Designing Ambiguity Sets for Distributionally Robust Optimization Using Structural Causal Optimal Transport

Ahmad-Reza Ehyaei [1], **Golnoosh Farnadi** [2] **Samira Samadi** [1]
[1] Max Planck Institute for Intelligent Systems, Tübingen AI Center, Germany
[2] Mila - Québec AI Institute, Université de Montréal
Montréal, Canada
ahmadreza.ehyaei@uni-tuebingen.de, farnadig@mila.quebec, ssamadi@tuebingen.mpg.de

Abstract

Distributionally robust optimization tackles out-of-sample issues like overfitting and distribution shifts by adopting an adversarial approach over a range of possible data distributions, known as the ambiguity set. To balance conservatism and accuracy, these sets must include realistic probability distributions by leveraging information from the nominal distribution. Assuming that nominal distributions arise from a structural causal model with a directed acyclic graph $\mathcal{G}$ and structural equations, previous methods such as adapted and $\mathcal{G}$-causal optimal transport have only utilized causal graph information in designing ambiguity sets. In this work, we propose incorporating structural equations, which include causal graph information, to enhance ambiguity sets, resulting in more realistic distributions. We introduce structural causal optimal transport and its associated ambiguity set, demonstrating their advantages and connections to previous methods. A key benefit of our approach is a relaxed version, where a regularization term replaces the complex causal constraints, enabling an efficient algorithm via difference-of-convex programming to solve structural causal optimal transport. We also show that when structural information is absent and must be estimated, our approach remains effective and provides finite sample guarantees. Lastly, we address the radius of ambiguity sets, illustrating how our method overcomes the curse of dimensionality in optimal transport problems, achieving faster shrinkage with dimension-free order.

Introduction

Distributionally Robust Optimization (DRO) is a data-driven framework designed to address out-of-sample challenges, such as distribution overfitting and distributional shifts, by minimizing potential discrepancies between in-sample expected loss and out-of-sample expected loss. DRO achieves this by defining a distributional ambiguity set (DAS) that encompasses a range of possible data distributions around the estimated true probability, ensuring that the DAS contains the unknown true underlying distribution with certainty or at least with high confidence. To guarantee the model's performance over out-of-sample distributions, DRO employs an adversarial approach that minimizes the worst-case loss to identify the optimal model (Blanchet et al. 2024). Ambiguity sets are typically categorized into two groups: discrepancy-based and moment-based. Discrepancy-based sets include distributions close to a nominal distribution according to a discrepancy measure, while moment-based sets include distributions whose moments satisfy certain properties (Rahimian and Mehrotra 2022). Among discrepancy-based sets, the Wasserstein distance is often preferred, as it quantifies the discrepancy by the minimal transportation cost, ensuring computational tractability through strong dual formulations and convergence guarantees. Additionally, Wasserstein ambiguity sets are robust against outliers and can handle both continuous and discrete distributions, unlike the KL divergence ball, which is limited to discrete distributions (Guo, Hong, and Yang 2017).

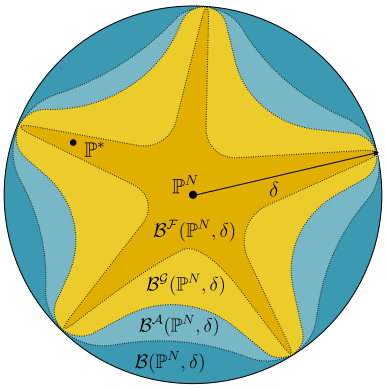

Figure 1: The underlying distribution $\mathbb{P}^*$ originates from a causal structure. The dark blue region represents $\mathcal{B}(\mathbb{P}, \delta)$ the ambiguity set for the classical OT. The light blue region corresponds to $\mathcal{B}^{\mathcal{A}}(\mathbb{P}, \delta)$ the DAS for adapted optimal transport. The light yellow region denotes $\mathcal{B}^{\mathcal{G}}(\mathbb{P}, \delta)$ the DAS for $\mathcal{G}$-causal OT, and the dark yellow region represents $\mathcal{B}^{\mathcal{F}}(\mathbb{P}, \delta)$ the DAS for our structural causal OT with diameter δ.

In designing the DAS two points need to be considered (Rahimian and Mehrotra 2022):

P1. What distributional information should the DAS include?

P2. How large should the radius of the DAS?

To understand the significance of **P1**, consider the corresponding DAS of classical Wasserstein DRO. It includes all probability distributions within a specified Wasserstein

Optimal Transport Variant	Information Utilization	Wasserstein Distance	Ambiguity Set
Classical (Villani et al. 2009; Peyré, Cuturi et al. 2017; Ambrosio et al. 2021)	No Constraints	$W(\mathbb{P}, \mathbb{Q})$	$\mathcal{B}(\mathbb{P}, \delta)$
Adopted (Backhoff et al. 2017; Lassalle 2018; Xu et al. 2020)	Preserves Causal Order	$W^{\mathcal{A}}(\mathbb{P}, \mathbb{Q})$	$\mathcal{B}^{\mathcal{A}}(\mathbb{P}, \delta)$
$\mathcal{G}$-Causal (Cheridito and Eckstein 2023)	Preserves Causal Graph	$W^{\mathcal{G}}(\mathbb{P}, \mathbb{Q})$	$\mathcal{B}^{\mathcal{G}}(\mathbb{P}, \delta)$
Structural Causal	Preserves Structural Equations	$W^{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$	$\mathcal{B}^{\mathcal{F}}(\mathbb{P}, \delta)$
Relaxed Structural Causal	Partially Preserves Structural Equations with Penalty Term	$W^{\mathcal{F}_\varepsilon}(\mathbb{P}, \mathbb{Q})$	$\mathcal{B}^{\mathcal{F}_\varepsilon}(\mathbb{P}, \delta)$

Table 1: Comparison of optimal transport variants using nominal distribution information to design ambiguity sets, with corresponding notations for Wasserstein distance and ambiguity set diameter δ For two probability distributions $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{X})$.

distance from the empirical distribution, usually based on a metric like the ℓ_p norm. While this approach works well for unstructured data, it fails for data with special structures, such as temporal patterns or causal relationships. In such cases, the Wasserstein DAS should be refined to exclude unrealistic scenarios. Without this refinement, models become overly conservative, reducing accuracy.

Previous works address this issue by introducing optimal transport (OT) with additional constraints using partial distribution information. *Adapted OT* emphasizes the temporal structure of features (Backhoff et al. 2017). In causal structures, features are arranged according to a directed acyclic graph (DAG) $\mathcal{G}$. While adapted OT preserves the feature hierarchy, it fails to capture causal models. To improve this, (Cheridito and Eckstein 2023) introduce Wasserstein distances for causal models that respect the DAG $\mathcal{G}$. The resulting $\mathcal{G}$-causal OT's DAS is a subset of the adapted DAS, considering all structural causal models with graph $\mathcal{G}$ as candidates. Although this set is narrower than the classical DAS, it still includes unrealistic scenarios by overlooking functional relationships between features. For instance, if a variable $\mathbf{X}_i$ weakly affects its descendant $\mathbf{X}_j$, this weak relationship is not considered in the $\mathcal{G}$-causal OT.

Our Contribution. To tackle this challenge, we propose a novel variant of OT that takes into account the structural equations of causal models. This approach enables us to refine the DAS by limiting it to probability distributions derived from structural causal models with identical structural equations but potentially varying noise variables. This concept facilitates the connection between endogenous and exogenous spaces in constructing the DAS. The resulting duality allows us to define the DAS within the exogenous space, where variable independence can be assumed, simplifying the design process. We can subsequently map back to the feature space to craft our preferred DAS.

Another advantage of our approach is the flexibility to define a relaxed version by replacing causal constraints with an entropy regularization term. This allows the implementation of an efficient algorithm combining difference-of-convex (DC) programming and Sinkhorn's method to solve structural causal OT, a capability not available in $\mathcal{G}$-causal OT.

Since the design of the ambiguity set relies on structural equations, we demonstrate that, in real-world applications where the structural equations are estimated, the solution of the structural causal OT converges to the true solution. This offers a convergence guarantee for finite sample scenarios.

Moreover, we address **P2** by determining the optimal radius necessary to ensure that the true probability distribution is included within the DAS. Our approach mitigates the curse of dimensionality commonly encountered in data-driven ambiguity set design for many OT problems. In the numerical study, we demonstrate the impact of our method and its properties. Additional theoretical results, algorithm descriptions, and numerical outcomes are provided in the appendix, while proofs of our assertions are included in the supplementary material.

Related Work

There are various methods that address how to design DAS, how to add constraints to achieve desirable OT, and discuss the magnitude of DAS. In Tab. 1 and Fig. 1, the main methods, along with the information used in designing DAS and their relationships, are summarized. These methods include:

Discrepancy-based OT. Various discrepancy-based methods exist, such as ϕ-divergences (Love and Bayraksan 2015; Lam 2016; Duchi and Namkoong 2021) and goodness-of-fit tests (Bertsimas, Gupta, and Kallus 2018). However, we focus on the Wasserstein OT due to its advantages over other discrepancies (Gao, Chen, and Kleywegt 2017; Mohajerin Esfahani and Kuhn 2018; Blanchet, Kang, and Murthy 2019).

Temporal OT. Temporal OT (Backhoff et al. 2017; Lassalle 2018; Xu et al. 2020; Bartl, Beiglböck, and Pammer 2021; Backhoff-Veraguas and Pammer 2022) mainly addresses OT for stochastic processes by preserving temporal structure. However, this approach does not encompass all information about the distribution derived from the causal structure. In (Eckstein and Pammer 2024), Sinkhorn's algorithms for adapted OT are proposed by introducing a relaxed version.

$\mathcal{G}$-Causal OT. In (Cheridito and Eckstein 2023), the authors propose $\mathcal{G}$-causal OT to preserve the causal graph $\mathcal{G}$ in the

OT. However, this work does not provide a computational method for estimating $\mathcal{G}$-causal solutions.

Structured ambiguity Set. In the case where all features are independent, our work relates to factored multi-marginal OT (Tran et al. 2021) and CO-OT (Titouan et al. 2020). In this scenario, the ambiguity set also connects to the ambiguity hyperrectangle (Chaouach, Boskos, and Oomen 2022).

Diameter of Ambiguity Set. For the estimation of $W(\mathbb{P}, \mathbb{P}^N)$ for finite samples, various works exist (Bolley, Guillin, and Villani 2007; Fournier and Guillin 2015; Dedecker and Merlevède 2019). The work (Weed and Bach 2019a) provides the sharp order. Additionally, (Chaouach, Boskos, and Oomen 2022; Chaouach, Oomen, and Boskos 2023) discuss the diameter of the ambiguity set under the independent condition of features.

Preliminary Knowledge

Data Model. Let $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n = \mathcal{X}$ denote the n-dimensional features, the set of N observations $\{x^i\}_{i=1}^N$ used to construct the empirical distribution $\mathbb{P}^N$, defined as $\mathbb{P}^N := \frac{1}{N}\sum_{i=1}^N \delta_{x^i}$, where δ_x is the Dirac delta function.

Assume the feature space is modeled by a *structural causal model (SCM)* $\mathcal{M} = \langle \mathcal{G}, \mathbf{X}, \mathcal{F}, \mathbf{U}, \mathbb{P}_\mathbf{U} \rangle$ (Pearl 2009). This includes **structural equations** $\mathcal{F} = \{f_1, f_2, \dots, f_n\}$ where $\{\mathbf{X}_i := f_i(\mathbf{X}_{\mathrm{Pa}(i)}, \mathbf{U}_i)\}_{i=1}^n$, which describe the causal relationships between an endogenous variable $\mathbf{X}_i$, its causal predecessors $\mathbf{X}_{\mathrm{Pa}(i)}$, and an exogenous variable $\mathbf{U}_i$ representing unobservable factors in space $\mathcal{U}_i$, and $\mathbf{U} = (\mathbf{U}_1, \dots, \mathbf{U}_n)$ lives on the whole exogenous space is $\mathcal{U} = \mathcal{U}_1 \times \cdots \times \mathcal{U}_n$. The causal relations are represented by a directed acyclic **causal graph** $\mathcal{G}$.

A DAG imposes a **causal order** (topological order), refers to the sequence in which variables can be arranged such that each variable is only affected by the variables preceding it in the order (Pearl 2009; Peters, Janzing, and Schölkopf 2017). Causal order Only specifies the sequence of variables, indicating which variables can potentially affect others, but not the detailed nature of those effects.

By *causal sufficiency* and *no hidden confounders*, we can suppose exogenous variables to be mutually independent, allowing $\mathbb{P}_\mathbf{U}$ to be written as $\prod_{i=1}^n \mathbb{P}_{\mathbf{U}_i}$ (Peters, Janzing, and Schölkopf 2017).

Since perturbations in SCMs are utilized by counterfactuals, it is necessary for SCMs to be counterfactually identifiable to ensure that counterfactuals can be learned from sample data. One prominent family of counterfactually identifiable models is the *Bijective Generation Mechanism (BGM)* (Nasr-Esfahany, Alizadeh, and Shah 2023). In BGM, the structural equations $\mathcal{F}$ have a **reduced-form mapping** $g : \mathcal{U} \to \mathcal{V}$, where $\mathbf{X}$ can be expressed as a bijective function of the exogenous space, i.e., $\mathbf{X} = g(\mathbf{U})$. This bijective ensures no information is lost from exogenous to endogenous variables.

An important example of BGM is the **Additive Noise Models (ANM)**, where structural equations are given as:

$$\{\mathbf{X}_i := f_i(\mathbf{X}_{\mathrm{pa}(i)}) + \mathbf{U}_i\}_{i=1}^n \implies$$
$$\mathbf{U} = (I - f)(\mathbf{X}) \implies \quad \mathbf{X} = (I - f)^{-1}(\mathbf{U})$$

As seen in the above equation, the reduced-form mapping is $g = (I - f)^{-1}$, where $I(x) = x$ is the identity function. ANM is often preferred over general SCMs due to its simplicity, interpretability, and effective handling of noise, making it ideal for fields such as statistics, causal inference, signal processing, image processing, economics, and social sciences, where additive noise is prevalent. In this work, we focus on the ANM, but our results are extendable to BGMs.

Structured Ambiguity Set. In the variants of Wasserstein OT, the ambiguity set is typically defined through coupling. Let $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{X})$ be probability distributions; the distribution $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$ is called a *coupling* or **plan** if for all measurable subsets $A, B \subset \mathcal{X}$, $\pi(A \times \mathcal{X}) = \mathbb{P}(A)$ and $\pi(\mathcal{X} \times B) = \mathbb{Q}(B)$. Let $\Pi(\mathbb{P}, \mathbb{Q})$ represent the set of all couplings between $\mathbb{P}$, and $\mathbb{Q}$. Each $\pi \in \Pi(\mathbb{P}, \mathbb{Q})$ shows how $\mathbb{P}$ transforms into $\mathbb{Q}$. All plans starting from $\mathbb{P}$ are denoted by $\Pi(\mathbb{P}, *) := \bigcup_{\mathbb{Q} \in \mathcal{P}(\mathcal{X})} \Pi(\mathbb{P}, \mathbb{Q})$.

The structured DAS for $\mathbb{P}$ is a subset of plans $\Pi(\mathbb{P}, *)$ that meet specific constraints $\mathcal{C} = \{c_k\}$, typically defined by parameters $\boldsymbol{\delta} = \{\delta_k\}$:

$$\Pi^\mathcal{C}(\mathbb{P}, \boldsymbol{\delta}) := \{\pi \in \Pi(\mathbb{P}, *) : c_k(\pi, \delta_k), \forall c_k \in \mathcal{C}\}$$

The desirable property of $\Pi^c(\mathbb{P}, \mathbb{Q})$ is its closeness under the weak topology in probability space, which guarantees the existence of solutions in OT theorems. The corresponding ambiguity set for $\Pi^\mathcal{C}(\mathbb{P}, \boldsymbol{\delta})$ is derived by finding the marginal distribution of each plan on the second coordinate (Marg_2):

$$\mathcal{B}^\mathcal{C}(\mathbb{P}, \boldsymbol{\delta}) = \{\mathrm{Marg}_2(\pi) : \pi \in \Pi^\mathcal{C}(\mathbb{P}, \boldsymbol{\delta})\}. \tag{1}$$

In DRO, the worst-case loss is obtained by taking the expectation of a given function $\psi : \mathcal{X} \to \mathbb{R}$ over the ambiguity set. The following alternative formulation, by definition, often facilitates computations.

$$\sup_{\mathbb{Q} \in \mathcal{B}^\mathcal{C}(\mathbb{P}, \boldsymbol{\delta})} \left\{ \mathop{\mathbb{E}}_{y \sim \mathbb{Q}}[\psi(y)] \right\} = \sup_{\pi \in \Pi^\mathcal{C}(\mathbb{P}, \boldsymbol{\delta})} \left\{ \mathop{\mathbb{E}}_{(.,y) \sim \pi}[\psi(y)] \right\} \tag{2}$$

Similar to constrained plans, if there are constraints on the space of probability measures, the corresponding subset is denoted by $\mathcal{P}^c(\mathcal{X})$. Let $c(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to [0, \infty]$ be a transportation cost function that is non-negative and upper-semi-continuous. For each $p \in [1, \infty]$, the set of p-integrable distributions $\mathcal{P}^p(\mathcal{X})$ with respect to c is defined as:

$$\left\{ \mathbb{P} \in \mathcal{P}(\mathcal{X}) \;\middle|\; \int_\mathcal{X} c(x, x_0)^p \mathbb{P}(\mathrm{d}x) < \infty \text{ for some } x_0 \in \mathcal{X} \right\}$$

The $\mathcal{C}$-Wasserstein distance between $\mathbb{P} \in \mathcal{P}^c(\mathcal{X}) \cap \mathcal{P}^p(\mathcal{X})$ and $\mathbb{Q} \in \mathcal{P}(\mathcal{X}) \cap \mathcal{P}^p(\mathcal{X})$ is defined by finding the lowest-cost constrained transport plan:

$$W^c(\mathbb{P}, \mathbb{Q}) := \left(\inf_{\pi \in \Pi^\mathcal{C}(\mathbb{P}, \mathbb{Q})} \left\{ \mathop{\mathbb{E}}_{(x,y) \sim \pi}[c^p(x, y)] \right\} \right)^{\frac{1}{p}} \tag{3}$$

In cases where $\Pi^c(\mathbb{P}, \mathbb{Q}) = \emptyset$, the $W^c(\mathbb{P}, \mathbb{Q})$ is set to ∞.

Wasserstein Ambiguity Set. In classical OT (Villani et al. 2009; Peyré, Cuturi et al. 2017), transport plans and distributions are unconstrained, except for the transportation cost. For two probability measures $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{X})$, the Wasserstein distance $W(\mathbb{P}, \mathbb{Q})$ represents the optimal cost of transporting one distribution to the other. The corresponding ambiguity set, defined by a perturbation radius $\delta \in \mathbb{R}^+$, is given by:

$$\Pi(\mathbb{P}, \delta) = \{\pi \in \Pi(\mathbb{P}, *) : \mathop{\mathbb{E}}_{(x,y)\sim\pi}[c(x, y)] \leq \delta\}$$

By computing the second marginal distribution on $\Pi(\mathbb{P}, \delta)$, the ambiguity set can be expressed as:

$$\mathcal{B}(\mathbb{P}, \delta) = \{\mathbb{Q} \in \mathcal{P}(\mathcal{X}) : W(\mathbb{P}, \mathbb{Q}) \leq \delta\}.$$

Adopted Ambiguity Set. Adapted OT (Backhoff et al. 2017) was introduced to address the limitations of classical OT in preserving temporal constraints for two stochastic processes. Given two discrete-time stochastic processes $\mathbb{P}$ and $\mathbb{Q}$ with indices $i \in [n]$, the coupling π must respect the temporal structure. Consequently, the plans should satisfy the temporal conditional distribution constraints as follows:

$$\Pi^{\mathcal{A}}(\mathbb{P}, \mathbb{Q}) = \big\{\pi \in \Pi(\mathbb{P}, \mathbb{Q}) : \pi - \text{almost sure } \forall x, y \in \mathcal{X},$$
$$\pi(\mathrm{d}y_i \mid \mathrm{d}x_1, \ldots, \mathrm{d}x_n) = \pi(\mathrm{d}y_i \mid \mathrm{d}x_1, \ldots, \mathrm{d}x_i), \forall\, i \in [n]\big\}$$

The adapted ambiguity set and Wasserstein distance are denoted by $\mathcal{B}^{\mathcal{A}}(\mathbb{P}, \delta)$ and $W^{\mathcal{A}}(\mathbb{P}, \mathbb{Q})$, respectively, similar to Eq. 1 and Eq. 3.

$\mathcal{G}$-Compatible Ambiguity Set. Since adapted OT cannot preserve complex structures like causal graph dependencies, (Cheridito and Eckstein 2023) proposed the $\mathcal{G}$-causal OT framework. A probability $\mathcal{P} \in \mathcal{P}(\mathcal{X})$ is called compatible with a sorted causal graph $\mathcal{G}$ if there exists a random variable $\mathbf{X} \sim \mathbb{P}$, along with measurable functions $f_i : \mathcal{X}_{\mathrm{pa}(i)} \times \mathbb{R}^{d_i} \to \mathcal{X}_i$ for $i = 1, \ldots, n$, and independent random variables $\mathbf{U}_i$ for all $i \in [n]$ such that:

$$\mathbf{X}_i = f_i(\mathbf{X}_{\mathrm{pa}(i)}, \mathbf{U}_i) \quad \text{for all } i = 1, \ldots, n.$$

The set of $\mathcal{G}$-compatible measures is denoted by $\mathcal{P}^{\mathcal{G}}(\mathcal{X})$. A $\mathcal{G}$-compatible plans, which capture $\mathcal{G}$ graph, is defined as:

$$\Pi^{\mathcal{G}}(\mathbb{P}, \mathbb{Q}) = \big\{\pi \in \Pi(\mathbb{P}, \mathbb{Q}) : \forall\, i \in [n],\ \pi\text{-almost all } (x, y),$$
$$\pi(\mathrm{d}x_1, \mathrm{d}y_1, \ldots, \mathrm{d}x_n, \mathrm{d}y_n) = \bigotimes_{i=1}^{n} \pi(\mathrm{d}x_i, \mathrm{d}y_i \mid x_{\mathrm{pa}(i)}, y_{\mathrm{pa}(i)})$$
$$\text{and} \quad \pi(\mathrm{d}x_i \mid x_{\mathrm{pa}(i)}, y_{\mathrm{pa}(i)}) = \mathbb{P}(\mathrm{d}x_i \mid x_{\mathrm{pa}(i)})\big\}. \tag{4}$$

The corresponding DAS and Wasserstein metric are denoted by $\mathcal{B}^{\mathcal{G}}(\mathbb{P}, \delta)$ and $W^{\mathcal{G}}(\mathbb{P}, \mathbb{Q})$. Since preserving the causal graph inherently includes preserving the causal order of features, $\mathcal{G}$-compatible plans are a subset of adapted plans.

By reviewing the main definitions and notations, we are prepared to present our method, which addresses the limitations of previous methods in fully capturing the information of causal models.

Structural Causal Ambiguity Sets

The adopted $\mathcal{G}$-causal ambiguity set retains only the causal graph structure without requiring the full details of the causal model. For example, if the nominal distribution indicates a weak relationship between a parent $\mathbf{X}_i$ and its child $\mathbf{X}_j$ (e.g., $\mathbf{X}_i = \alpha\mathbf{X}_j$ with $\alpha \approx 0$), the $\mathcal{G}$-causal set neglects this weak dependency. To address this limitation, we propose a new OT variant that incorporates structural equations from the SCM, thereby capturing these dependencies and utilizing more information than the causal graph $\mathcal{G}$ alone.

Before presenting our method, we highlight a natural assumption. Let c be the cost function on the feature space. Given an invertible SCM, there exists a bijective map g such that $x = g(u)$. We define the push-forward cost $\tilde{c} = c \circ (g \times g)$ on the exogenous space as $\tilde{c}(u, u') = c(g(u), g(u'))$. Since the variables in the exogenous space are mutually independent, each $\mathcal{U}_i \times \mathcal{U}_i$ can have its own cost function $\tilde{c}_i$, allowing $\tilde{c}$ to be decomposed into components. As a result, $\tilde{c}$ is expected to have a simpler form. In summary, we make the following assumptions.

Assumption 1. *(i) $\mathcal{M}$ is a ANM, with structural equations $\mathcal{F} = \{f_i\}$ and g is bijective reduced-form mapping.*

(ii) The random variables $\mathbf{U}_i$ are independent and take values in the $\mathcal{U}_i \subseteq \mathbb{R}^{d_i}$ that is equipped by the norm $\tilde{c}_i$.

(iii) The push-forward of the cost function c to the exogenous space has the form:

$$\tilde{c}(u, u') = \left(\sum_{i=1}^{n} \tilde{c}_i(u_i, u_i')^p\right)^{\frac{1}{p}}, \forall u_i, u_i' \in \mathcal{U}_i \text{ and } p \geq 1.$$

Now, we are prepared to present our constraints in both probability space and plans.

Definition 1 (**$\mathcal{F}$-Compatible Measures**). *A measure is compatible with the structural equations $\mathcal{F}$ if its g^{-1} push-forward distribution over the exogenous space is factored,*

$$\mathcal{P}^{\mathcal{F}}(\mathcal{X}) = \left\{\mathbb{P} \in \mathcal{P}(\mathcal{X}) : g_{\#}^{-1}\mathbb{P} = \bigotimes_{i=1}^{n} \tilde{\mathbb{P}}_i, \quad \tilde{\mathbb{P}}_i \in \mathcal{P}(\mathcal{U}_i)\right\},$$

where $\bigotimes$ means the product of measures.

Another useful definition of $\mathcal{P}^{\mathcal{F}}(\mathcal{X})$ is that it includes all g-pushforward distributions of $\bigotimes_{i=1}^{n} \tilde{\mathbb{P}}_i$ where $\tilde{\mathbb{P}}_i \in \mathcal{P}(\mathcal{U}_i)$. This duality facilitates conversion between spaces, simplifying our results. The following lemma outlines the properties of $\mathcal{F}$-compatible measures.

Proposition 1. *Let $\mathbb{P} \in \mathcal{P}^{\mathcal{F}}(\mathcal{X})$, then:*

(i) There exists a random variable $\mathbf{X} \sim \mathbb{P}$ along with independent random variables $\mathbf{U}_1, \ldots, \mathbf{U}_n$ in the space $\mathcal{U}_i$ such that,

$$\mathbf{X}_i = f_i(\mathbf{X}_{\mathrm{pa}(i)}, \mathbf{U}_i) \quad \textit{for all } i = 1, \ldots, n.$$

(ii) The measure $\mathbb{P}$ can be decomposed as

$$\mathbb{P}(\mathrm{d}x_1, \ldots, \mathrm{d}x_n) = \bigotimes_{i=1}^{n} \mathbb{P}\left(\mathrm{d}x_i \mid x_{\mathrm{pa}(i)}\right),$$

which means variables are conditionally independent of their non-descendants given their parents.

We are now ready to address **P1** in our method, which determines the specific information that should be considered in the design of the ambiguity set.

Definition 2 (**$\mathcal{F}$-Compatible Plans**). *The plan π is called $\mathcal{F}$-compatible if its pushforward map $\tilde{\pi} = (g^{-1} \times g^{-1})_{\#}\pi$ under $g^{-1} \times g^{-1}$ is factored in the exogenous space as follows:*

$$\Pi^{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \Bigg\{$$

$$\pi \in \Pi(\mathbb{P}, \mathbb{Q}) : \textit{for } \tilde{\pi}\textit{-almost sure}, \forall (u, w) \in \mathcal{U} \times \mathcal{U},$$

$$(g^{-1} \times g^{-1})_{\#}\pi(\mathrm{d}u, \mathrm{d}w) = \bigotimes_{i=1}^{n} \tilde{\pi}_i(\mathrm{d}u_i, \mathrm{d}w_i),$$

$$\textit{such that } \tilde{\pi}_i \in \Pi\left(\mathrm{Marg}_i(g_{\#}^{-1}\mathbb{P}), \mathrm{Marg}_i(g_{\#}^{-1}\mathbb{Q})\right) \Bigg\}.$$

where Marg_i *is the marginal distribution over the coordinate* $\mathcal{X}_i$.

The intuition behind this definition is straightforward: we consider the plans π whose push-forward in the exogenous space decomposes onto $\mathcal{U}_i \times \mathcal{U}_i$, as we have mutually independent noise by the assumption. Now we investigate the distributional properties implied by the definition of $\mathcal{F}$-compatible plans.

Proposition 2. *Let $\mathbb{P}, \mathbb{Q} \in \mathcal{P}^{\mathcal{F}}(\mathcal{X})$ and $\pi \in \Pi^{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$, then we have:*

(i) if $(X, Y) \sim \pi$ then there exists measurable functions

$$h_i \colon \mathcal{X}_i \times \mathcal{X}_{\mathrm{pa}(i)} \times \mathcal{X}_{\mathrm{pa}(i)} \times \mathbb{R}^{d_i} \to \mathcal{X}_i, \quad i \in [n],$$

and $\mathbb{R}$-valued random variables $\mathbf{V}_1, \ldots, \mathbf{V}_n$ such that $\mathbf{X}, \mathbf{V}_1, \ldots, \mathbf{V}_n$ are mutually independent and

$$\mathbf{Y}_i = h_i(\mathbf{X}_i, \mathbf{X}_{\mathrm{pa}(i)}, \mathbf{Y}_{\mathrm{pa}(i)}, \mathbf{V}_i) \quad \textit{for all } i \in [n].$$

(ii) for all $i = 1, \ldots, n$ and π-almost all $(x, y) \in \mathcal{X} \times \mathcal{Y}$

$$\pi(\mathrm{d}x_1, \mathrm{d}y_1, \ldots, \mathrm{d}x_n, \mathrm{d}y_n) = \bigotimes_{i=1}^{n} \pi(\mathrm{d}x_i, \mathrm{d}y_i \mid x_{\mathrm{pa}(i)}, y_{\mathrm{pa}(i)})$$

$$\textit{and} \quad \pi(\mathrm{d}x_i, \mathrm{d}y_i \mid x_{\mathrm{pa}(i)}, y_{\mathrm{pa}(i)}) \in \Pi_i, \quad \forall i \in [n],$$

where $\Pi_i = \Pi\left(\mathbb{P}(\mathrm{d}x_i \mid x_{\mathrm{pa}(i)}), \mathbb{Q}(\mathrm{d}y_i \mid y_{\mathrm{pa}(i)})\right)$.

Proposition 2 ensures that $\mathcal{F}$-compatible plans preserve the causal graph structure in conditional distributions. Here, we present the main properties of $\Pi^{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$.

Proposition 3. *If $\mathbb{P}, \mathbb{Q} \in \mathcal{P}^{\mathcal{F}}(\mathcal{X})$, then $\Pi^{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$ are non-empty and weakly closed.*

By demonstrating the properties of definitions, we present our new OT problem, which minimizes transport costs over plans that preserve the structural equations $\mathcal{F}$.

Definition 3 (**Structural Causal OT**). *For $p \in [1, \infty)$ and $\mathbb{P}, \mathbb{Q} \in \mathcal{P}^{\mathcal{F}}(\mathcal{X}) \cap \mathcal{P}^{p}(\mathcal{X})$, the structural causal Wasserstein distance is finding the minimum-cost $\mathcal{F}$-compatible plans between $\mathbb{P}$ and $\mathbb{Q}$:*

$$W^{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) \coloneqq \left(\inf_{\pi \in \Pi^{\mathcal{F}}(\mathbb{P}, \mathbb{Q})} \left\{ \mathop{\mathbb{E}}_{(x,y)\sim\pi} [c^p(x, y)] \right\} \right)^{\frac{1}{p}} \tag{5}$$

The result below outlines the fundamental properties of the structural causal Wasserstein distance.

Proposition 4. *$W^{\mathcal{F}}$ is a semi-metric on $\mathcal{P}^{\mathcal{F}}(\mathcal{X})$ and attains its minimum.*

Now we are ready to introduce **structural causal ambiguity set** $W^{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$ which is defined as

$$\mathcal{B}^{\mathcal{F}}(\mathbb{P}, \delta) \coloneqq \{\mathbb{Q} \in \mathcal{P}^{\mathcal{F}}(\mathcal{X}) : W^{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) \leq \delta\} \tag{6}$$

Since our constraints involve the $\mathcal{G}$-causal information, which encompasses causal order, we intuitively expect the corresponding DAS to be nested sets. This intuition is confirmed in the following proposition.

Proposition 5. *Let $\mathbb{P}, \mathbb{Q} \in \mathcal{P}^{\mathcal{F}}(\mathcal{X})$, then for different definitions of the ambiguity set, we have:*

(i) $W^{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) \geq W^{\mathcal{G}}(\mathbb{P}, \mathbb{Q}) \geq W^{\mathcal{A}}(\mathbb{P}, \mathbb{Q}) \geq W(\mathbb{P}, \mathbb{Q})$,

(ii) $\mathcal{B}^{\mathcal{F}}(\mathbb{P}, \delta) \subseteq \mathcal{B}^{\mathcal{G}}(\mathbb{P}, \delta) \subseteq \mathcal{B}^{\mathcal{A}}(\mathbb{P}, \delta) \subseteq \mathcal{B}(\mathbb{P}, \delta)$.

Let $\mathcal{F}^0$ represent the zero structural equations, i.e., $f_i \equiv 0, \forall i$, implying no causal structure in the SCM. We finish this section by explaining the duality that demonstrates the correspondence between the DAS in the feature space with causal structure $\mathcal{F}$ and the DAS in the exogenous space with structural equations $\mathcal{F}^0$. In the proposition below, to highlight that the cost functions differ in the two spaces, we embed the cost function in the notation of the ambiguity set.

Proposition 6. *Let $\mathcal{F}$ be structural equations with bijective reduced-form mapping g and $\mathbb{P} \in \mathcal{P}^{\mathcal{F}}(\mathcal{X})$, then*

$$\mathcal{B}_c^{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = g_{\#}\mathcal{B}_{c\circ(g\times g)}^{\mathcal{F}^0}(g_{\#}^{-1}\mathbb{P}, \delta).$$

This property plays a crucial role in designing the relaxed OT problem.

Relaxed Structural Causal Optimal Transport

One challenge in defining new variants of OT is developing efficient algorithms to compute the Wasserstein distance. In both classical and adapted OT, entropic regularization provides an efficient solution by adding an entropy penalty term to the original problem (Cuturi 2013; Eckstein and Pammer 2024). To provide a fast computation method, we introduce a relaxed version of structural causal OT. This modification transforms the original problem into a difference-of-convex optimization problem, making it more computationally feasible. As a result, iterative algorithms like the Sinkhorn-Knopp (Benamou et al. 2015) can solve the problem efficiently, significantly reducing computation time and enabling the handling of large-scale problems.

Definition 4 (**Relaxed Structural Causal OT**). *Given $\varepsilon \geq 0$ and probability measures $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{X})$, we define the relaxed structural causal OT by solving the following optimization problem:*

$$W^{\mathcal{F}_\varepsilon}(\mathbb{P}, \mathbb{Q})^p \coloneqq \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \mathop{\mathbb{E}}_{(x,y)\sim\pi} [c^p(x, y)] + \varepsilon D_{KL}\left(\pi \| \pi_{\otimes}\right)$$

where $\pi_{\otimes} \in \Pi(\mathbb{P}, \mathbb{Q})$ is obtained by the following mapping:

$$\pi_{\otimes} = (g \times g)_{\#} \left(\bigotimes_{i=1}^{n} \mathrm{Marg}_i \left((g^{-1} \times g^{-1})_{\#}\pi \right) \right) \tag{7}$$

The intuition behind the definition of $\pi_{\otimes}$ is straightforward: it acts as a projection π onto the space $\Pi^{\mathcal{F}}(\mathbb{P},\mathbb{Q})$. Thus, if $\pi \in \Pi^{\mathcal{F}}(\mathbb{P},\mathbb{Q})$, then $\pi_{\otimes} = \pi$. A small penalty value $D_{\mathrm{KL}}(\pi \| \pi_{\otimes})$ indicates that π is close to the set $\Pi^{\mathcal{F}}(\mathbb{P},\mathbb{Q})$.

The relaxed version simplifies the problem by shifting the search for an optimal solution from the constrained plans $\Pi^{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ to the simpler space $\Pi(\mathbb{P},\mathbb{Q})$, while preserving the $\mathcal{F}$ structure via a regularizer. The following proposition shows that the relaxed version converges to the structural causal OT as ϵ approaches infinity. This result guarantees the effectiveness of the relaxed solution in finding the structural causal OT.

Proposition 7. *Let π_ε be the minimizer of $W^{\mathcal{F}_\varepsilon}(\mathbb{P},\mathbb{Q})$ then:*
(i) when $\epsilon \to \infty$ then $W^{\mathcal{F}_\varepsilon}(\mathbb{P},\mathbb{Q}) \to W^{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ and every cluster point in the set $\{\pi_\varepsilon\}$ is the optimal solution of $W^{\mathcal{F}}(\mathbb{P},\mathbb{Q})$.
(ii) when $\epsilon \to 0$ then $W^{\mathcal{F}_\varepsilon}(\mathbb{P},\mathbb{Q}) \to W(\mathbb{P},\mathbb{Q})$ and every cluster point of the set $\{\pi_\varepsilon\}$ is the optimal solution of $W(\mathbb{P},\mathbb{Q})$.

Hopefully, not only does $\mathcal{B}^{\mathcal{F}_\varepsilon}(\mathbb{P},\delta)$ converge to $\mathcal{B}^{\mathcal{F}}(\mathbb{P},\delta)$ as $\varepsilon \to \infty$, but $\mathcal{B}^{\mathcal{F}}(\mathbb{P},\delta)$ is also always a subset of $\mathcal{B}^{\mathcal{F}_\varepsilon}(\mathbb{P},\delta)$, aiding in efficiently estimating this set from above. The proposition below formalizes this result.

Proposition 8. *For relaxed structural causal OT and $\mathbb{P},\mathbb{Q} \in \mathcal{P}^{\mathcal{F}}(\mathcal{X})$ we have:*
(i) $W^{\mathcal{F}_\varepsilon}(\mathbb{P},\mathbb{Q})$ attains its minimum.
(ii) $W^{\mathcal{F}}(\mathbb{P},\mathbb{Q}) \geq W^{\mathcal{F}_\varepsilon}(\mathbb{P},\mathbb{Q}) \geq W(\mathbb{P},\mathbb{Q})$,
(iii) $\mathcal{B}^{\mathcal{F}}(\mathbb{P},\delta) \subseteq \mathcal{B}^{\mathcal{F}_\varepsilon}(\mathbb{P},\delta) \subseteq \mathcal{B}(\mathbb{P},\delta)$.

The duality between the relaxed Wasserstein distance in feature space with structural equations $\mathcal{F}$ and in exogenous space with structural equations $\mathcal{F}^0$ is key to designing efficient computational methods for determining structural causal distance.

Proposition 9. *For $\mathbb{P},\mathbb{Q} \in \mathcal{P}^{\mathcal{F}}(\mathcal{X})$ we have:*

$$W_c^{\mathcal{F}_\varepsilon} = W_{c\circ(g\times g)}^{\mathcal{F}^0_\varepsilon}(g_{\#}^{-1}\mathbb{P}, g_{\#}^{-1}\mathbb{Q}) \tag{8}$$

Now we are ready to design our algorithm. If we consider $\tilde{\mathbb{P}},\tilde{\mathbb{Q}},\tilde{c},\tilde{\pi}$ as the push-forwards of $\mathbb{P},\mathbb{Q},c$, and π by g^{-1}, then by Prop. 9, we can express $W_{\tilde{c}}^{\mathcal{F}^0_\varepsilon}(\tilde{\mathbb{P}},\tilde{\mathbb{Q}})$ as:

$$\inf_{\tilde{\pi}\in\Pi(\tilde{\mathbb{P}},\tilde{\mathbb{Q}})} \left\{ \mathop{\mathbb{E}}_{(u,v)\sim\tilde{\pi}}[c(u,v)] + H(\tilde{\pi}) - \sum_{i=1}^{n} H_i(\tilde{\pi}) \right\}, \tag{9}$$

Here, $H(\tilde{\pi})$ is the entropy of the plan $\tilde{\pi}$, and $H_i(\tilde{\pi})$ is the entropy of the i-th marginal distribution $\tilde{\pi}_i$ over the coordinate $(i, i+n)$. For example, if $(u_1,\ldots,u_n,v_1,\ldots,v_n) \sim \tilde{\pi}$, then $\tilde{\pi}_1$ represents the marginal distribution on the coordinates (u_1,v_1). Thus, $H_i(\tilde{\pi})$ can be written as $\mathbb{E}_{\tilde{\pi}}[\log(\mathrm{Marg}_i(\tilde{\pi})(u_i,v_i))]$. Since $H(\tilde{\pi})$ and $H_i(\tilde{\pi})$ are convex functions, the optimization problem in Eq. 9 is a difference of convex functions. Therefore, by applying the DC algorithm, we can estimate the value of $\tilde{\pi}$. In the DC algorithm, the convex term $\sum_i H_i(\tilde{\pi})$ is iteratively replaced by its linear approximation, converting Eq. 9 into a convex problem. DC algorithm implies if we define:

$$G(\tilde{\pi}) \in \partial(\sum H_i)(\tilde{\pi}),$$

we can reformulate the problem as a convex optimization:

$$\inf_{\tilde{\pi}\in\Pi(\tilde{\mathbb{P}},\tilde{\mathbb{Q}})} \{\langle \tilde{c} - \varepsilon G(\tilde{\pi}), \tilde{\pi}\rangle + H(\tilde{\pi})\}. \tag{10}$$

Eq. 10 can be solved using the Sinkhorn method for multi-marginal OT (Benamou et al. 2015) to find the minimum cost plan, since for $\tilde{\pi} \in \Pi(\tilde{\mathbb{P}},\tilde{\mathbb{Q}})$, we have:

$$\mathrm{Marg}_i(\tilde{\pi}) = \begin{cases} \mathrm{Marg}_i(\tilde{\mathbb{P}}) & i \leq n, \\ \mathrm{Marg}_{i-n}(\tilde{\mathbb{Q}}) & i > n. \end{cases}$$

In the case where $P = (p_k)_k \in \mathbb{R}^N$ corresponds to feature values $(x^k)_k$ and $Q = (q_r)_r \in \mathbb{R}^M$ corresponds to feature values $(y^r)_r$ instead of $\mathbb{P}$ and $\mathbb{Q}$, we first estimate the structural equations using sample data points. After estimating the reduced-form mapping g, we map the sample data to the exogenous space to obtain $u = g^{-1}(x)$ and $v = g^{-1}(y)$. Hence, we can express

$$p_k = P(u_1^k,\ldots,u_n^k), \quad q_r = P(v_1^r,\ldots,v_n^r).$$

By summing over the other coordinates, we can calculate the marginal distribution P_i as:

$$\tilde{\mathbb{P}}_i(u_i^k) = \sum_{u_1^{j_1},\ldots,u_{i-1}^{j_{i-1}},u_{i+1}^{j_{i+1}},\ldots u_n^{j_n}} P(u_1^{j_1},\ldots,u_i^k,\ldots,u_n^{j_n}).$$

Similarly, we can compute the marginal $\tilde{\mathbb{Q}}_i$. Since we know the cost function in the exogenous space by assumption 1, we can calculate the cost tensor. Then, we apply the Sinkhorn algorithm to find the tensor π. The above steps are summarized in Alg. 1.

Since designing the relaxed structural causal OT requires estimating structural equations, we need assurance that using sample data to estimate these equations will converge to the optimal plan. The next theorem confirms this property.

Theorem 1 (Finite Sample Guarantee). *Let assumption 1 hold and let $\mathcal{F} = \{f_i\}$ represent the continuous structural equations, with $\hat{\mathcal{F}} = \{\hat{f}_i\}$ denoting the estimated structural equations. Suppose $\mathbb{P},\mathbb{Q} \in \mathcal{P}^{\mathcal{F}}(\mathcal{X})$ have compact support. Then, for every $\epsilon > 0$, there exists a $\delta > 0$ such that if $\|f_i - \hat{f}_i\|_\infty < \delta$, then*

$$\left| W^{\mathcal{F}_\varepsilon}(\mathbb{P},\mathbb{Q}) - W^{\hat{\mathcal{F}}_\varepsilon}(\mathbb{P},\mathbb{Q}) \right| \leq \epsilon,$$

where $\|\cdot\|_\infty$ denotes the supremum norm.

Concentration Inequality In Presence of SCM

Determining the ambiguity set radius (**P2**) in DRO is crucial for balancing robustness and sample sensitivity. A smaller radius increases sensitivity to noise and reduces robustness, while a larger one enhances robustness but may overlook the true distribution's behavior. The optimal choice involves estimating the magnitude of $W(\mathbb{P}^N,\mathbb{P})$.

Numerous studies explore the concentration of $W(\mathbb{P}^N,\mathbb{P})$. For example, (Fournier and Guillin 2015) provides convergence bounds, (Dedecker and Merlevède 2019) examines dependence conditions, and (Weed and Bach 2019b) offers sharp inequalities. Below, we adapt (Fournier and Guillin 2015, Theorem 1) and tailor the results to our non-metric cost function.

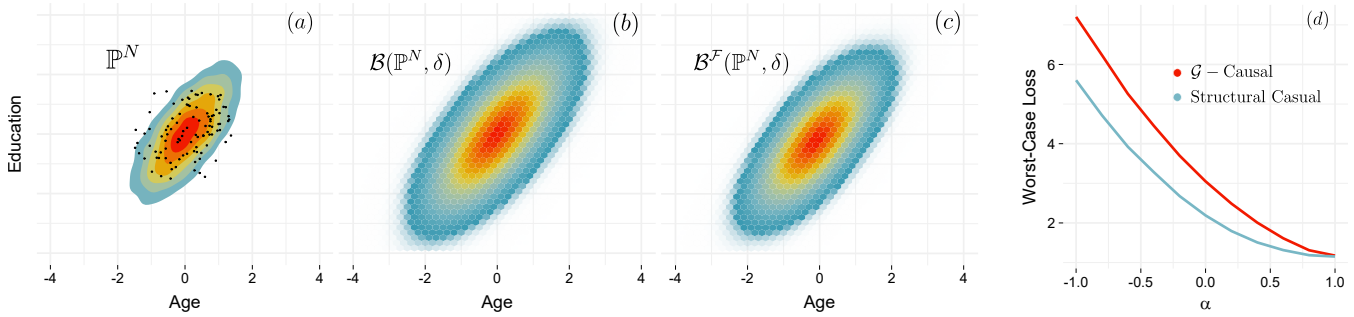

Figure 2: (a) Empirical estimation of true probability distribution for the model $\mathbf{E} = \mathbf{A} + \mathbf{U_E}$ (Age and Education is normalized). (b) Ambiguity set obtained via classical OT with radius 0.5. (c) Structural causal ambiguity set with radius 0.5. (d) Comparing Worst-case losses for the structural causal and $\mathcal{G}$-causal DAS with radius $\delta = 0.5$ and function $\psi(x, y) = (x - y)^2$.

Algorithm 1: Relaxed Structural Causal Optimal Plan

Input: Probability measures $P = (p_k)_k \in \mathbb{R}^N$ for feature values $(x^k)_k$ and $Q = (q_r)_r \in \mathbb{R}^M$ for feature values $(y^r)_r$, $\tilde{c}$ cost function over exogenous space and regularization parameter ϵ.
Output: Tensor $\pi \in \Pi(\mathbb{P}, \mathbb{Q})$.

1. Estimate structural equations $\hat{\mathcal{F}}$ and obtain reduced-form mappings $\hat{g}, \hat{g}^{-1}$.
2. Calculate exogenous values $(u^k)_k$ and $(v^r)_r$ with $u^k = \hat{g}^{-1}(x^k)$, $v^r = \hat{g}^{-1}(y^r)$.
3. Compute marginal distributions $\tilde{P}_i$ and $\tilde{Q}_i$ for $i \in [n]$.
4. Calculate the cost tensor on the exogenous space $C = \{\tilde{c}(u, w)\}$ where $u = (u_1^{i_1}, \ldots, u_n^{i_n})$, $v = (v_1^{j_1}, \ldots, v_n^{j_n})$ and $i_k \in [N], j_k \in [M]$.
5. While not converged:
   - Gradient step: compute the gradient of the convex term $G^{(t)} = \sum_i \nabla_\pi H_i(\pi^{(t)})$.
   - Sinkhorn step: Estimate $\pi^{(t+1)}$

   $$\pi^{(t+1)} = \arg\min_{\pi \in \Pi(\tilde{\mathbb{P}}, \tilde{\mathbb{Q}})} \langle C - \varepsilon G^{(t)}, P \rangle + \varepsilon H(P),$$

   by the Sinkhorn Algorithm.
6. Output the tensor π corresponding to the probability values of $(\hat{g}(u), \hat{g}(w)) \in \mathcal{X} \times \mathcal{X}$.

Proposition 10 (Concentration Inequality). *Let $\mathbb{P}^{\otimes} = \mathbb{P} \otimes \mathbb{P} \otimes \cdots$ for the product measure on $\mathcal{X}^N$, the space of all sequences of observations and Let $\mathbb{P} \in \mathcal{P}(\mathcal{X})$ compactly support and satisfy Assumption 1. Then for every $N \geq 1$ and any confidence level $1 - \varepsilon$ with $\varepsilon \in (0, 1)$, there exists δ that holds.*

$$\mathbb{P}^{\otimes}(\mathbb{P} \in \mathcal{B}_p(\hat{\mathbb{P}}^N, \delta)) \geq 1 - \varepsilon,$$

where the radius $\delta(N, \varepsilon)$ satisfies:

$$\delta(N, \varepsilon) \lesssim \left(N \ln(C\varepsilon^{-1})\right)^{-1/\max\{d, 2p\}}, \qquad (11)$$

where C is constant depends only to $\mathbb{P}$ and d dimension of feature space. Moreover, if $d \geq 2p$ and $\mathbb{P}$ have a density function such that its support is compact convex, then for every $t < d$,

$$\delta(N, \varepsilon) \gtrsim N^{-1/t}, \qquad (12)$$

Prop. 10 shows that in high-dimensional spaces, the radius decreases slowly at a rate of $N^{-1/d}$, which is non-improvable. Thus, merely increasing the sample size offers limited improvement in approximating the true distribution or shrinking the ambiguity ball. To overcome this, we exploit the independence of components in the exogenous space, enabling a more refined ambiguity set than traditional Wasserstein sets. This approach mitigates the curse of dimensionality and ensures performance regardless of dimension d.

The key idea of the theorem is to leverage the causal structure to construct the empirical distribution instead of directly constructing $\mathbb{P}^N$. Given samples $(x^i)_i$, we first derive the corresponding exogenous samples $(u^i)_i$ and then construct the empirical distribution $\mathbb{P}^N_{\mathbf{U}_k} := \frac{1}{N}\sum_{i=1}^N \delta_{u_k^i}$ for each exogenous component. By independence assumption, $\hat{\mathbb{P}}^N_{\mathbf{U}}$ is obtained as $\hat{\mathbb{P}}^N_{\mathbf{U}} = \mathbb{P}^N_{\mathbf{U}_1} \otimes \cdots \otimes \mathbb{P}^N_{\mathbf{U}_n}$. Finally, by mapping back to the feature space, we construct $\hat{\mathbb{P}}^N_{\otimes} = g_{\#}\hat{\mathbb{P}}^N_{\mathbf{U}}$.

Proposition 11. *Let $\mathbb{P} \in \mathcal{P}(\mathcal{X})$ compactly support and satisfy the assumption 1. Then for every $N \geq 1$ and any confidence level $1 - \varepsilon$ with $\varepsilon \in (0, 1)$, there exists δ that holds.*

$$\mathbb{P}^{\otimes}(\mathbb{P} \in \mathcal{B}^{\mathcal{F}}(\hat{\mathbb{P}}^N_{\otimes}, \delta)) \geq 1 - \varepsilon,$$

where the radius $\delta(N, \varepsilon)$ satisfies:

$$\delta(N, \varepsilon) \lesssim \left(N \ln(Cn\varepsilon^{-1})\right)^{-1/\max\{d^*, 2p\}}, \qquad (13)$$

where $d^* = \max_{i=1}^n d_i$ and C is constant depends only to $\mathbb{P}$ and d_i.

We conclude this section with the following corollary, which demonstrates that the convergence rate in structural causal models does not depend on the dimension of the space.

Corollary 1. *If $d_i = 1$ and $d \geq 2p + 1$, then $W(\hat{\mathbb{P}}^N, \mathbb{P}) \lesssim N^{-1/d}$, however $W^{\mathcal{F}}(\hat{\mathbb{P}}^N_{\otimes}, \mathbb{P}) \lesssim N^{-1/2p}$. This implies that the dependence is only on $\mathbb{P}$, allowing us to break the curse of dimensionality.*

$\psi(x,y)$	$\delta=0.1$	$\delta=0.2$	$\delta=0.3$	$\delta=0.4$
$\lvert x-y\rvert$	2.70	6.35	9.95	13.6
$(x-y)^2$	5.54	12.9	22.0	29.6
$\lvert x+y\rvert$	0.722	0.730	0.830	1.05
$(x+y)^2$	1.10	1.32	1.68	1.86
x^2+y^2	2.46	5.15	7.53	11.2

Table 2: Shows the additional percentage of worst-case loss $\sup_{\mathbb{Q}\in\mathcal{B}^{\mathcal{G}}(\mathbb{P},\delta)}\mathbb{E}[\psi]$ relative to $\sup_{\mathbb{Q}\in\mathcal{B}^{\mathcal{F}}(\mathbb{P},\delta)}\mathbb{E}[\psi]$ for different ambiguity radius δ values and functions. A positive value indicates the fact $\mathcal{B}^{\mathcal{F}}(\mathbb{P},\delta)\subseteq\mathcal{B}^{\mathcal{G}}(\mathbb{P},\delta)$.

Experimental Evaluation

As our method is a novel variant of OT, it is applicable in any scenario where OT or Wasserstein distance has been previously employed, particularly when the data model is derived from causal structures. Fields such as transfer learning, reinforcement learning, algorithmic fairness, generative adversarial networks, and clustering (see additional applications in (Montesuma, Mboula, and Souloumiac 2023; Khamis et al. 2024)) could benefit from our approach. Therefore, a comprehensive numerical demonstration of its applications requires further independent and follow-up work.

In this section, we demonstrate that even with the simplest causal structures in data, different OT variants can produce varying results. Consider a super simple model involving two demographic variables, Age ($\mathbf{A}$) and Education ($\mathbf{E}$), which are common features in real datasets with a known causal relationship. We model this using the simple linear SCM: $\mathbf{A} := \mathbf{U_A}; \mathbf{E} := \alpha\mathbf{A} + \mathbf{U_E}$, where $\mathbf{U_E}$ and $\mathbf{U_A}$ are standard normal distributions (as we normalize age and education in our data). We simulate this model for varying $\alpha \in [-1, 1]$ and compute the classical, structural causal and $\mathcal{G}$-causal DAS for different radii $\delta \in \{0.1, 0.2, 0.3, 0.4\}$ to illustrate the differences between OT variants. To quantify the difference of ambiguity sets, we compute the worst-case loss (Eq. 2) for the functions $\psi(x, y) = |x - y|$, $(x - y)^2$, $|x + y|$, $(x + y)^2$, and $x^2 + y^2$.

To compute structural causal OT, we use Prop. 6, which reformulates the model into a CO-OT (Tran et al. 2021) problem. The structural causal distance is then calculated using the COOT Python package available on (Flamary and contributors 2023). We generated 10,000 distributions to explore the structural causal ambiguity set.

We have a challenge in computing the $\mathcal{G}$-causal distance due to the lack of direct computational methods. To overcome this, we randomly generated 4-dimensional Gaussian plans that preserve the $\mathcal{G}$-causal structure, producing 10,000 distributions as samples for the $\mathcal{G}$-causal DAS. For generating the classical OT DAS, tools like (Lab 2024) are useful in computing the Wasserstein distance.

In Fig. 2(a), the empirical density is displayed. We compare our DAS with classical OT ambiguity sets by generating 1000 points from each probability measure within the DAS, aggregating the points, and plotting a heatmap. In part (b), the classical OT ambiguity set is depicted, which is larger than the structural causal DAS. Unlike classical OT, which expands in all directions disregarding causal structure, the structural causal DAS maintains causal relations.

As shown in Table 2, the worst-case loss is consistently lower for the structural causal DAS compared to the $\mathcal{G}$-causal DAS (supports Prop.5). Notably, in scenarios like $(x-y)^2$, as illustrated in Fig. 2(d), the loss difference is significant because the $\mathcal{G}$-causal DAS does not maintain causal links as effectively as the structural causal DAS. This alignment in structural causal DAS reduces loss when designing the ambiguity set around causal structural equations.

Discussion and Limitations

The main focus of this work is to establish a theoretical framework for a new variant of OT that incorporates not only the causal graph but also the magnitude of relationships between features. We address key aspects (**P1** and **P1**) of designing the new DAS and demonstrate its advantages compared to previous definitions. In the numerical section, we illustrate the impact of our method, even with the simplest causal structure.

To demonstrate the advantages of our method, further independent work focusing on real-world applications, including transfer learning, algorithmic fairness, GANs, etc., is essential to complete the theoretical aspects of our research.

To enhance our method for real-world problems, it is essential to establish a strong duality theorem to convert the DRO problem into a more computationally tractable form, which is a focus of our future work. We also aim to extend these results to general SCM models and relax our assumptions.

Acknowledgments

The authors thank the Max Planck Institute for Intelligent Systems, Tübingen AI Center, for supporting this project. Partial funding support was also provided by the Canada CIFAR AI Chair program.

References

Ambrosio, L.; Brué, E.; Semola, D.; et al. 2021. *Lectures on optimal transport*, volume 130. Springer.

Backhoff, J.; Beiglbock, M.; Lin, Y.; and Zalashko, A. 2017. Causal transport in discrete time and applications. *SIAM Journal on Optimization*, 27(4): 2528–2562.

Backhoff-Veraguas, J.; and Pammer, G. 2022. Stability of martingale optimal transport and weak optimal transport. *The Annals of Applied Probability*, 32(1): 721–752.

Bartl, D.; Beiglböck, M.; and Pammer, G. 2021. The Wasserstein space of stochastic processes. *arXiv preprint arXiv:2104.14245*.

Benamou, J.-D.; Carlier, G.; Cuturi, M.; Nenna, L.; and Peyré, G. 2015. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2): A1111–A1138.

Bertsimas, D.; Gupta, V.; and Kallus, N. 2018. Data-driven robust optimization. *Mathematical Programming*, 167: 235–292.

Blanchet, J.; Kang, Y.; and Murthy, K. 2019. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3): 830–857.

Blanchet, J.; Li, J.; Lin, S.; and Zhang, X. 2024. Distributionally robust optimization and robust statistics. *arXiv preprint arXiv:2401.14655*.

Bolley, F.; Guillin, A.; and Villani, C. 2007. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137: 541–593.

Chaouach, L. M.; Boskos, D.; and Oomen, T. 2022. Uncertain uncertainty in data-driven stochastic optimization: towards structured ambiguity sets. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, 4776–4781. IEEE.

Chaouach, L. M.; Oomen, T.; and Boskos, D. 2023. Comparing structured ambiguity sets for stochastic optimization: Application to uncertainty quantification. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, 8274–8279. IEEE.

Cheridito, P.; and Eckstein, S. 2023. Optimal transport and Wasserstein distances for causal models. *arXiv preprint arXiv:2303.14085*.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Dedecker, J.; and Merlevède, F. 2019. Behavior of the empirical Wasserstein distance in $\mathbb{R}^d$ under moment conditions. *Electronic Journal of Probability*, 24(none): 1 – 32.

Duchi, J. C.; and Namkoong, H. 2021. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3): 1378–1406.

Eckstein, S.; and Pammer, G. 2024. Computational methods for adapted optimal transport. *The Annals of Applied Probability*, 34(1A): 675–713.

Flamary, R.; and contributors. 2023. CO-Optimal Transport (COOT). https://github.com/PythonOT/COOT. GitHub repository.

Fournier, N.; and Guillin, A. 2015. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(1-2): 707–738.

Gao, R.; Chen, X.; and Kleywegt, A. J. 2017. Wasserstein distributionally robust optimization and variation regularization. *arXiv preprint arXiv:1712.06050*.

Guo, X.; Hong, J.; and Yang, N. 2017. Ambiguity set and learning via Bregman and Wasserstein. *arXiv preprint arXiv:1705.08056*.

Khamis, A.; Tsuchida, R.; Tarek, M.; Rolland, V.; and Petersson, L. 2024. Scalable Optimal Transport Methods in Machine Learning: A Contemporary Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lab, N. 2024. Distributionally Robust Optimization (DRO). https://github.com/namkoong-lab/dro. Accessed: 2024-08-14.

Lam, H. 2016. Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research*, 41(4): 1248–1275.

Lassalle, R. 2018. Causal transport plans and their Monge–Kantorovich problems. *Stochastic Analysis and Applications*, 36(3): 452–484.

Love, D.; and Bayraksan, G. 2015. Phi-divergence constrained ambiguous stochastic programs for data-driven optimization. *Technical report, Department of Integrated Systems Engineering, The Ohio State University, Columbus, Ohio*.

Mohajerin Esfahani, P.; and Kuhn, D. 2018. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1): 115–166.

Montesuma, E. F.; Mboula, F. N.; and Souloumiac, A. 2023. Recent advances in optimal transport for machine learning. *arXiv preprint arXiv:2306.16156*.

Nasr-Esfahany, A.; Alizadeh, M.; and Shah, D. 2023. Counterfactual identifiability of bijective causal models. In *International Conference on Machine Learning*, 25733–25754. PMLR.

Pearl, J. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Peters, J.; Janzing, D.; and Schölkopf, B. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Peyré, G.; Cuturi, M.; et al. 2017. Computational optimal transport. *Center for Research in Economics and Statistics Working Papers*, 2017-86.

Rahimian, H.; and Mehrotra, S. 2022. Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 3: 1–85.

Titouan, V.; Redko, I.; Flamary, R.; and Courty, N. 2020. Co-optimal transport. *Advances in neural information processing systems*, 33: 17559–17570.

Tran, Q. H.; Janati, H.; Redko, I.; Flamary, R.; and Courty, N. 2021. Factored couplings in multi-marginal optimal transport via difference of convex programming. *arXiv preprint arXiv:2110.00629*.

Villani, C.; et al. 2009. *Optimal transport: old and new*, volume 338. Springer.

Weed, J.; and Bach, F. 2019a. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A): 2620 – 2648.

Weed, J.; and Bach, F. 2019b. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A): 2620 – 2648.

Xu, T.; Wenliang, L. K.; Munn, M.; and Acciaio, B. 2020. Cot-gan: Generating sequential data via causal optimal transport. *Advances in neural information processing systems*, 33: 8798–8809.