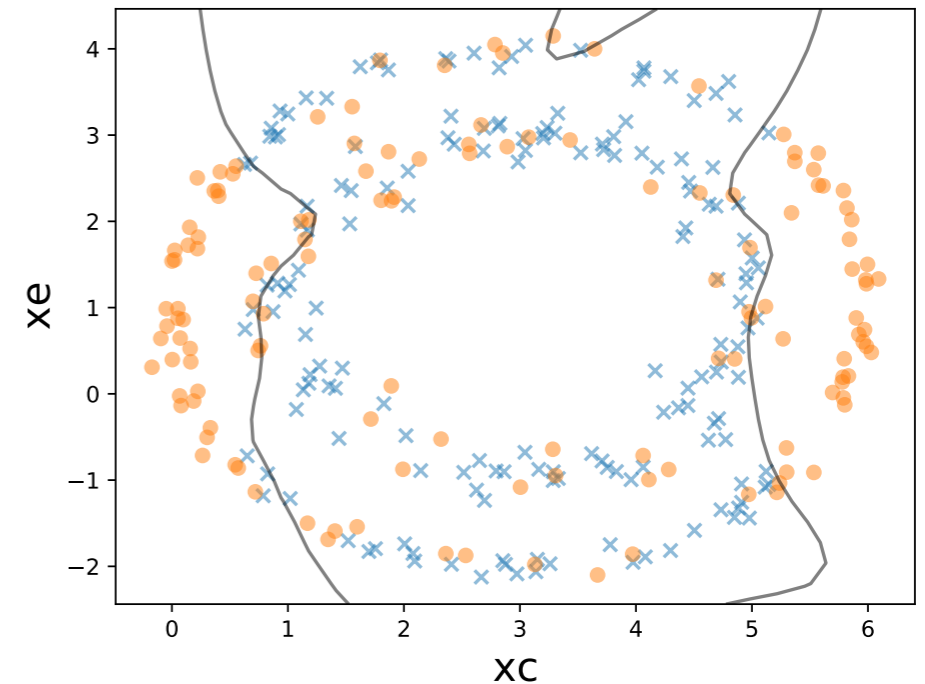


2. $\{x_j^t\}_{j=1}^{N^t}$ 上にソース標本の付随する変数の
重心を輸送 $(X^t, \text{diag}(\gamma^\top \mathbf{1}))^{-1} \gamma^\top z^s)$

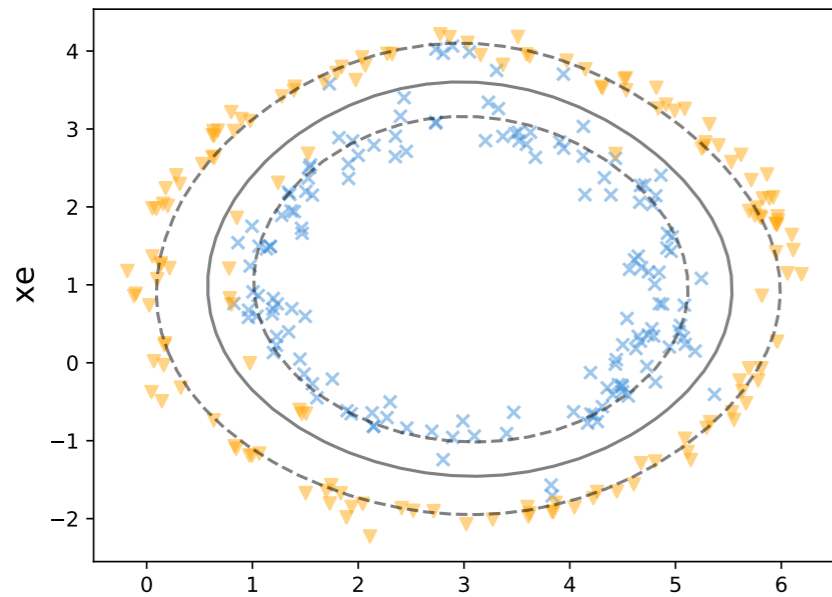


ニューラルネットワークを 用いた場合の実験

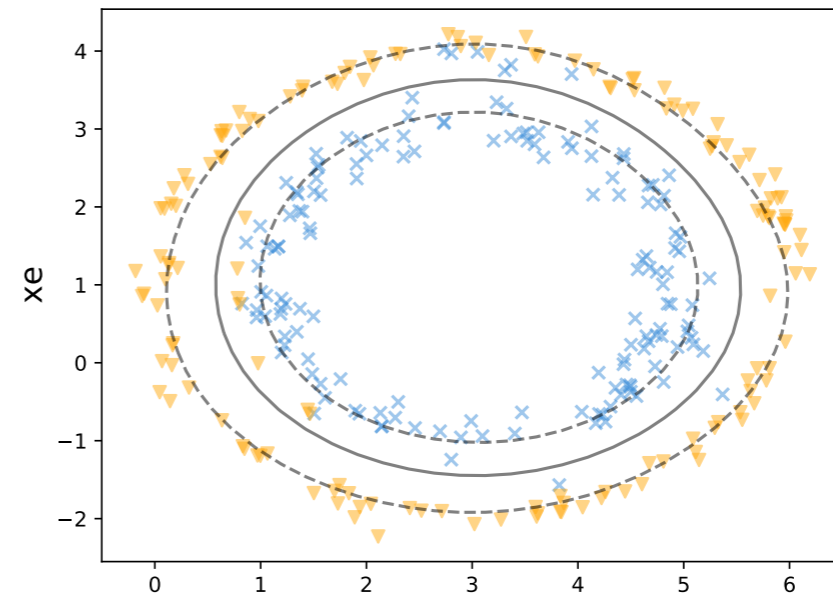
- 複雑なモデルを用いると
誤ったラベルへの過剰適合
が起こる
- 交互最適化によるラベル
予測精度の改善とモデル
学習が難しい



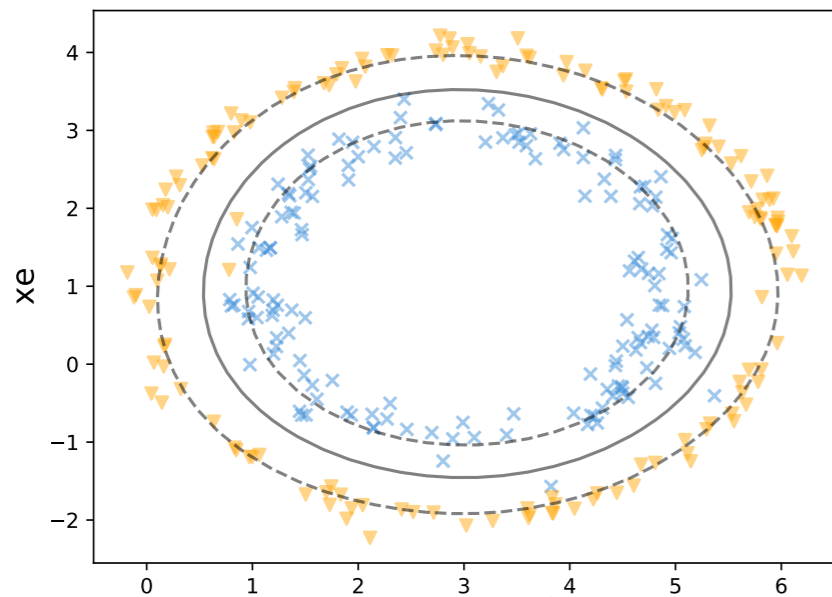
パラメータ α の影響



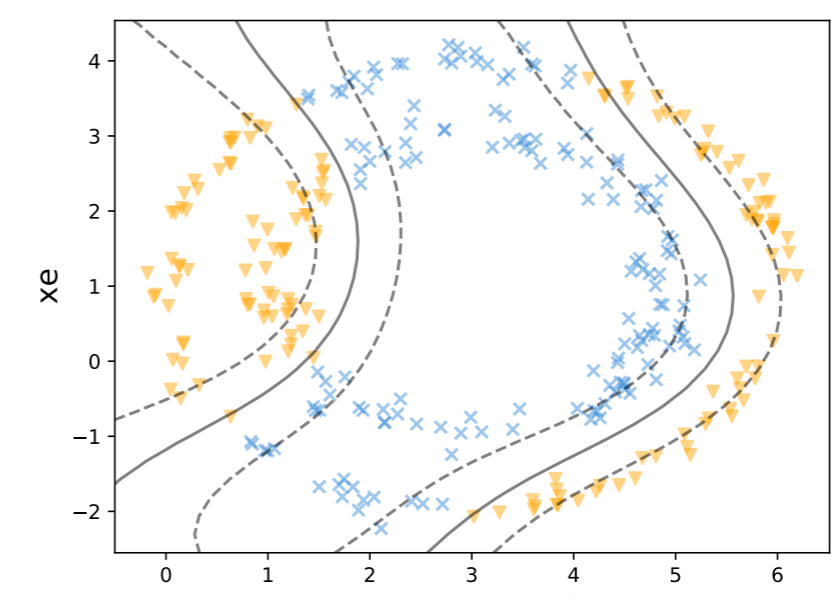
$$\alpha = 0.01$$



$$\alpha = 0.1$$



$$\alpha = 1$$



$$\alpha = 10$$

テスト誤差の評価

輸送に関する誤差の評価

- 仮説集合 \mathcal{F} に属する f について

ターゲットドメインでの誤差を評価する

$$\begin{aligned} \text{err}_T(f) &= \mathbb{E}_{(\mathbf{x}^c, \mathbf{x}^e, y) \sim \mathcal{P}_T} \mathcal{L}(y, f(\mathbf{x}^c, \mathbf{x}^e)) \\ \text{err}_{S'}(f) &= \mathbb{E}_{(\mathbf{x}^c, \mathbf{x}^e, y) \sim \mathcal{P}_{S'}} \mathcal{L}(y, f(\mathbf{x}^c, \mathbf{x}^e)) \end{aligned}$$

- 予測誤差関数 \mathcal{L} は対称で三角不等式を満たし

$$k\text{-リップシッツ連続} \quad |\mathcal{L}(a, b) - \mathcal{L}(a, c)| \leq k|b - c|$$

- $\text{err}_{S'}(f^*) + \text{err}_T(f^*)$ を最小化するモデル f^* が有界

$$|f^*(\mathbf{x}_s^c, \mathbf{x}_s^e) - f^*(\mathbf{x}_t^c, \mathbf{x}_t^e)| \leq M$$

- モデル f^* と $\mathcal{P}_{S'}$ から \mathcal{P}_T^f への最適輸送計画 Π^* が

ϕ -確率的転移リップシッツ性を満たす

$$\Pr_{((\mathbf{x}_s^c, \mathbf{x}_s^e), (\mathbf{x}_t^c, \mathbf{x}_t^e)) \sim \Pi^*} [|f^*(\mathbf{x}_s^c, \mathbf{x}_s^e) - f^*(\mathbf{x}_t^c, \mathbf{x}_t^e)| > \lambda d((\mathbf{x}_s^c, \mathbf{x}_s^e), (\mathbf{x}_t^c, \mathbf{x}_t^e))] \leq \phi(\lambda)$$

輸送に関する誤差の評価

$1 - \delta$ 以上の確率で

$$\begin{aligned} \text{err}_T(f) \leq & \underbrace{W_1(\hat{\mathcal{P}}_S, \hat{\mathcal{P}}_T^f)}_{\text{red}} + \sqrt{\frac{2}{c'} \log\left(\frac{2}{\delta}\right) \left(\frac{1}{\sqrt{N_S}} + \frac{1}{\sqrt{N_T}}\right)} \\ & + \underbrace{\text{err}_{S'}(f^*)}_{\text{blue}} + \underbrace{\text{err}_T(f^*)}_{\text{green}} + kM\phi(\lambda) \end{aligned}$$

- ソース分布 \mathcal{P}_S とターゲット分布 $\mathcal{P}_T(x^c, x^e, y)$ を近似する分布 $\mathcal{P}_T^f(x^c, x^e, f(x^c, x^e))$ 間の Wasserstein 距離
- 輸送コストは $\alpha d(x_s^c, x_t^c) + \mathcal{L}(y_s, f(x_t^c, x_t^e))$
- $\text{err}_{S'}(f^*) + \text{err}_T(f^*)$ を最小化するモデル f^* による誤差
- \mathcal{L} の連続性, 仮説 f^* の上界および (確率的な) 連続性によって決まる定数

先行研究の評価の利点と欠点

- 利点
 - 推定に用いる最適輸送問題の距離が上界に出現
 - 欠点
 - モデル f の学習が考慮されていない（モデルは任意）
 - ラベル輸送可能性とは別に，ソース分布とターゲットの推定分布の距離が離れると上界の評価が過大になる
 - 実際には考慮しないソース，ターゲット両方の誤差の和を最小化するモデル f^* の評価が上界に含まれる
- モデル f の学習を考慮した上界を導出

予測モデルのテスト誤差上界

- ターゲット変数の空間 Ω_t は正定値カーネル K を備え \mathcal{H} を K に対応する再生核ヒルベルト空間であるとする
- カーネルが有界 $\sup_{x \in \Omega_t} K(x, x) \leq \Lambda^2$
- モデル集合 $\mathcal{F} = \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq a\}$
- 損失関数 \mathcal{L} は以下を満たす
 - 対称性 $\mathcal{L}(y_1, y_2) = \mathcal{L}(y_2, y_1)$
 - 三角不等式 $\mathcal{L}(y_1, y_2) + \mathcal{L}(y_2, y_3) \geq \mathcal{L}(y_1, y_3)$
 - リプシッツ連続性 $|\mathcal{L}(y_1, y_2) - \mathcal{L}(y_1, y_3)| \leq k|y_2 - y_3|$
 - 有界性 $L_0 = \sup_{y \in \mathcal{C}} \mathcal{L}(0, y) < \infty$

予測モデルのテスト誤差上界

- 最適なモデル $f_0 \in \mathcal{F}_0$ は任意の $\mathbf{x}_1^c, \mathbf{x}_1^e, \mathbf{x}_2^c, \mathbf{x}_2^e$ に対して
$$|f_0(\mathbf{x}_1^c, \mathbf{x}_1^e) - f_0(\mathbf{x}_2^c, \mathbf{x}_2^e)| \leq M$$
- 最適なモデル f_0 と $\mathcal{P}_{\hat{T}}$ から \mathcal{P}_T への最適輸送計画
 π^* は ϕ -確率的転移リップシットツ性をみたす
 - $\mathcal{P}_{\hat{T}}$: ターゲット分布を近似する分布
提案法では最適輸送により推定される分布
 - \mathcal{P}_T : 真のターゲット分布

予測モデルのテスト誤差上界

- 前ページまでの仮定のもと

$\hat{f}_s \equiv \arg \min_{f \in \mathcal{F}} \widehat{\text{err}}_{\hat{T}}(f)$ とする

任意の $\lambda > 0$ および $\alpha = k\lambda$ のもとと任意の $\delta \in (0, 1)$ に対して少なくとも確率 $1 - \delta$ で

$$\begin{aligned} \text{err}_T(y, \hat{f}_s) &\leq \widehat{\text{err}}_{\hat{T}}(\hat{f}_s, y) + \frac{2ka\Lambda}{\sqrt{N_t}} + (L_0 + ka\Lambda) \sqrt{\frac{\ln(1/\delta)}{N_t}} \\ &\quad + W(\mathcal{P}_{\hat{T}}, \mathcal{P}_T) + 2 \text{err}_T(y, f_0) + kM\phi(\lambda) \end{aligned}$$

- $\widehat{\text{err}}_{\hat{T}}(f) = \mathbb{E}_{(\mathbf{x}^c, \mathbf{x}^e, y) \sim \mathcal{P}_{\hat{T}}} \mathcal{L}(y, f(\mathbf{x}^c, \mathbf{x}^e))$

予測モデルのテスト誤差上界

$$\begin{aligned} \text{err}_T(y, \hat{f}_s) &\leq \widehat{\text{err}}_{\hat{T}}(\hat{f}_s, y) + \frac{2ka\Lambda}{\sqrt{N_t}} + (L_0 + ka\Lambda) \sqrt{\frac{\ln(1/\delta)}{N_t}} \\ &\quad + W(\mathcal{P}_{\hat{T}}, \mathcal{P}_T) + 2 \text{err}_T(y, f_0) + kM\phi(\lambda) \end{aligned}$$

- 利点
 - 推定したテスト分布で学習したモデルに対する誤差の評価を行うことで上界がよりタイトになる
 - ソース, ターゲット共通のモデルの評価は不要
- 欠点
 - ターゲット分布の推定方法についての評価は行わない
 - Wasserstein距離は推定で考える輸送の距離とは異なる