Conditional Generative Learning from Invariant Representations in Multi-Source: Robustness and Efficiency

Guojun Zhu

School of Mathematical Sciences,
University of Chinese Academy of Sciences

Joint work with Sanguo Zhang(University of Chinese Academy of Sciences)
and Mingyang Ren(Shanghai Jiao Tong University)

November 30, 2024

Outline

Outline

Background

- Classical methods: based on Pre-trained Fine-tuning model.
- Transfer Learning, a rapidly growing field in statistics, offers a powerful approach to enhance data analysis, particularly when the target data is limited:
  - Generative Model[Damodaran et al., 2018]
  - Graph Model[Ren et al., 2023]
  - Linear Regression[Tian et al., 2023]
  - Semiparametric Regression[He et al., 2024]

Background

- Classical methods: based on Pre-trained Fine-tuning model.
- Transfer Learning, a rapidly growing field in statistics, offers a powerful approach to enhance data analysis, particularly when the target data is limited:
  - Generative Model[Damodaran et al., 2018]
  - Graph Model[Ren et al., 2023]
  - Linear Regression[Tian et al., 2023]
  - Semiparametric Regression[He et al., 2024]
- Theoretical Insights for GAN. More and more statisticians exploring the learning theory from diverse perspectives:
  - Approximation Error & Statistical Error Framwork [Huang et al., 2022]
  - WGAN for Regression [Song et al., 2023]
  - Adaptive WGAN Architecture [Tan et al., 2024]

Limitation

The Pre-trained Fine-tuning model has limitations in terms of theoretical guarantees and tabular medical data.

Limitation

The Pre-trained Fine-tuning model has limitations in terms of theoretical guarantees and tabular medical data.

- For traditional datasets, such as tabular medical data, pre-trained models are not exist...
  Additionally, there is a lack of large-scale datasets for pre-training.
- Fine-tuning Pre-trained model introduces theoretical challenges...
  The complex adjustments required to align the generator and discriminator make it difficult to derive rigorous theoretical guarantees.

Limitation

The Pre-trained Fine-tuning model has limitations in terms of theoretical guarantees and tabular medical data.

- For traditional datasets, such as tabular medical data, pre-trained models are not exist...
  Additionally, there is a lack of large-scale datasets for pre-training.
- Fine-tuning Pre-trained model introduces theoretical challenges...
  The complex adjustments required to align the generator and discriminator make it difficult to derive rigorous theoretical guarantees.

We need to establish a new method that does not rely on pre-trained fine-tuning models!

Outline

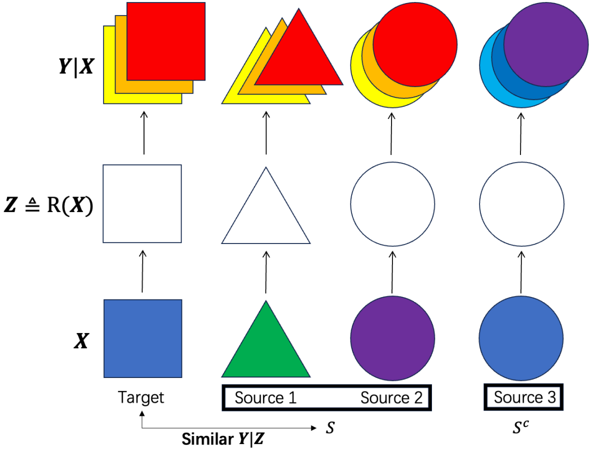
Motivation

Can we learn the "good" representation to effectively utilize the source data?

- What is the "good" representation?
  - **Domain Invariance.** To reduce the distribution discrepancy between the source and target domains.
  - **Dimension Reduction.** To retain all the necessary information for learning the conditional distribution

Motivation

- Why we need the "good" representation?



Motivation

- How to learn the "good" representation?

Domain Adaptation. Similar challenges have been addressed with a number of well-developed and effective methods in classification task.

- **Asymmetric.** Transform the features of the source domain to match those of the target domain.[Hoffman et al., 2014, Courty et al., 2017]
- **Symmetric.** Project both domains into a shared latent space, aligning their distributions.[Damodaran et al., 2018, Shen et al., 2018]

Outline

Data and Modeling Framework

- **Data.** T Sources $\left\{\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)}\right\}_{i=1}^{t}$ and Target $\left\{\boldsymbol{x}_i^{(0)}, \boldsymbol{y}_i^{(0)}\right\}_{i=1}^{(0)}$
- **Idea.** Find Domain Invariant representation $R : \mathcal{X} \mapsto Z$
- **Similarity Measure.** Based on the IPM metric,

$$d_{\mathcal{F}_B^1}\left(P_{Y|Z}^{(t)}, P_{Y|Z}^{(0)}\right) = \sup_{f \in \mathcal{F}_B^1}\left\{\mathbb{E}_{P_{Y|Z}^{(t)}} f(\boldsymbol{y}) - \mathbb{E}_{P_{Y|Z}^{(0)}} f(\boldsymbol{y})\right\},$$

Data and Modeling Framework

- **Data.** T Sources $\left\{\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i^{(t)}\right\}_{i=1}^{t}$ and Target $\left\{\boldsymbol{x}_i^{(0)}, \boldsymbol{y}_i^{(0)}\right\}_{i=1}^{(0)}$
- **Idea.** Find Domain Invariant representation $R : \mathcal{X} \mapsto Z$
- **Similarity Measure.** Based on the IPM metric,

$$d_{\mathcal{F}_B^1}\left(P_{Y|Z}^{(t)}, P_{Y|Z}^{(0)}\right) = \sup_{f \in \mathcal{F}_B^1}\left\{\mathbb{E}_{P_{Y|Z}^{(t)}} f(\boldsymbol{y}) - \mathbb{E}_{P_{Y|Z}^{(0)}} f(\boldsymbol{y})\right\},$$

- **Reliable Source S.** To keep robustness, we consider the threshold h,

$$\forall t \in S, \mathbb{E}_{P_Z^{(t)}} d_{\mathcal{F}_B^1}\left(P_{Y|Z}^{(t)}, P_{Y|Z}^{(0)}\right) \leq h.$$

- **Goal.** Finding a generation function $G(\eta, \boldsymbol{z})$, such that

$$G(\eta, \mathbf{z}) \sim P_{Y|Z=\mathbf{z}}^{(0)}, \mathbf{z} \in Z.$$

- *Question. How to estimate reliable source S and make full use of them?*

Error Decomposition

- **If S is known**, we can consider pool-training.
  Decompose error $d_{\mathcal{F}_B^1}\left(P_{\hat{Z},\hat{G}}, P_{\hat{Z},Y}^{(0)}\right)$, where $P_{\hat{Z},Y} = \sum_{t\in S\cup\{0\}} \frac{\mathrm{n}_t}{n} P_{\hat{Z},Y}^{(t)}$
  - **Learning bias**. $\mathrm{d}_{\mathcal{F}_\mathrm{B}^1}\left(\mathrm{P}_{\hat{\mathrm{Z}},\hat{\mathrm{G}}}, \mathrm{P}_{\mathrm{Z},\mathrm{Y}}\right)$,
    We focus on $\hat{R}$, $\hat{G}$ by Conditional WGAN.
  - **Transfer bias**. $d_{\mathcal{F}_B^1}\left(P_{\hat{Z},Y}, P_{\hat{Z},Y}^{(0)}\right)$
    We focus on $\hat{R}$ by Optimal Transport Regularization.
  - Combine the two parts, we get the loss:

$$\mathcal{L}_1(R,G,D;S) = \mathbb{E}_{P_X P_\eta} D(X, G(\eta, R(X))) - \mathbb{E}_{P_{X,Y}} D(X,Y) + \sum_{t\in S} \lambda_t \inf_\gamma \int \left\| \left(R\left(X^{(t)}\right), Y^{(t)}\right) - \left(R\left(X^{(0)}\right), Y^{(0)}\right) \right\|_1 d\gamma.$$

Error Decomposition

- **If S is known**, we can consider pool-training.
  Decompose error $d_{\mathcal{F}_B^1}\left(P_{\hat{Z},\hat{G}}, P^{(0)}_{\hat{Z},Y}\right)$, where $P_{\hat{Z},Y} = \sum_{t\in S\cup\{0\}} \frac{\mathrm{n}_t}{n} P^{(t)}_{\hat{Z},Y}$
  - **Learning bias**. $\mathrm{d}_{\mathcal{F}_\mathrm{B}^1}\left(\mathrm{P}_{\hat{\mathrm{Z}},\hat{\mathrm{G}}}, \mathrm{P}_{\mathrm{Z},\mathrm{Y}}\right)$,
    We focus on $\hat{R}$, $\hat{G}$ by Conditional WGAN.
  - **Transfer bias**. $d_{\mathcal{F}_B^1}\left(P_{\hat{Z},Y}, P^{(0)}_{\hat{Z},Y}\right)$
    We focus on $\hat{R}$ by Optimal Transport Regularization.
  - Combine the two parts, we get the loss:

$$\mathcal{L}_1(R,G,D;S) = \mathbb{E}_{P_X P_\eta} D(X, G(\eta, R(X))) - \mathbb{E}_{P_{X,Y}} D(X,Y)$$
$$+ \sum_{t\in S} \lambda_t \inf_\gamma \int \left\| \left(R\left(X^{(t)}\right), Y^{(t)}\right) - \left(R\left(X^{(0)}\right), Y^{(0)}\right) \right\|_1 d\gamma.$$

- **If S is unknown**, *can we develop a data-driven selection criterion to estimate the subset S?*

Data-driven Selection Criterion

- **If S is unknown**, *can we develop a data-driven selection criterion to estimate the subset S?*

Pre-train full model to learn the "rough" domain invariant representation $\tilde{Z}$

$$\hat{S} = \left\{ t : W_1 \left(P^{n_t}_{\tilde{Z},Y'}, P^{n_0}_{\tilde{Z},Y} \right) \leq C \left(\max \left\{ n^{-1/(r+q)}, n_0^{-1/r} \right\} \right) \right\}$$

Outline

Notations

- $\boldsymbol{x}_i^{(t)} \in \mathcal{X} \subset \mathbb{R}^d$ is drawn according to distribution $P_X^{(t)}$ over $\mathcal{X}$, and then $\boldsymbol{y}_i^{(t)} \in \mathcal{Y} \subset \mathbb{R}^q$ is drawn according to the conditional distribution $P^{(t)}_{Y|X=\boldsymbol{x}_i^{(t)}}$, $t \in [T]$. Besides, we assume a low-dimensional subspace $\mathcal{Z} \subset \mathbb{R}^r$, $r << d$.
- The overall architecture of the network is characterized by its width, denoted as $W = \max\{p_1, \ldots, p_H\}$, and its depth, represented by H. To facilitate discussion, we denote a neural network with input dimension p_0, output dimension p_{H+1}, a maximum width of W, and a maximum depth of H as $\mathcal{NN}(p_0, p_{H+1}, W, H)$.

Notations

- For the generator network class G_θ : Let $\mathcal{G} \equiv NN\left(r+m, q, W_G, H_G\right)$ be a class of ReLU-activated FNNs, $G_\theta : \mathbb{R}^{d+m} \rightarrow \mathbb{R}^{\mathrm{q}}$, with parameter $\boldsymbol{\theta}$, width W_G, and depth H_G.
- For the discriminator network class D_ϕ : Let $\mathcal{D} \equiv \mathcal{N}\mathcal{N}\left(r+q, 1, W_D, H_D\right) \cap \mathcal{F}^1_{\text{Lip}}$ be a class of ReLU-activated FNNs, $f_\delta : \Omega \rightarrow \mathbb{R}$, with parameter ϕ, width $W_{\mathcal{D}}$, and depth $H_{\mathcal{D}}$.
- For the representation network class R_ω : Let $\mathcal{R} \equiv \mathcal{N}\mathcal{N}\left(d, r, W_R, H_R\right)$ be a class of ReLU-activated FNNs, $R_\omega : \mathbb{R}^d \rightarrow \mathbb{R}^r$, with parameter ω, width W_R, and depth H_R.

Conditions and Assumptions

- **Assumption 1.** The similarity measure between the outlier sources and the target domain is assumed to be of a much larger order than h,

$$\forall t \in S^c, \mathbb{E}_{P^{(t)}_{\tilde{Z}}} d_{\mathcal{F}^1_B}\left(P^{(t)}_{Y|\tilde{Z}}, P^{(0)}_{Y|\tilde{Z}}\right) = O(h^\alpha), \alpha > 1.$$

- **Assumption 2.** For some $\delta > 0, (\hat{Z}, Y)$ satisfies the first-order moment tail condition, for any $n \geq 1$,

$$\mathbb{E}\left[\|(\hat{Z}, Y)\|\mathbb{I}_{\{\|(\hat{Z}, Y)\| > \log n\}}\right] = O\left(n^{-(\log n)^\delta/(r+q)}\right).$$

- **Assumption 3.** The noise distribution P_η is absolutely continuous with respect to the Lebesgue measure.

- **Assumption 4.** The IPM distance between the conditional distributions of reliable source domains and the target domain is bounded in expectation, for some $h = O\left(\max\left\{n^{-1/r+q}, n_0^{-1/r}\right\}\right)$,

$$\forall t \in S \subset [T], \quad \mathbb{E}_{P^{(t)}_{\hat{Z}}} d_{\mathcal{F}^1_B}\left(P^{(t)}_{Y|\hat{Z}}, P^{(0)}_{Y|\hat{Z}}\right) \leq h.$$

Conditions and Assumptions

- **Assumption 5.** The conditional distribution of the target domain satisfies a certain Lipschitz condition under the Total Variation (TV) distance, for some $K > 1$:

$$\forall \hat{\boldsymbol{z}}_1, \hat{\boldsymbol{z}}_2, d_{TV}\left(P^{(0)}_{Y|\hat{Z}=\hat{z}_1}, P^{(0)}_{Y|\hat{Z}=\hat{z}_2}\right) \leq \frac{K-1}{2B}\left\|\hat{\boldsymbol{z}}_1 - \hat{\boldsymbol{z}}_2\right\|_1,$$

  where $d_{TV}(\cdot,\cdot)$ is the TV distance.

- **Assumption 6.** The conditional distribution of the target domain satisfies a certain Lipschitz condition under the 1-Wasserstein distance, for some $K > 1$:

$$\forall \tilde{\boldsymbol{z}}_1, \tilde{\boldsymbol{z}}_2 \in \mathcal{Z}, W_1\left(P^{(0)}_{Y|\tilde{Z}=\tilde{z}_1}, P^{(0)}_{Y|\tilde{Z}=\tilde{z}_2}\right) \leq (K-1)\left\|\tilde{\boldsymbol{z}}_1 - \tilde{\boldsymbol{z}}_2\right\|_1,$$

  where $W_1(\cdot,\cdot)$ is the 1-Wasserstein distance.

Statistical Properties

If S is known

Theorem 1 Suppose Assumptions $2-5$ hold. Let (W_D, H_D) of $D_{\boldsymbol{\phi}}$, (W_G, H_G) of $G_{\boldsymbol{\theta}}$ and (W_R, H_R) of $R_{\boldsymbol{\omega}}$ be specified such that $W_D H_D = \lceil \sqrt{n} \rceil$, $W_G^2 H_G = c_1 qn$. and $W_R^2 H_R = c_2$ rn for some constants $12 \leq c_1, c_2 \leq 384$. Let $n = \sum_{t \in S \cup \{0\}} n_t$, we have:

$$\mathbb{E}_{\hat{G}} d_{\mathcal{F}_B^1} \left(P_{\hat{Z}, \hat{G}}, P_{\hat{Z}, Y}^{(0)} \right) \lesssim n^{-1/(r+q)} \log n + n_0^{-1/r}.$$

Statistical Properties

If S is known

Theorem 1 Suppose Assumptions $2-5$ hold. Let (W_D, H_D) of $D_{\boldsymbol{\phi}}$, (W_G, H_G) of $G_{\boldsymbol{\theta}}$ and (W_R, H_R) of $R_{\boldsymbol{\omega}}$ be specified such that $W_D H_D = \lceil \sqrt{n} \rceil$, $W_G^2 H_G = c_1 qn.$ and $W_R^2 H_R = c_2$ rn for some constants $12 \leq c_1, c_2 \leq 384$. Let $n = \sum_{t \in S \cup \{0\}} n_t$, we have:

$$\mathbb{E}_{\hat{G}} d_{\mathcal{F}_B^1} \left(P_{\hat{Z}, \hat{G}}, P_{\hat{Z}, Y}^{(0)} \right) \lesssim n^{-1/(r+q)} \log n + n_0^{-1/r}.$$

When $P_{Z,Y}$ has a bounded support, we can drop the logarithm factor.

Theorem 2 Suppose that $P_{\hat{Z},Y}$ is supported on $[-U, U]^{r+q}$ for some $U > 0$ and Assumptions 3-5 hold. (W_D, H_D) of $D_{\boldsymbol{\phi}}$, (W_G, H_G) of $G_{\boldsymbol{\theta}}$ and (W_R, H_R) of $R_{\boldsymbol{\omega}}$ be specified such that $W_D H_D = \lceil \sqrt{n} \rceil$, $W_G^2 H_G = c_1 qn$ and $W_R^2 H_R = c_2 rn$ for some constants $12 \leq c_1, c_2 \leq 384$. Let the output of $G_{\boldsymbol{\theta}}$ be on $[-U, U]^q$. Let $n = \sum_{t \in S \cup \{0\}} n_t$, we have:

$$\mathbb{E}_{\hat{G}} d_{\mathcal{F}_B^1} \left(P_{\hat{Z}, \hat{G}}, P_{\hat{Z}, Y}^{(0)} \right) \lesssim n^{-1/(r+q)} + n_0^{-1/r}.$$

Statistical Properties

If S is known

Theorem 2 Suppose that $P_{\hat{Z},Y}$ is supported on $[-U, U]^{r+q}$ for some $U > 0$ and Assumptions 3-5 hold. (W_D, H_D) of $D_{\boldsymbol{\phi}}$, (W_G, H_G) of $G_{\boldsymbol{\theta}}$ and (W_R, H_R) of $R_{\boldsymbol{\omega}}$ be specified such that $W_D H_D = \lceil \sqrt{n} \rceil$, $W_G^2 H_G = c_1 qn$ and $W_R^2 H_R = c_2 rn$ for some constants $12 \leq c_1, c_2 \leq 384$. Let the output of $G_{\boldsymbol{\theta}}$ be on $[-U, U]^q$. Let $n = \sum_{t \in S \cup \{0\}} n_t$, we have:

$$\mathbb{E}_{\hat{G}} d_{\mathcal{F}_B^1} \left(P_{\hat{Z},\hat{G}}, P_{\hat{Z},Y}^{(0)} \right) \lesssim n^{-1/(r+q)} + n_0^{-1/r}.$$

Corollary 3 Under the conditions of Theorem 2 we have

$$\mathbb{E}_{\hat{G}} \mathbb{E}_{P_{\hat{Z}}^{(0)}} d_{\mathcal{F}_B^1} \left(P_{\hat{G}|\hat{Z}}, P_{Y|\hat{Z}}^{(0)} \right) \lesssim n^{-1/(r+q)} + n_0^{-1/r}.$$

Statistical Properties

If S is unknown

Theorem 4 Suppose that $P_{\tilde{Z},Y}, P_{\hat{Z},Y}$ is supported on $[-U, U]^{r+q}$ for some $U > 0$ and Assumptions 1,3-6 hold. Let $\left(W_{\tilde{D}}, H_{\tilde{D}}\right)$ of $D_{\tilde{\phi}}$, $\left(W_{\tilde{G}}, H_{\tilde{G}}\right)$ of $G_{\tilde{\theta}}$ and $\left(W_{\tilde{R}}, H_{\tilde{R}}\right)$ of $R_{\tilde{\omega}}$ be specified such that $W_{\tilde{D}} H_{\tilde{D}} = \lceil \sqrt{n} \rceil$, $W_{\tilde{G}}^2 H_{\tilde{G}} = c_1 qn$ and $W_{\tilde{R}}^2 H_{\tilde{R}} = c_2 rn$ for some constants $12 \leq c_1, c_2 \leq 384$, where $n = \sum_{t \in [T] \cup \{0\}} n_t$. Let the output of $G_{\tilde{\boldsymbol{\theta}}}$ be on $[-U, U]^q$ and threshold $C\left(\max\left\{n^{-1/(r+q)}, n_0^{-1/r}\right\}\right) = h$, we have:

$$\mathbb{E}_{\hat{G}} \mathbb{E}_{P_{\hat{Z}}^{(0)}} d_{\mathcal{F}_B^1}\left(P_{\hat{G}|\hat{Z}}, P_{Y|\hat{Z}}^{(0)}\right) \precsim n^{-1/(r+q)} + n_0^{-1/(r+q)}.$$

Outline

Simulation data: conditional density estimation

- For methods to compare, consider:
  - **S**elected **T**ransfer-**WGAN**(**STWGAN**): The method we proposed,
  - **T**arget-**O**nly(**TO**): a method trained exclusively on the target domain without representation learning,
  - **Pool**: an ablation variant where $\lambda_t = 0$,
  - **Pr**e-**t**rained **F**ine-**t**uning(**PT-FT**): pretrained on source data.
- The samples are generated as:
  - **Model 1 (M1).** A nonlinear model:

$$Y = X_1 + \exp\left(X_2 + X_3/3\right) + \sin\left(X_4 + X_5\right) + \varepsilon,$$

    where $\varepsilon \sim N(0, X_1^2)$.
  - **Model 2 (M2).** A model with a multiplicative error:

$$Y = (2 + X_1^2/3 + X_2^2 + X_3^2 + X_4^2 + X_5^2)/3 \times \varepsilon,$$

    where $\varepsilon \sim N(X_3, 1)$.
  - **Model 3 (M3).** A mixture model:

$$\begin{aligned} Y =& \mathbb{I}_{\{U \leq 1/3\}} N\left(-3 - X_1/3 - X_2^2, 0.25\right) \\ &+ \mathbb{I}_{\{U > 1/3\}} N\left(3 + X_1/3 + X_2^2, 1\right), \end{aligned}$$

Simulation data: conditional density estimation

The covariate vector X is generated from $N(\boldsymbol{\mu}^{(t)}, \mathbf{I}_{100})$ in the t-th domain. So the ambient dimension of X is 100, but (M1) and (M2) only depend on the first 5 components of X and (M3) only depends on the first 2 components of X.

Table: The value of $\mu^{(t)}$, where the index (t) represents the domain, with (0) denoting the target domain.

(t)	$\boldsymbol{\mu}^{(t)}$	posterior drift
(0)	$(2, 1, 0, \ldots, 0)^\top$	-
(1)	$(0, 0, 0, \ldots, 0)^\top$	No
(2)	$(5, 5, 5, \ldots, 5)^\top$	No
(3)	$(-5, \ldots, -5)^\top$	No
(4)	$(2, 1, 0, \ldots, 0)^\top$	Yes
(5)	$(2, 1, 0, \ldots, 0)^\top$	Yes

Simulation data: conditional density estimation

We consider the posterior drift in the fourth and fifth source domains across different data generation models.

- **Model 1 (M1).** A nonlinear model with an additive error term:

$$(4): \quad Y = 5X_1 + \exp\left(X_2 + X_3/3 + 2\right) + \cos\left(X_4 + X_5\right) + \varepsilon + 5,$$
$$(5): \quad Y = X_1/5 + \exp\left(X_2 + X_3/3 - 2\right) + \cos\left(X_4 + X_5\right) + \varepsilon - 5,$$

where $\varepsilon \sim N(0, X_1^2)$.

- **Model 2 (M2).** A model with a multiplicative Gaussian error term:

$$(4): \quad Y = (7 + X_1^3/3 + X_2^3 + X_3^3 + X_4^3 + X_5^3) \times \varepsilon + 5,$$
$$(5): \quad Y = (-3 + X_1 + X_2 + X_3 + X_4 + X_5) \times \varepsilon - 5,$$

where $\varepsilon \sim N(X_3, 1)$

- **Model 3 (M3).** A mixture of two normal distributions:

$$(4): \quad Y = \mathbb{I}_{\{U \leq 1/3\}} N\left(-8 - X_1^3 - X_2, 0.25\right) + \mathbb{I}_{\{U > 1/3\}} N\left(8 + X_1^3 + X_2, 1\right),$$
$$(5): \quad Y = \mathbb{I}_{\{U \leq 1/3\}} N\left(2 - X_1 - X_2, 0.25\right) + \mathbb{I}_{\{U > 1/3\}} N\left(-2 + X_1 + X_2, 1\right),$$

where $U \sim \text{Unif}(0, 1)$ and is independent of X.

Evaluation

Similar to the experiments conducted by Liu et al. [2021], Zhou et al. [2023], we consider the **mean squared error (MSE)** of the estimated conditional mean $\mathbb{E}(Y \mid X)$ and the estimated conditional standard deviation $\mathrm{SD}(Y \mid X)$.
We use a test data set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K\}$ of size $K = 2000$. For the proposed method, we first generate $J = 10000$ samples $\{\eta_j : j = 1, \ldots, J\}$ from the reference distribution P_η and calculate conditional samples $\left\{\hat{G}\left(\eta_j, \boldsymbol{x}_k\right), j \in [J], k \in [K]\right\}$.

- The MSE of the estimated conditional mean:

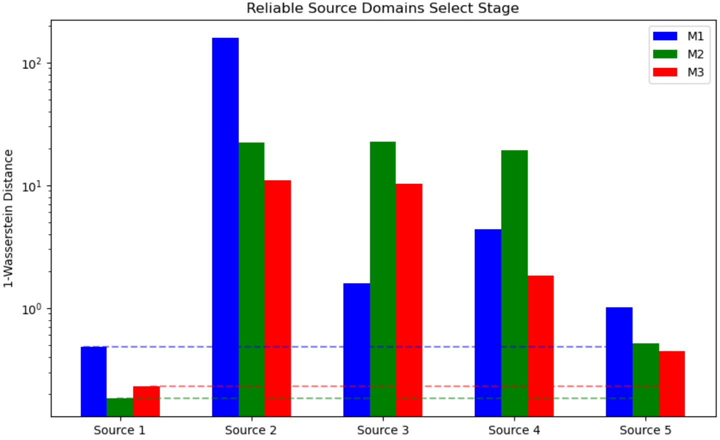
$$\mathrm{MSE(mean)} = (1/K)\sum_{k=1}^{K}\{\widehat{\mathbb{E}}\left(Y \mid X = \boldsymbol{x}_k\right) - \mathbb{E}\left(Y \mid X = \boldsymbol{x}_k\right)\}^2,$$

- The MSE of the estimated conditional standard deviation:

$$\mathrm{MSE(sd)} = (1/K)\sum_{k=1}^{K}\{\widehat{\mathrm{SD}}(Y \mid X = \boldsymbol{x}_k) - \mathrm{SD}(Y \mid X = \boldsymbol{x}_k)\}^2.$$

Selection results

In all three data simulated models, the first source domain is considered a reliable source domain, while the others are identified as outlier source domains.



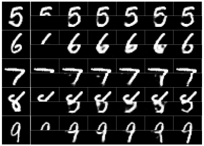
Simulation results

we considered different sample sizes for the source domains while keeping $n_0 = 10,000$ fixed, as shown in Table.

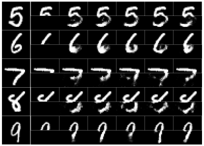
			STWGAN	Target-only	Pool	PT-FT
$n_t = 20,000$	M1	Mean	**15.77**(1.29)	21.49(1.24)	16.90(1.63)	77.87(11.27)
		SD	4.43(1.48)	8.21(2.84)	**1.89**(0.45)	2.17(1.22)
	M2	Mean	4.40(1.10)	9.51(3.63)	6.75(2.35)	**3.83**(2.82)
		SD	1.95(0.30)	**1.39**(0.14)	1.84(0.18)	2.08(0.33)
	M3	Mean	**2.22**(0.99)	25.75(4.10)	3.07(1.42)	3.07(1.02)
		SD	**0.47**(0.10)	10.14(5.20)	0.75(0.10)	9.94(1.69)
$n_t = 40,000$	M1	Mean	**10.64**(2.07)	17.06(1.91)	16.94(2.94)	81.76(11.55)
		SD	6.69(4.47)	7.69(3.33)	**1.37**(0.22)	1.66(0.46)
	M2	Mean	**3.12**(1.14)	7.10(3.01)	5.15(1.77)	5.01(2.81)
		SD	1.90(0.38)	**1.53**(0.23)	2.10(0.34)	2.67(1.04)
	M3	Mean	**2.09**(1.39)	26.89(7.13)	2.32(1.55)	2.96(0.82)
		SD	**0.54**(0.13)	7.72(4.06)	0.61(0.51)	8.09(2.72)
$n_t = 60,000$	M1	Mean	**10.73**(1.16)	24.40(2.84)	17.56(1.69)	85.43(14.43)
		SD	2.84(1.59)	9.84(2.56)	1.61(0.56)	**1.60**(0.81)
	M2	Mean	**2.22**(1.30)	7.73(4.01)	6.96(1.42)	5.27(3.10)
		SD	2.37(1.16)	**1.51**(0.20)	2.05(0.13)	3.46(1.36)
	M3	Mean	**1.68**(1.34)	20.97(3.40)	2.32(1.55)	2.66(1.03)
		SD	**0.56**(0.06)	5.67(2.67)	0.61(0.51)	8.41(2.19)

Outline

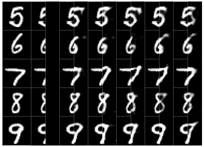
Image reconstruction: MNIST dataset



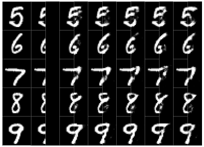
(a) STWGAN



(b) Target Only



(c) STWGAN



(d) Target Only

Figure: Comparison of STWGAN and Target Only: (a) and (b) show results of upper2lower, while figures (c) and (d) show results of left2right.

Image-to-Image translation

(a) shoes2edges (b) edges2shoes

Summary and Discussion

We proposed **S**elected **T**ransfer-**WGAN(STWGAN)**, a robust transfer approach designed to address the challenges of multi-source conditional generation models. This is achieved through a two-stage training process that maintains the training stability of WGAN. Our algorithm does not rely on pre-trained models from large datasets and provides both non-asymptotic error bounds and asymptotic guarantees.

Future work will discuss how neural networks can learn complex dimensionality reduction structures and retain useful information.

THANKS!

Reference

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in neural information processing systems*, 30, 2017.

Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 447–463, 2018.

Baihua He, Huihang Liu, Xinyu Zhang, and Jian Huang. Representation transfer learning for semiparametric regression. *arXiv preprint arXiv:2406.13197*, 2024.

Judy Hoffman, Erik Rodner, Jeff Donahue, Brian Kulis, and Kate Saenko. Asymmetric and category invariant feature transformations for domain adaptation. *International journal of computer vision*, 109:28–41, 2014.

Jian Huang, Yuling Jiao, Zhen Li, Shiao Liu, Yang Wang, and Yunfei Yang. An error analysis of generative adversarial networks for learning distributions. *Journal of machine learning research*, 23(116):1–43, 2022.

Shiao Liu, Xingyu Zhou, Yuling Jiao, and Jian Huang. Wasserstein generative learning of conditional distribution. *arXiv preprint arXiv:2112.10039*, 2021.

Mingyang Ren, Xin He, and Junhui Wang. Structural transfer learning of non-gaussian dag. *arXiv preprint arXiv:2310.10239*, 2023.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Shanshan Song, Tong Wang, Guohao Shen, Yuanyuan Lin, and Jian Huang. Wasserstein generative regression. *arXiv preprint arXiv:2306.15163*, 2023.

Zhiyao Tan, Ling Zhou, and Huazhen Lin. Generative adversarial learning with optimal input dimension and its adaptive generator architecture. *arXiv preprint arXiv:2405.03723*, 2024.

Ye Tian, Yuqi Gu, and Yang Feng. Learning from similar linear representations: Adaptivity, minimaxity, and robustness. *arXiv preprint arXiv:2303.17765*, 2023.

Xingyu Zhou, Yuling Jiao, Jin Liu, and Jian Huang. A deep generative approach to conditional sampling. *Journal of the American Statistical Association*, 118(543): 1837–1848, 2023.