PARTIAL TRANSPORTABILITY FOR DOMAIN GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning prediction models that generalize to related domains is one of the most fundamental challenges in artificial intelligence. There exists a growing literature that argues for learning invariant associations using data from multiple source domains. However, whether invariant predictors generalize to a given target domain depends crucially on the assumed structural changes between domains. Using the perspective of transportability theory, we show that invariance learning, and the settings in which invariant predictors are optimal in terms of worst-case losses, is a special case of a more general *partial* transportability task. Specifically, the partial transportability task seeks to identify / bound a conditional expectation $\mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}]$ in an unseen domain π^* using knowledge of qualitative changes across domains in the form of causal graphs and data from source domains $\pi^1, \ldots, \pi^k$. We show that solutions to this problem have a much wider generalization guarantee that subsumes those of invariance learning and other robust optimization methods that are inspired by causality. For computations in practice, we develop an algorithm that provably provides tight bounds asymptotically in the number of data samples from source domains for any partial transportability problem with discrete observables and illustrate its use on synthetic datasets.

1 INTRODUCTION

Generalization guarantees are central to the design of reliable machine learning models as the predictions and conclusions obtained in one or several source domains $\pi^1, \ldots, \pi^k$ (e.g. in controlled laboratory circumstances, from a specific study or population, etc.) are transported and applied elsewhere, in a domain π^* that may differ in several aspects from that of source domains. It is apparent that what structure and what assumptions are imposed on the relationship between domains determines whether a model will generalize as intended. For example, if the target environment is arbitrary, or substantially different from the study environment, transporting predictions is difficult or even impossible.

A structural account of causation provides suitable semantics for reasoning about the structural invariances across different domains (Pearl, 2009; Pearl & Bareinboim, 2011; Bareinboim et al., 2013). Each domain π^i is associated with a different structural causal model (SCM) M_{π^i} that differs in one or more of its component parts with respect to other domains and defines different distributions over the observed variables. In practice, the SCMs are usually not fully observable, which leads to the transportability challenge of using data from one (or more) SCMs to make inference about distributions from another SCM. A query, e.g. $\mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}]$, is said to be point identified if it can be uniquely computed given available data (from one or more domains) and qualitative knowledge about the causal changes between domains in the form of selection diagrams. However, in problems of transportability, especially when no data in the target domain can be collected, the combination of qualitative assumptions and data often does not permit one to uniquely determine a given query, which is said to be non-identifiable. In such cases, partial identification methods deal with bounding a given query e.g. $l < \mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}] < u$ in non-identifiable problems and may still serve an informative purpose for decision-making if $0 < l < u < 1$. Both settings have been studied in the literature. In particular, there exists an extensive set of graphical conditions and algorithms for the identifiability of observational, interventional, and counterfactuals distributions across domains from a combination of datasets in various settings (Pearl & Bareinboim, 2011; Bareinboim et al., 2013; Bareinboim & Pearl,

2014; 2016; Lee et al., 2020; Correa & Bareinboim, 2019). Several methods exist also for partial identification of causal effects and counterfactuals (Balke & Pearl, 1997; Chickering & Pearl, 1996; Zhang et al., 2021). Despite the generality of these results, there is still no treatment or algorithms for the partial identification of transportability queries.

In the machine learning literature, notably, a version of the transportability task is also widely studied as the problem of *domain generalization* (Wang et al., 2022). The objective is to learn a prediction function with a minimum performance guarantee on any distribution in some uncertainty set that includes potential test / target distributions (Ben-Tal et al., 2009; Gulrajani & Lopez-Paz, 2020). This problem has implicit connections to causality and SCMs if uncertainty sets of distributions are defined on the basis of "invariant correlations", such as stable conditional expectations $\mathbb{E}_{P_{\pi^1}}[y \mid \mathbf{x}] = \cdots = \mathbb{E}_{P_{\pi^k}}[y \mid \mathbf{x}]$ across training domains $\pi^1, \ldots, \pi^k$, to be used for prediction in a target domain π^* and that may be learned from data sampled across sufficiently many different domains with statistical tests (Peters et al., 2016; Subbaswamy et al., 2019; Subbaswamy & Saria, 2020) or custom loss functions (Magliacane et al., 2018; Arjovsky et al., 2019; Rojas-Carulla et al., 2018; Bellot & van der Schaar, 2020). For instance, Arjovsky et al. (2019) argue for learning representations that define an invariant optimal classifier across several training datasets. Subbaswamy et al. (2019); Subbaswamy & Saria (2020) use causal graphs and identifiable interventional distributions to define invariant prediction rules across domains. Notwithstanding their wide applicability, there is little theoretical understanding of the extrapolation guarantees that can be expected from invariant prediction rules given a finite set of domains. Correlations invariant across source domains need not be invariant in a target domain; and performance guarantees, in general, depend on the structural invariances assumed for their respective SCMs.

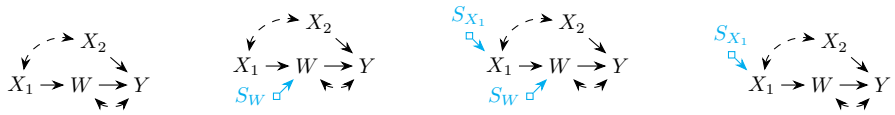
In this paper, we start by describing the conditions under which invariant prediction rules can be expected to perform well in an arbitrary target domain from first principles using the semantics of structural causal models (Pearl, 2009; Pearl & Bareinboim, 2011). We then introduce a broader optimization problem – the task of partial transportability – whose objective is to bound, instead of point estimate, a query in an arbitrary target domain of interest, such as $\mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}]$, given data from one or more source domains and qualitative knowledge about the causal changes between domains in the form of selection diagrams. We demonstrate that solutions to this problem subsume various instantiations of invariant predictors (in the conditions where these are adequate) and have a wider distributional robustness guarantee to any distribution in the target domain that is compatible with the assumed selection diagrams. For computations in practice, we show that the partial transportability task can be solved approximately for systems of variables with finite domains with a Markov Chain Monte Carlo sampling approach. The resulting bounds are sound and tight, and provide the most informative inference on a target query given the available information.

1.1 Preliminaries

We introduce in this section some basic notations and definitions that will be used throughout the paper. We use capital letters to denote variables (X), small letters for their values (x), bold letters for sets of variables ($\mathbf{X}$) and their values ($\mathbf{x}$), and use Ω to denote their domains of definition ($x \in \Omega_X$). A conditional independence statement in distribution P is written as $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_P$. A d-separation statement in some graph $\mathcal{G}$ is written as $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}}$. For convenience, we denote by $P(\mathbf{x})$ probabilities $P(\mathbf{X} = \mathbf{x})$, and $\mathbb{1}\{\cdot\}$ for the indicator function equal to 1 is the statement in $\{\cdot\}$ evaluates to true, and equal to 0 otherwise. All proofs are given in the Appendix.

We use the language of structural causal models (SCMs) (Definition 7.1.1 (Pearl, 2009)) to define the semantics of causality. An SCM M is a tuple $M = \langle \mathbf{V}, \mathbf{U}, \mathbf{F}, P \rangle$ where $\mathbf{V}$ is a set of endogenous variables and $\mathbf{U}$ is a set of exogenous variables. Each exogeneous variable $U \in \mathbf{U}$ is distributed according to a probability measure $P(u)$. $\mathbf{F}$ is a set of functions where each $f_V \in \mathbf{F}$ determines the deterministic dependencies of V on other parts of the system. That is, $v := f_V(\boldsymbol{pa}_V, \boldsymbol{u}_V)$, with $\mathbf{Pa}_V \subset \mathbf{V}$, and $\mathbf{U}_V \subset \mathbf{U}$, the exogeneous sources of variation that influence V. With this construction, we define the potential response $\mathbf{V}(\boldsymbol{u})$ to be the solution of $\mathbf{V}$ in the model M given $\mathbf{U} = \boldsymbol{u}$. Moreover, drawing values of exogenous variables $\mathbf{U}$ following the probability measure P induces a joint distribution over observables given by,

$$P(\mathbf{v}) = \int_{\Omega_{\mathbf{U}}} \prod_{V \in \mathbf{V}} \mathbb{1}\{f_V(\boldsymbol{pa}_V, \boldsymbol{u}_V) = v\} dP(\boldsymbol{u}). \tag{1}$$



(a) $\mathcal{G}^*$ (b) $\mathcal{G}^{*,a}$ (c) $\mathcal{G}^{*,b}$ (d) $\mathcal{G}^{a,b}$

Figure 1: Example of graphs: (a) Causal graph of target domain π^*, (b) selection diagram that compares domains π^* with π^a, (c) selection diagram that compares domains π^* with π^b, (d) selection diagram that compares domains π^a with π^b.

An SCM induces a causal graph $\mathcal{G}$ in which each variable in $\mathbf{V}$ is associated with a node; we draw a directed arrow between two variables $X \to Y$ if $X \in \mathbf{Pa}_Y$ appears as an argument of f_Y in the SCM, and a bidirected arrow $X \leftrightarrow Y$ if $\mathbf{U}_X \cap \mathbf{U}_Y \neq \emptyset$, that is X and Y share an unobserved confounder. The set of parent nodes of $\mathbf{X}$ in $\mathcal{G}$ is denoted by $pa(\mathbf{X})_{\mathcal{G}} = \cup_{X \in \mathbf{X}} pa(X)_{\mathcal{G}}$. Its capitalized version Pa includes the argument as well, e.g. $Pa(\mathbf{X})_{\mathcal{G}} = pa(\mathbf{X})_{\mathcal{G}} \cup \mathbf{X}$. Similar definitions are used for children ch, descendants de, etc.

2 Domain generalization through the lens of transportability

We adopt the setting of domain generalization. We assume access to k observable / source domains $\pi^1, \pi^2, \ldots, \pi^k$ with associated data distributions over a common set of variables $\mathbf{V}$ denoted $P_{\pi^1}(\mathbf{v}), P_{\pi^2}(\mathbf{v}), \ldots, P_{\pi^k}(\mathbf{v})$. Our focus is on a query, such as $\mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}]$, to be evaluated in a target domain π^* (potentially) different from source domains, where typically Y is an outcome variable, $\mathbf{X}$ is a set of covariates, and $Y \cup \mathbf{X} = \mathbf{V}$.

For concreteness, consider a medical study where patient data was collected under different treatment protocols in an attempt to assess, in a target hospital π^*, the prognosis of neurodegenerative diseases such as Alzheimer's in patients with a number of existing conditions. In the causal graph in Fig. 1a, X_1 and X_2 are treatments for hypertension and clinical depression, respectively, both known to be causes of neurodegenerative diseases Y. In the case of hypertension, the effect is mediated by blood pressure W, whose effect on neurodegenerative diseases is confounded, since both conditions share important confounding factors such physical activity levels and diet patterns (Skoog & Gustafson, 2006) (graphically encoded through the bidirected arrows). Hypertension and clinical depression are not known to affect each other (no direct link between them), although it's common for patients with clinical depression to simultaneously be at risk of hypertension (Meng et al., 2012). In this example, we have access to an observational study conducted in a hospital π^a, and to a randomized clinical trial π^b for X_1, both of which however are known to have a different high blood pressure W incidence than that in π^*. These differences are called domain discrepancies.

Definition 1 (Domain Discrepancy)**.** *Let π_a and π_b be domains associated, respectively, with SCMs M^a and M^b and causal diagrams $\mathcal{G}^a$ and $\mathcal{G}^b$. We denote by $\Delta^{a,b} \subset \mathbf{V}$ a set of variables such that, for every $V_i \in \Delta^{a,b}$, there might exist a discrepancy if $f^a_{V_i} \neq f^b_{V_i}$ or $P^a(\mathbf{U}_i) \neq P^b(\mathbf{U}_i)$.*

Definition 2 (Selection diagram)**.** *Given domain discrepancies $\Delta^{a,b}$ between two domains π^a and π^b and a causal graph $\mathcal{G}^a = (\mathbf{V}, \mathbf{E})$, let $\mathbf{S} = \{S_V : V \in \Delta^{a,b}\}$ be called selection nodes. Then, a selection diagram $\mathcal{G}^{a,b}$ is defined as a graph $(\mathbf{V} \cup \mathbf{S}, \mathbf{E} \cup \{S_V \to V\}_{S_V \in \mathbf{S}})$.*

The S-nodes locate the mechanisms where structural discrepancies between the two domains are suspected to take place. The absence of a selection node pointing to a variable represents the assumption that the mechanism responsible for assigning value to that variable is identical in both domains. In the medical example above, Fig. 1b shows a selection diagram comparing domains π^a and π^* in which the S_W node indicates a structural difference in the assignment of W, either $f^*_W \neq f^a_W$ and / or $P_{\pi^*}(u_W) \neq P^a(u_W)$, but not in the assignment of other variables, for instance $f^*_Y = f^a_Y$ and $P_{\pi^*}(u_Y) = P^a(u_Y)$. Fig. 1c and Fig. 1d are selection diagrams that compare domains (π^b, π^*) and (π^a, π^b) respectively.

2.1 INVARIANCE LEARNING FOR DOMAIN GENERALIZATION

It is apparent that there is a degree of unidentifiability in optimal prediction rules in a target domain depending on the structural differences between it and the available data. In the literature on domain generalization, a natural objective for a chosen prediction function is to minimize worst-case losses over an uncertainty set of potential target distributions that are compatible with a set of selection diagrams $\{\mathcal{G}^{i,*} : i = 1, \dots, k\}$,

$$\arg\min_{f} \max_{M \in \mathcal{M}(\mathcal{G}^*)} \mathbb{E}_{P_M}[(y - f(\mathbf{x}))^2], \tag{2}$$

where $\mathcal{M}(\mathcal{G}^*)$ is the family of SCMs compatible with the causal graph $\mathcal{G}^*$. In the literature, selection diagrams $\{\mathcal{G}^{i,*} : i = 1, \dots, k\}$ are mostly implicit. In particular, invariant predictors (Arjovsky et al., 2019; Peters et al., 2016; Lu et al., 2021; Bellot & van der Schaar, 2020; Rojas-Carulla et al., 2018; Magliacane et al., 2018) have been developed to be agnostic of the underlying causal structure, e.g. instead optimizing over distributions $P \in \mathcal{P}$ in a chosen uncertainty set of distributions (for example, defined as a convex hull of source data distributions). Using the language of selection diagrams to specify structural invariances, a statistically invariant predictor is defined as follows.

Definition 3 (Statistically invariant predictor)**.** *A statistically invariant predictor is given by the conditional expectation $\mathbb{E}[y \mid \mathbf{z}]$, where $\mathbf{Z} \subset V$ is the maximal subset of the Markov blanket of Y such that $(Y \perp\!\!\!\perp \mathbf{S} \mid \mathbf{Z})_{\mathcal{G}^{i,j}}$ for all $i, j = 1, \dots, k$.*[1]

In words, statistically invariant predictors seek the best prediction rule that is stable across source domains. Such a prediction rule, however, does not necessarily exist. For instance, for the purpose of neurodegenrative disease prediction in Fig. 1, no set $\mathbf{Z}$ d-separates Y from $\mathbf{S}$. It holds also that even if it did, invariance among source domains does not guarantee invariance in the target domain. For example, $\mathbb{E}_{P_{\pi^1}}[y \mid x_1, x_2] = \mathbb{E}_{P_{\pi^2}}[y \mid x_1, x_2]$ in $\mathcal{G}^{1,2} = \{S_{X_1} \to X_1 \to Y \to X_2\}$ qualifies as a statistically invariant predictor but $\mathbb{E}_{P_{\pi^1}}[y \mid x_1, x_2] \neq \mathbb{E}_{P_{\pi^*}}[y \mid x_1, x_2]$ if $\mathcal{G}^{1,*} = \{X_1 \to Y \to X_2 \leftarrow S_{X_2}\}$. This example shows that the set of distributions we can expect a statistically invariant predictor to generalize to is constrained and depends on the kind of changes to be expected in the target domain. The following Theorem states the conditions on $\mathcal{M}(\mathcal{G}^*)$ precisely.

Theorem 1 (Optimality of statistically invariant predictors)**.** *Statistically invariant predictors are a solution to the robust optimization problem in Eq. (2) over all SCMs M with graph $\mathcal{G}^*$ such that $(Y \perp\!\!\!\perp \mathbf{S} \mid \mathbf{Z})_{\mathcal{G}^{*,i}}$ for all $i = 1, \dots, k$.*

In other words, a conditional expectation is only statistically invariant in the target domain by assumption. Note, similarly, that there is no guarantee that causal parents form predictors that generalize, i.e. $(Y \perp\!\!\!\perp \mathbf{S} \mid \mathbf{Pa}_Y)_{\mathcal{G}^{*,i}}$ is not always guaranteed. For example, $\mathbb{E}_{P_{\pi^a}}[y \mid \boldsymbol{pa}_Y] \neq \mathbb{E}_{P_{\pi^b}}[y \mid \boldsymbol{pa}_Y] \neq \mathbb{E}_{P_{\pi^*}}[y \mid \boldsymbol{pa}_Y]$ in the neurological disease prediction example.

Corollary 1 (Optimality of causal parents)**.** *A prediction function of the form $\mathbb{E}[y \mid \boldsymbol{pa}_Y]$ is a solution to the robust optimization problem in Eq. (2) over all SCMs M with graph $\mathcal{G}^*$ such that $\mathbf{Pa}_Y$ is the maximal subset of the Markov blanket of Y in $\mathcal{G}^*$, such that $(Y \perp\!\!\!\perp \mathbf{S} \mid \mathbf{Pa}_Y)_{\mathcal{G}^{i,j}}$ and $(Y \perp\!\!\!\perp \mathbf{S} \mid \mathbf{Pa}_Y)_{\mathcal{G}^{*,i}}$ for all $i, j = 1, \dots, k$.*

A strictly more general approach to defining invariant prediction rules has been proposed by Subbaswamy et al. (2019); Subbaswamy & Saria (2020). The authors advocate for the use of intervention invariant conditional expectations if such an expression can be identified.

Definition 4 (Intervention invariant predictor)**.** *The conditional expectation $\mathbb{E}_{P_{\pi^i}}[y \mid \mathbf{z} \backslash \mathbf{m}, do(\mathbf{m})]$ is called an intervention invariant predictor if there exists π^i among training domains $\pi^1, \dots, \pi^k$ such that $(Y \perp\!\!\!\perp \mathbf{S} \mid \mathbf{Z} \backslash \mathbf{M})_{\mathcal{G}^{*,i}_{\overline{M}}}$ and the expectation itself can be computed from available data*[2]*.*

For example, in Fig. 1b, although $\mathbb{E}_{P_{\pi^a}}[y \mid x_2] \neq \mathbb{E}_{P_{\pi^*}}[y \mid x_2]$, the conditional expectation in a model where W has been intervened on $\mathbb{E}_{P_{\pi^a}}[y \mid x_2, do(w)] = \mathbb{E}_{P_{\pi^*}}[y \mid x_2, do(w)]$ because

[1] Other invariance assumptions have also been made, e.g. $(\mathbf{Z} \perp\!\!\!\perp \mathbf{S})_{\mathcal{G}^{i,j}}$ and $(Y \perp\!\!\!\perp \mathbf{S} \mid \mathbf{X})_{\mathcal{G}^{i,j}}$ for problems where only the distribution of some covariates is expected to change across domains (Muandet et al., 2013), or such as $(\mathbf{Z} \perp\!\!\!\perp \mathbf{S} \mid Y)_{\mathcal{G}^{i,j}}$ (Li et al., 2018). Problems where the magnitude of changes is assumed to be bounded, i.e. $\mid P_{\pi^*}(y \mid \mathbf{x}) - P_{\pi^i}(y \mid \mathbf{x}) \mid \leqslant c$, instead of restricting the d-separation statements involving $\mathbf{S}$ have been studied by (Rothenhäusler et al., 2021).

[2] $\mathcal{G}_{\overline{W}}$ is the graph $\mathcal{G}$ in which incoming edges into W have been removed.

X_2 d-separates Y from all selection nodes in the intervened graph, i.e. $(Y \perp\!\!\!\perp \mathbf{S} \mid X_2)_{\mathcal{G}^{*,a}_{\overline{W}}}$. In this example, however, the interventional expectation is not identified, and in fact no intervention invariant predictor exists. If identified, intervention invariant predictors have the following generalization guarantee (Subbaswamy et al., 2019).

Theorem 2 (Optimality of intervention invariant predictors)**.** *Let $\mathbf{M} = ch(\mathbf{S})_{\mathcal{G}^i}$ and $\mathbf{Z} \subset V$ be the maximal subset of the Markov blanket of Y in $\mathcal{G}^i$ where π^i is a training domain chosen such that $(Y \perp\!\!\!\perp \mathbf{S} \mid \mathbf{Z} \backslash \mathbf{M})_{\mathcal{G}^{*,i}_{\overline{M}}}$. If identified, the intervention invariant predictor $\mathbb{E}_{P_{\pi^i}}[y \mid \mathbf{z} \backslash \mathbf{m}, do(\mathbf{m})]$ is a solution to the robust optimization problem in Eq. (2) over all SCMs M with graph $\mathcal{G}^*$ compatible with $\mathcal{G}^{*,i}$.*

In both Thms. 1 and 2, the set $\mathcal{M}(\mathcal{G}^*)$ over which predictors are optimal is relatively constrained. Whether it is of interest or plausible depends on the application at hand. Moreover, solutions to a worst-case optimization problem say nothing about the *range* of values optimal prediction functions $\mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}]$ may take in other distributions P_{π^*} away from the worst-case.

3 Partial Transportability of Statistical Relations

The uncertainty and inherent under-identifiability of solutions to domain generalization problems is not quantified by solutions to worst-case optimization problem. This motivates us to define the task of *partial* transportability that extends the literature on domain generalization by considering bounds on the value of arbitrary queries $\mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}]$ in arbitrary target domains π^* defined by a set of selection diagrams $\{\mathcal{G}^{i,*} : i = 1, \ldots, k\}$.

Task (Partial Transportability). *Derive a tight bound $[l, u]$ over a query of the form $\mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}]$ with knowledge of selection diagrams $\{\mathcal{G}^{*,i} : i = 1, \ldots, k\}$, a corresponding collection of data distributions $\{P_i(\mathbf{v}) : i = 1, \ldots, k\}$, and set of intervals $\{I_j : V_j \in \bigcup_i \Delta^{*,i}\}$ that define potential constraints on probabilities in the target domain. Algorithmically, this may be written as a solution to the following optimization problem,*

$$\min_{M \in \mathcal{M}(\mathcal{G}^*)} / \max \mathbb{E}_{P_M}[y \mid \mathbf{x}], \quad \text{such that} \quad \forall V \notin \Delta^{*,i} : f_V^{(*)} = f_V^{(i)}, P_{\pi^*}(\boldsymbol{u}_V) = P_{\pi^i}(\boldsymbol{u}_V),$$
$$\text{and} \quad \forall V \in \bigcup_i \Delta^{*,i} : P_{\pi^*}(v \mid \boldsymbol{pa}_V) \in I_V. \tag{3}$$

In words, the task is to evaluate the minimum and maximum values over all possible SCMs M compatible with $\{\mathcal{G}^{*,i} : i = 1, \ldots, k\}$ that define the structurally invariant mechanisms in the system and (potentially uninformative, i.e., $I_V \in [0, 1]$) assumptions about target-specific probabilities.

For example, given the causal description of the protocols presented in the introductory medical example and Fig. 1, the question might be how to combine these various datasets to predict an individual's risk of developing neurodegenerative diseases in π^*? The optimal prediction function is given by $\mathbb{E}_{P_{\pi^*}}[y \mid w, x_1, x_2]$ (under mean squared error losses), which may be written as,

$$\mathbb{E}_{P_{\pi^*}}[y \mid w, x_1, x_2] = \sum_y y P_{\pi^*}(y, w, x_1, x_2) / \sum_y P_{\pi^*}(y, w, x_1, x_2), \tag{4}$$

where $P_{\pi^*}(y, w, x_1, x_2)$ is equal to,

$$\int_{\Omega_{\mathbf{U}}} \underbrace{\mathbb{1}\{f_Y^{(*)}(w, x_2, u_{wy}) = y\}}_{\text{matches RCT } \pi^b} \underbrace{\mathbb{1}\{f_W^{(*)}(x_1, u_{wy}) = w\}}_{\text{specific to } \pi^*} \underbrace{\mathbb{1}\{f_{X_1,X_2}^{(*)}(u_{x_1x_2}) = x_1, x_2\}}_{\text{matches hospital } \pi^a} dP_{\pi^*}(\boldsymbol{u}). \tag{5}$$

This is a mixture of terms for which data from source domains can be leveraged, for example $f_Y^{(*)}(w, x_2, u_{wy}) = f_Y^{(a)}(w, x_2, u_{wy})$ (superscripts denote domain), but also involves unobserved confounders u_{wy} and $u_{x_1x_2}$ that cannot be marginalized out, and terms that are specific to π^*. In addition, although $P_{\pi^*}(w \mid x_1)$ is known to differ in our target medical study it may be the case that we have some domain knowledge that constrains it, e.g. $P_{\pi^*}(w \mid x_1) \in [0.2, 0.7] =: I_w$ (if left undetermined, $I_w := [0, 1]$). The following propositions show that the partial transportability task extends the literature on invariance learning in the sense that invariant predictors (as defined in Defs. 3 and 4) are special solutions to the transportability task in Eq. (3) when the selection diagrams and target graphs are specified as in Thms. 1 and 2 and the query is identifiable.

Proposition 1 (Statistical invariance as partial transportability)**.** *Let $\mathbf{Z} \subset V$ be the maximal subset of the Markov blanket of Y such that $(Y \perp\!\!\!\perp \mathbf{S} \mid \mathbf{Z})_{\mathcal{G}^{i,j}}$ for all $i, j = 1, \dots, k$. For a given graph $\mathcal{G}^*$, if $\min_{M \in \mathcal{M}(\mathcal{G}^*)} \mathbb{E}_{P_M}[y \mid \mathbf{z}] = \max_{M \in \mathcal{M}(\mathcal{G}^*)} \mathbb{E}_{P_M}[y \mid \mathbf{z}]$ then $f(\mathbf{z}) := \min_{M \in \mathcal{M}(\mathcal{G}^*)} \mathbb{E}_{P_M}[y \mid \mathbf{z}]$ is the optimal statistically invariant predictor.*

Proposition 2 (Intervention invariance as partial transportability)**.** *Let $\mathbf{M} = ch(\mathbf{S})_{\mathcal{G}^i}$ and $\mathbf{Z} \subset V$ be the maximal subset of the Markov blanket of Y in $\mathcal{G}^i$ where π^i is a training domain chosen such that $(Y \perp\!\!\!\perp \mathbf{S} \mid \mathbf{Z} \backslash \mathbf{M})_{\mathcal{G}^{*,i}_{\overline{M}}}$. If $\min_{M \in \mathcal{M}(\mathcal{G}^*)} \mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{z} \backslash \mathbf{m}, do(\mathbf{m})] = \max_{M \in \mathcal{M}(\mathcal{G}^*)} \mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{z} \backslash \mathbf{m}, do(\mathbf{m})]$ then $f(\mathbf{z} \backslash \mathbf{m}, \mathbf{m}) := \min_{M \in \mathcal{M}(\mathcal{G}^*)} \mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{z} \backslash \mathbf{m}, do(\mathbf{m})]$ is the optimal intervention invariant predictor.*

The following theorem presents a wider generalization guarantee for solutions to the partial transportability task. In particular, there is no need for requirements on sets $\mathbf{Z}$ or $\mathbf{M}$ or any assumed separation statements or identifiability criteria, and therefore the set of SCMs $\mathcal{M}(\mathcal{G}^*)$ is more general than considered in the literature.

Theorem 3 (General worst-case solutions as partial transportability)**.** *Let $\mathbf{X} = \mathbf{V} \backslash Y$. Given a query $\mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}]$ and an arbitrary set of selection diagrams $\{\mathcal{G}^{*,i} : i = 1, \dots, k\}$, let $[\hat{l}(\mathbf{x}), \hat{u}(\mathbf{x})]$ be the solution to the partial transportability task in Eq. (3). Then, $f(\mathbf{x}) := \hat{l}(\mathbf{x})$ is the solution to the worst-case loss minimization problem in Eq. (2) over all SCMs M with graph $\mathcal{G}^*$.*

In words, lower bounds computed from the partial transportability task are a solution to the robust optimization problem in Eq. (2) for *any* chosen set $\mathcal{M}(\mathcal{G}^*)$. Fig. 2 sketches the relative constraints imposed by different methods on the space of potential target SCMs $\mathcal{M}(\mathcal{G}^*)$. In red we illustrate the requirements of statistically invariant predictors, i.e. $P_{\pi^*}(y \mid \mathbf{z}) = P_{\pi^1}(y \mid \mathbf{z}) = \cdots = P_{\pi^k}(y \mid \mathbf{z})$; in blue the requirements of intervention invariant predictors (applicable in a strictly larger set). In turn, solutions to the partial transportability task are robust in the whole space of SCMs compatible with the causal graph even if, for instance, $P_{\pi^*}(y \mid \mathbf{z}) \neq P_{\pi^i}(y \mid \mathbf{z}, do(\mathbf{w})) \neq P_{\pi^j}(y \mid \mathbf{z})$ for all source domains π^i, π^j.

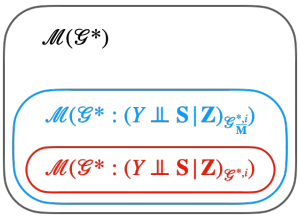


Figure 2: Relative robustness of domain generalization methods.

4 Algorithms for partial transportability

This section presents algorithms to solve the partial transportability task for SCMs with *discrete observables*, that is each $V \in \mathbf{V}$ taking values in a finite space of outcomes, while each $U \in \mathbf{U}$ associated with an arbitrary probability density function $P(\mathbf{u})$.

A first step in our argument will be to decompose a chosen query into smaller factors so as to infer which factors can be matched across domains and point-identified from data, and subsequently re-parameterize unmatched factors by a special family of SCMs to make the bounding problem tractable. We use the concept of c-components and C-factors developed by Tian & Pearl (2002). The set $\mathbf{V}$ can be partitioned into c-components such that two variables are assigned to the same set $\mathbf{C} \subset \mathbf{V}$ if and only if they are connected by a bi-directed path in $\mathcal{G}$. In addition let $\mathbf{U_C} = \bigcup_{V_i \in \mathbf{C}} \mathbf{U}_i$ denote the set of exogeneous variables that are parents of any $V \in \mathbf{C}$. For example, the graph in Fig. 1a induces c-components $\{X_1, X_2\}$ and $\{W, Y\}$.

For any set $\mathbf{C} \subseteq \mathbf{V}$, let $Q_{\pi^i}[\mathbf{C}](pa(\mathbf{C}))$ denote the C-factor of $\mathbf{C}$ in domain π^i which is defined by,

$$Q_{\pi^i}[\mathbf{C}](\boldsymbol{pa_c}) = \int_{\Omega_{\mathbf{U_C}}} \prod_{V \in \mathbf{C}} \mathbb{1}\{f_V^{(i)}(\boldsymbol{pa}_V, \boldsymbol{u}_V) = v\} dP_{\pi^i}(\boldsymbol{u_C}). \tag{6}$$

Moreover, let $\mathbb{C}$ denote the collection of c-components, then $P(\mathbf{v}) = \prod_{\mathbf{C} \in \mathbb{C}} Q[\mathbf{C}]$ and $Q[\mathbf{C}] = P(\mathbf{c} \mid \boldsymbol{pa_C})$ (we omit the dependence of each C-factor on $pa(\mathbf{C})$ for readability). This construction is useful because the joint distribution may be factorized according to the c-components of $\mathcal{G}$ and its factors matched across domains (Tian & Pearl, 2002; Correa & Bareinboim, 2019).

Lemma 1. *Let $\mathcal{G}^{a,b}$ be a selection diagram for the SCMs M^a and M^b, then $Q^a[\mathbf{C}] = Q^b[\mathbf{C}]$ if $\mathcal{G}^{a,b}$ does not contain selection nodes S_V pointing to any variable in $V \in \mathbf{C}$.*

For example, for the selection diagram in Fig. 1b, $P_{\pi^*}(\mathbf{v}) = Q^*[X_1, X_2]Q^*[W, Y]$ where by Lem. 1 $Q_{\pi^*}[X_1, X_2] = Q_{\pi^a}[X_1, X_2] = P_{\pi^a}(x_1, x_2)$, since the is no S-node pointing to X_1 or X_2. In turn, $Q_{\pi^*}[W, Y] \neq Q_{\pi^a}[W, Y]$ because of the S-node pointing to W. Note, however, that $Q_{\pi^*}[W, Y]$ defined as in Eq. (6) involves terms, e.g. $\mathbb{1}\{f_Y(w, x_1, u_{wy}) = y\}$, that are invariant across domains since the absence of an S-node into Y denotes invariance in causal mechanisms, and for which $P_{\pi^a}(\mathbf{v})$ may be used for estimation. We discuss next a re-parameterization of C-factors $Q_{\pi^*}[\mathbf{C}]$ that cannot be matched across domains with the goal of defining a tractable constrained optimization problem to bound $Q_{\pi^*}[\mathbf{C}]$.

Theorem 4. *Let M be an arbitrary SCM with graph $\mathcal{G}$ and let $\mathbf{C}$ be any c-component. Then, there exists a corresponding SCM N with finite exogeneous domain compatible with $\mathcal{G}$ such that $Q_M[\mathbf{C}] = Q_N[\mathbf{C}]$, where for every exogenous variable $U \in \mathbf{U_C}$, its cardinality $d_U = \left|\Omega_{Pa(\mathbf{C})}\right|$.*

This Theorem shows that SCMs with discretely-valued exogeneous variables are expressive enough to represent C-factors $Q[\mathbf{C}]$ irrespective of the true underlying data generating mechanism. From an optimization perspective, this is useful because it allows us to consistently parameterize C-factors and make inference on its distribution in a well-defined latent variable model (Rosset et al., 2017; Zhang et al., 2021). As an example, consider the introductory example with $\{X_1, X_2, Y, W\}$ binary and causal graphs in Fig. 1. $Q^*[W, Y]$ defined using Eq. (6) can also be written as:

$$\sum_{u_{wy}, u_y, u_w} \mathbb{1}\{f_Y^{(a)}(w, x_2, u_{wy}, u_y) = y\}\mathbb{1}\{f_W^{(*)}(x_1, u_{wy}, u_w) = w\}P_{\pi^a}(u_{wy}, u_y)P_{\pi^*}(u_w), \quad (7)$$

where $|\Omega_{U_{wy}}| = |\Omega_{U_w}| = |\Omega_{U_y}| = |\Omega_{X_1}| \cdot |\Omega_{X_2}| \cdot |\Omega_W| \cdot |\Omega_Y| = 16$; the function f_V is a mapping between finite domains $\Omega_{\mathbf{Pa}_V} \times \Omega_{\mathbf{U}_V} \mapsto \Omega_V$ for $V \in \{W, Y\}$. Moreover, we have used the structural invariances encoded by the selection diagrams in Fig. 1 to match causal mechanisms and exogeneous probabilities between domains. In particular, $P_{\pi^a}(u_y) = P_{\pi^b}(u_y) = P_{\pi^*}(u_y)$ by definition of the selection diagrams $\mathcal{G}^{a,*}$ and $\mathcal{G}^{b,*}$. Although discretely-valued causal mechanisms and exogeneous probabilities imply well-defined parameters to optimize over, the partial transportability task remains a difficult constrained optimization problem.

4.1 Approximations via Gibbs sampling

We follow (Chickering & Pearl, 1996; Zhang et al., 2021; Bellot et al., 2022) and take a Bayesian perspective to approximating bounds $[l, u]$. We evaluating credible intervals $P(l < \mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}] < u \mid \bar{\mathbf{v}}) = 1 - \alpha$ on the posterior of $\mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}]$ by approximating the expectation,

$$\mathbb{E}[\mathbb{1}\{l < \mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}] < u\} \mid \bar{\mathbf{v}}] = P(l < \mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}] < u \mid \bar{\mathbf{v}}) \quad (8)$$

provided with finite samples $\bar{\mathbf{v}} := (\bar{\mathbf{v}}_{\pi^1}, \ldots, \bar{\mathbf{v}}_{\pi^k})$, where $\bar{\mathbf{v}}_{\pi^i} = \{\mathbf{v}_{\pi^i}^{(j)} : j = 1, \ldots, n_i\}$ are n_i independent sampled collected in domain π^i and a set of selection diagrams $\{\mathcal{G}^{*,i} : i = 1, 2, \ldots, k\}$ using Gibbs sampling. Following the arguments in the previous section the query may be reduced to bounding a C-factor of the form,

$$\omega_{obj} := Q_{\pi^*}[\mathbf{C}] = \sum_{U \in \mathbf{U_C}} \sum_{u=1,\ldots,d_U} \prod_{V \in \mathbf{C}} \mathbb{1}\{\xi_V^{(\boldsymbol{pa}_V, \boldsymbol{u}_V)} = v\} \prod_{U \in \mathbf{U_C}} \theta_u. \quad (9)$$

that are parameterized by $\boldsymbol{\xi} = \{\xi_V^{(\boldsymbol{pa}_V, \boldsymbol{u}_V)} : V \in \mathbf{C}, \mathbf{Pa}_V \subset \mathbf{V}, \mathbf{U}_V \subset \mathbf{U_C}\}$ and $\boldsymbol{\theta} = \{\theta_u : U \in \mathbf{U_C}\}$ that represent causal functional assignments and exogeneous probabilities, respectively. We have dropped the domain indicator "*" from the definition of parameters for readability.

The Gibbs sampler starts with some initial value for all latent quantities $(\boldsymbol{u}, \boldsymbol{\xi}, \boldsymbol{\theta})$ in the expression of ω_{obj}, and iterates over the following sampling steps, each parameter conditioned on the current values of the remaining terms in the parameter vector.

1. *Sample* $\mathbf{u}$. Let $u \in \Omega_U, U \in \mathbf{U_C}$. For each observed data example across all domains $\mathbf{v}^{(n)} \in \bar{\mathbf{v}}$, $n = 1, \ldots, \sum_i n_i$, we sample corresponding exogeneous variables $U \in \mathbf{U_C}$ from the conditional distribution,

$$P(\mathbf{u}^{(n)} \mid \mathbf{v}^{(n)}, \boldsymbol{\xi}, \boldsymbol{\theta}) \propto P(\mathbf{u}^{(n)}, \mathbf{v}^{(n)} \mid \boldsymbol{\xi}, \boldsymbol{\theta}) = \prod_{V \in \mathbf{C}} \mathbb{1}\{\xi_V^{(\boldsymbol{pa}_V^{(n)}, \boldsymbol{u}_V^{(n)})} = v^{(n)}\} \prod_{U \in \mathbf{U_C}} \theta_{\mathbf{u}}. \quad (10)$$

2. *Sample* $\boldsymbol{\xi}$. Parameters $\boldsymbol{\xi}$ define deterministic causal mechanisms. For a given parameter $\xi_V^{(\boldsymbol{pa}_V, \boldsymbol{u}_V)} \in \boldsymbol{\xi}$ its conditional distribution is given by $P(\xi_V^{(\boldsymbol{pa}_V, \boldsymbol{u}_V)} = v \mid \bar{\mathbf{v}}, \bar{\mathbf{u}}) = 1$ if there exists a sample $(\mathbf{v}^{(n)}, \mathbf{pa}_V^{(n)}, \mathbf{u}^{(n)})$ for some n, where n iterates over the samples of $\mathbf{u}$ from step 1 and $\mathbf{v}$ associated with the *subset* of domains in which exogeneous probabilities match the target domain, such that $\xi_V^{(\boldsymbol{pa}_V^{(n)}, \boldsymbol{u}_V^{(n)})} = v^{(n)}$. Otherwise, $P(\xi_V^{(\boldsymbol{pa}_V, \boldsymbol{u}_V)} = v \mid \bar{\mathbf{v}}, \bar{\mathbf{u}})$ is given by a uniform discrete distribution over its domain Ω_V.
3. *Sample* $\boldsymbol{\theta}$. Let $\theta_u = (\theta_{u_1}, \ldots, \theta_{u_d}) \in \boldsymbol{\theta}$ be the parameters that define the probability vector of possible values of variables $U \in \mathbf{U_C}$. Its conditional distribution is given by,

$$\theta_{u_1}, \ldots, \theta_{u_d} \mid \bar{\mathbf{v}}, \bar{\mathbf{u}} \sim \text{Dirichlet}\left(\sum_n \mathbb{1}\{u^{(n)} = u_1\}, \ldots, \sum_n \mathbb{1}\{u^{(n)} = u_d\}\right), \tag{11}$$

where, similarly, n iterates over the samples of $\mathbf{u}$ from step 1 associated with the subset of domains in which exogeneous probabilities match the target domain.

In the above, we have described the conditional distributions of parameters that can be matched across domains, and therefore estimated from the subset of relevant available data. By the definition of the partial transportability task, parameters that are specific to the target domain π^* are constrained to lie in an assumed interval, e.g. $P_{\pi^*}(v \mid \boldsymbol{pa}_V) = \sum_{\boldsymbol{u}_V} \mathbb{1}\{\xi_V^{(\boldsymbol{pa}_V, \boldsymbol{u}_V)} = v_i\} \prod_{U \in \mathbf{U}_V} \theta_u \in I_V \subseteq [0, 1]$, or else left unspecified. In the first case, parameters are sampled independently and uniformly in the space defined by the constraints and in the second case, they are sampled independently and uniformly in their domain of definition, i.e. $\xi_V^{(\boldsymbol{pa}_V, \boldsymbol{u}_V)} \in \Omega_V, \theta_u \in \Omega_U$, in every step of the sampler.

Iterating this procedure forms a Markov chain with the invariant distribution to be the target posterior distribution $P(\boldsymbol{u}, \boldsymbol{\xi}, \boldsymbol{\theta} \mid \bar{\mathbf{v}})$. $P(\omega_{obj} \mid \bar{\mathbf{v}})$ is then approximated by plugging the MCMC samples into Eq. (9). The upper and lower α quantile among T samples of $P(\omega_{obj} \mid \bar{\mathbf{v}})$, when combined with the identified C-factors that form $\mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}]$, gives us a $(1 - \alpha)$ credible interval $\hat{l}_\alpha < \mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}] < \hat{u}_\alpha$ defined by,

$$\hat{l}_\alpha = \sup\{x : \sum_t \mathbb{1}\{\mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}]^{(t)} \leqslant x\} = \alpha/2\}, \tag{12}$$

$$\hat{u}_\alpha = \inf\{x : \sum_t \mathbb{1}\{\mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}]^{(t)} \leqslant x\} = 1 - \alpha/2\}. \tag{13}$$

The following Theorem shows that credible intervals $[\hat{l}_0, \hat{u}_0]$ converge to the true bounds $[l, u]$ for the unknown query $\mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}]$ and are, moreover, maximally informative, in the sense that we can always construct two data generating mechanisms M^1, M^2 for the target domain that are compatible with our current knowledge of the world such that $\mathbb{E}_{P_{M^1}}[y \mid \mathbf{x}] = l$ and $\mathbb{E}_{P_{M^2}}[y \mid \mathbf{x}] = u$.

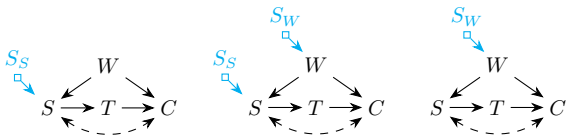
Theorem 5. *The solution* $[l, u]$ *to the partial transportability task defined over discrete SCMs is a tight bound over a target query* $\mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}]$. *The credible interval* $[\hat{l}_0, \hat{u}_0]$ *coincides with* $[l, u]$ *as* $n_i \to \infty$ *in all observable domains* π^i, $i = 1, \ldots, k$.

5 Lung cancer and smoking data

This example is inspired by the debate around the relationship between smoking and lung cancer in the 1950's (of Health & Services, 2014). We use a scientifically-grounded variation of the front-door graph that includes an individual's smoking status S, presence of tar in the lungs T, wealth W, and lung cancer status C, using the fact that smoking and lung cancer may be confounded by an individual's unobserved genetic profile. In this example, the objective is to make inference on cancer probability distributions in the French population π^{FR} from corresponding data in π^{UK} where the prevalence of smoking is known to be lower. The selection diagram is given in Fig. 4a.

Probability of cancer among smokers $P_{\pi^{\text{FR}}}(C = 1 \mid S = 1)$ **.** The C-factor decomposition and parameterization is given by the following derivations,

$$P_{\pi^{\text{FR}}}(c \mid s) = \frac{P_{\pi^{\text{FR}}}(c, s)}{\sum_c P_{\pi^{\text{FR}}}(c, s)} = \frac{\sum_{t,w} P_{\pi^{\text{FR}}}(c, s, t, w)}{\sum_{c,t,w} P_{\pi^{\text{FR}}}(c, s, t, w)} = \frac{\sum_{t,w} Q_{\pi^{\text{FR}}}[s, c] Q_{\pi^{\text{FR}}}[w] Q_{\pi^{\text{FR}}}[t]}{\sum_{c,t,w} Q_{\pi^{\text{FR}}}[s, c] Q_{\pi^{\text{FR}}}[w] Q_{\pi^{\text{FR}}}[t]}, \tag{14}$$



(a) $\mathcal{G}^{\text{FR,UK}}$ (b) $\mathcal{G}^{\text{FR,SW}}$ (c) $\mathcal{G}^{\text{SW,UK}}$

	π^{FR}
$\mathbb{E}_{P_{\pi^{\text{UK}}}}[c \mid t, w]$	.1502 (.005)
$\mathbb{E}_{P_{\pi^{\text{UK}}}}[c \mid t, w, s]$	.1448 (.004)
$\texttt{med}(\hat{l}_0, \hat{u}_0)$	.1361 (.005)
Lower bound $\hat{l}_0$	.1609 (.005)

(d) Performance comparisons.

Figure 4: (a-c) selection diagrams that compare domain π^{FR} with π^{UK}, π^{FR} with π^{SW}, and π^{SW} with π^{UK}, respectively. (d) gives mean squared error for cancer prediction on a sample of data from $P_{\pi^{\text{FR}}}$.

where $Q_{\pi^{\text{FR}}}[t] = Q_{\pi^{\text{UK}}}[t] = P_{\pi^{\text{UK}}}(t \mid s, w)$ and $Q_{\pi^{\text{FR}}}[w] = Q_{\pi^{\text{UK}}}[w] = P_{\pi^{\text{UK}}}(w)$, and,

$$Q_{\pi^{\text{FR}}}[s, c] = \sum_{u_{sc}, u_s} \mathbb{1}\{\xi_{C^{\text{UK}}}^{(w,t,u_{sc})} = c\}\mathbb{1}\{\xi_{S^{\text{FR}}}^{(w,u_{sc},u_s)} = s\}\theta_{u_{sc}}^{\text{UK}}\theta_{u_s}^{\text{FR}}.$$

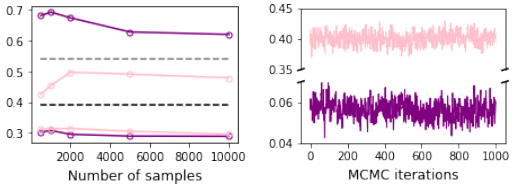


Figure 3: Bounding the probability of cancer.

In Fig. 3, we report estimated 100% credible intervals $\hat{l}_0 < P_{\pi^{\text{FR}}}(C = 1 \mid S = 1) < \hat{u}_0$ as a function of the number of samples without prior information (purple) and with the prior information that $P_{\pi^{\text{FR}}}(s \mid w)$ lies in an interval of width 0.1 around its true value (pink). The black and gray dotted lines are the actual values $P_{\pi^{\text{FR}}}(C = 1 \mid S = 1)$ and $P_{\pi^{\text{UK}}}(C = 1 \mid S = 1)$ respectively. Notice that a relatively small number of samples is required to converge to stable bounds, and that the prior information narrows the credible interval reflecting this additional constraint. We also show for illustration that our Gibbs sampler recovers the true values $P_{\pi^{\text{FR}}}(C = 1 \mid S = 1)$ (pink) and $P_{\pi^{\text{FR}}}(C = 1 \mid S = 0)$ (purple) when trained on data from π^{FR}, i.e. when probabilities are identified.

Prediction performance across domains. Consider the task of designing cancer prediction rules for optimal performance in the french population π^{FR}. We introduce an additional training domain to be able to define invariant predictors: data from the Swedish population π^{SW} whose structural differences with π^{UK} and with π^{FR} are given in Fig. 4. Across π^{UK} and π^{SW}, the optimal invariant predictor (Def. 3) is given by $\mathbb{E}_{P_{\pi^{\text{UK}}}}[c \mid t, w, s] = \mathbb{E}_{P_{\pi^{\text{SW}}}}[c \mid t, w, s]$ which, however, is not equal to $\mathbb{E}_{P_{\pi^{\text{FR}}}}[c \mid t, w, s]$ as no set blocks the open path between the selection node S_S and the cancer variable C in $\mathcal{G}^{\text{FR,UK}}$. We consider also the common strategy of using causal parents for prediction, i.e. using the prediction rule $\mathbb{E}_{P_{\pi^{\text{UK}}}}[c \mid t, w]$ (which, similarly, is not equal to $\mathbb{E}_{P_{\pi^{\text{FR}}}}[c \mid t, w]$). For comparison, we consider two plausible strategies that may be derived from bounds $[\hat{l}_0, \hat{u}_0]$ for the optimal prediction rule $\mathbb{E}_{P_{\pi^{\text{FR}}}}[c \mid t, w, s]$ computed using data from π^{UK} and π^{SW}: the median value $\texttt{med}(\hat{l}_0, \hat{u}_0)$ and the lower bound $\hat{l}_0$. We observe in Fig. 4d that indeed the prediction rule $\mathbb{E}_{P_{\pi^{\text{UK}}}}[c \mid t, w, s]$ underperforms in π^{FR}: for reference $\mathbb{E}_{P_{\pi^{\text{FR}}}}[c \mid t, w, s]$ has a mean error of .1220, cautioning against naively transporting invariant prediction rules across domains. Similarly, using causal parents for prediction underperforms. In contrast, the median of the derived bound proves to be a slightly better predictor in this case (although there are no guarantees for this behaviour in general). The lower bound solution performs worse for this particular dataset but has the guarantee of optimal performance in the "worst-case" domain compatible with the selection diagrams (Thm. 3).

6 Conclusions

This paper investigated the problem of domain generalization from the perspective of transportability theory. We introduced the task of partial transportability that seeks to bound the value of an arbitrary conditional expectation $\mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}]$ in an unseen domain π^* using selection diagrams and data from source domains. Using this formalism, we showed that invariant predictors and more general solutions to robust optimization problems derived in the literature are special cases of solutions to this task. Moreover, in systems of discrete observables, we showed that we can design provably consistent algorithms for inferring bounds that are sound and tight, and illustrated its performance on synthetic data.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

Elias Bareinboim and Judea Pearl. Transportability from multiple environments with limited experiments: Completeness results. *Advances in neural information processing systems*, 27, 2014.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

Elias Bareinboim, Sanghack Lee, Vasant Honavar, and Judea Pearl. Transportability from multiple environments with limited experiments. *Advances in Neural Information Processing Systems*, 26, 2013.

Alexis Bellot and Mihaela van der Schaar. Accounting for unobserved confounding in domain generalization. *arXiv preprint arXiv:2007.10653*, 2020.

Alexis Bellot, Junzhe Zhang, and Elias Bareinboim. Scores for learning discrete causal graphs with unobserved confounders. *Technical Report R-83, Causal AI Lab, Columbia University*, 2022.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton university press, 2009.

Constantin Carathéodory. Über den variabilitätsbereich der fourier'schen konstanten von positiven harmonischen funktionen. *Rendiconti Del Circolo Matematico di Palermo (1884-1940)*, 32(1): 193–217, 1911.

David M Chickering and Judea Pearl. A clinician's tool for analyzing non-compliance. In *Proceedings of the National Conference on Artificial Intelligence*, pp. 1269–1276, 1996.

Juan D Correa and Elias Bareinboim. From statistical transportability to estimating the effect of stochastic interventions. In *IJCAI*, pp. 1661–1667, 2019.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Sanghack Lee, Juan D Correa, and Elias Bareinboim. Generalized transportability: Synthesis of experiments from heterogeneous domains. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 2020.

Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Chaochao Lu, Yuhuai Wu, Jośe Miguel Hernández-Lobato, and Bernhard Schölkopf. Nonlinear invariant risk minimization: A causal approach. *arXiv preprint arXiv:2102.12353*, 2021.

Sara Magliacane, Thijs Van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. *Advances in neural information processing systems*, 31, 2018.

Lin Meng, Dongmei Chen, Yang Yang, Yang Zheng, and Rutai Hui. Depression increases the risk of hypertension incidence: a meta-analysis of prospective cohort studies. *Journal of hypertension*, 30 (5):842–851, 2012.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18. PMLR, 2013.

US Department of Health and Human Services. The health consequences of smoking—50 years of progress: a report of the surgeon general, 2014.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Twenty-fifth AAAI conference on artificial intelligence*, 2011.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

Denis Rosset, Nicolas Gisin, and Elie Wolfe. Universal bound on the cardinality of local hidden variables in networks. *arXiv preprint arXiv:1709.00707*, 2017.

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):215–246, 2021.

Herman Rubin and Oscar Wesler. A note on convexity in euclidean n-space. In *Proc. Amer. Math. Soc*, volume 9, pp. 522–523, 1958.

Ingmar Skoog and Deborah Gustafson. Update on hypertension and alzheimer's disease. *Neurological research*, 28(6):605–611, 2006.

Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 21(2):345–352, 2020.

Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3118–3127. PMLR, 2019.

Jin Tian and Judea Pearl. *A general identification condition for causal effects*. eScholarship, University of California, 2002.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Junzhe Zhang, Jin Tian, and Elias Bareinboim. Partial counterfactual identification from observational and experimental data. *arXiv preprint arXiv:2110.05690*, 2021.

A PROOFS

Recall the robust optimization problem of interest,

$$\arg\min_{f} \max_{M \in \mathcal{M}(\mathcal{G}^*)} \mathbb{E}_{(y,\mathbf{x}) \sim P_M}[(y - f(\mathbf{x}))^2], \tag{15}$$

where $\mathcal{M}(\mathcal{G}^*)$ is the family of SCMs compatible with the causal graph $\mathcal{G}^*$.

A statistically invariant predictor is given by the conditional expectation $\mathbb{E}[y \mid \mathbf{z}]$, where $\mathbf{Z} \subset V$ is the maximal subset of the Markov blanket of Y such that $(Y \perp\!\!\!\perp \mathbf{S} \mid \mathbf{Z})_{\mathcal{G}^{i,j}}$ for all $i, j = 1, \ldots, k$. This implies that $\mathbb{E}_{P_{\pi^1}}[y \mid \mathbf{z}] = \cdots = \mathbb{E}_{P_{\pi^k}}[y \mid \mathbf{z}]$.

Theorem 6 (Thm. 1 restated)**.** *Statistically invariant predictors are a solution to the robust optimization problem in Eq. (15) over all SCMs M with graph $\mathcal{G}^*$ such that $(Y \perp\!\!\!\perp \mathbf{S} \mid \mathbf{Z})_{\mathcal{G}^{*,i}}$ for all $i = 1, \ldots, k$.*

Proof. Under the condition that $(Y \perp\!\!\!\perp \mathbf{S} \mid \mathbf{Z})_{\mathcal{G}^{*,i}}$ for all $i = 1, \ldots, k$ it holds that $\mathbb{E}_{P_{\pi^1}}[y \mid \mathbf{z}] = \cdots = \mathbb{E}_{P_{\pi^k}}[y \mid \mathbf{z}] = \mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{z}]$. And since $\mathbf{Z} \subset V$ is defined as the maximal subset of the Markov blanket of Y such that the above equalities hold, it is the optimal solution to Eq. (15). If $\mathbf{Z}$ were not maximal we could construct a better solution by using other variables in the Markov blanket of Y, and if $\mathbf{Z}$ does not d-separate Y from $\mathbf{S}$ we can construct an SCM compatible with $\mathcal{G}^*$ that would make the squared error loss in π^* infinite. ■

Theorem 7 (Thm. 2 restated)**.** *Let $\mathbf{M} = ch(\mathbf{S})_{\mathcal{G}^i}$ and $\mathbf{Z} \subset V$ be the maximal subset of the Markov blanket of Y in $\mathcal{G}^i$ where π^i is a training domain chosen such that $(Y \perp\!\!\!\perp \mathbf{S} \mid \mathbf{Z} \backslash \mathbf{M})_{\mathcal{G}^{*,i}_{\overline{M}}}$. If identified, the intervention invariant predictor $\mathbb{E}_{P_{\pi^i}}[y \mid \mathbf{z} \backslash \mathbf{m}, do(\mathbf{m})]$ is a solution to the robust optimization problem in Eq. (15) over all SCMs M with graph $\mathcal{G}^*$ compatible with $\mathcal{G}^{*,i}$.*

Proof. The proof of Theorem 2 is given in (Subbaswamy et al., 2019, Thms. 3 and 4) that has been adapted to the notation used in this paper. ■

Proposition 3 (Prop. 1 restated)**.** *Let $\mathbf{Z} \subset V$ be the maximal subset of the Markov blanket of Y such that $(Y \perp\!\!\!\perp \mathbf{S} \mid \mathbf{Z})_{\mathcal{G}^{i,j}}$ for all $i, j = 1, \ldots, k$. For a given graph $\mathcal{G}^*$, if $\min_{M \in \mathcal{M}(\mathcal{G}^*)} \mathbb{E}_{P_M}[y \mid \mathbf{z}] = \max_{M \in \mathcal{M}(\mathcal{G}^*)} \mathbb{E}_{P_M}[y \mid \mathbf{z}]$ then $f(\mathbf{z}) := \min_{M \in \mathcal{M}(\mathcal{G}^*)} \mathbb{E}_{P_M}[y \mid \mathbf{z}]$ is the optimal statistically invariant predictor.*

Proof. Consider the conditions of Thm. 6 on the separation statement required for optimality of the invariant statistical predictor. If such a predictor is well-defined, then by definition of the partial transportability task under the same conditions, the lower and upper bound coincide and the recovered function is the invariant statistical predictor because it is unique. ■

Proposition 4 (Prop. 2 restated)**.** *Let $\mathbf{M} = ch(\mathbf{S})_{\mathcal{G}^i}$ and $\mathbf{Z} \subset V$ be the maximal subset of the Markov blanket of Y in $\mathcal{G}^i$ where π^i is a training domain chosen such that $(Y \perp\!\!\!\perp \mathbf{S} \mid \mathbf{Z} \backslash \mathbf{M})_{\mathcal{G}^{*,i}_{\overline{M}}}$. If $\min_{M \in \mathcal{M}(\mathcal{G}^*)} \mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{z} \backslash \mathbf{m}, do(\mathbf{m})] = \max_{M \in \mathcal{M}(\mathcal{G}^*)} \mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{z} \backslash \mathbf{m}, do(\mathbf{m})]$ then $f(\mathbf{z} \backslash \mathbf{m}, \mathbf{m}) := \min_{M \in \mathcal{M}(\mathcal{G}^*)} \mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{z} \backslash \mathbf{m}, do(\mathbf{m})]$ is the optimal intervention invariant predictor.*

Proof. The proof uses the same argument as that of Prop. 3. ■

Theorem 8 (Thm. 3 restated)**.** *Let $\mathbf{X} = \mathbf{V} \backslash Y$. Given a query $\mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}]$ and an arbitrary set of selection diagrams $\{\mathcal{G}^{*,i} : i = 1, \ldots, k\}$, let $[\hat{l}(\mathbf{x}), \hat{u}(\mathbf{x})]$ be the solution to the partial transportability task in Eq. (3). Then, $f(\mathbf{x}) := \hat{l}(\mathbf{x})$ is the solution to the worst-case loss minimization problem in Eq. (2) over all SCMs M with graph $\mathcal{G}^*$.*

Proof. Consider the following decomposition,

$$\max_{M \in \mathcal{M}(\mathcal{G}^*)} \mathbb{E}_{y|\mathbf{x} \sim P_M}[(y - f(\mathbf{x}))^2] = \max_{M \in \mathcal{M}(\mathcal{G}^*)} \mathbb{E}_{y|\mathbf{x} \sim P_M}[y^2 - 2yf(\mathbf{x}) + f(\mathbf{x})^2] \tag{16}$$

$$= \max_{M \in \mathcal{M}(\mathcal{G}^*)} \mathbb{E}_{y|\mathbf{x} \sim P_M}[y^2] - 2f(\mathbf{x}) \min_{M \in \mathcal{M}(\mathcal{G}^*)} \mathbb{E}_{y|\mathbf{x} \sim P_M}[y] + f(\mathbf{x})^2 \tag{17}$$

The function $\hat{f}(\mathbf{x})$ that minimizes the above is given by $\hat{f}(\mathbf{x}) = \min_{M \in \mathcal{M}(\mathcal{G}^*)} \mathbb{E}_{P_M}[y \mid \mathbf{x}]$ which is the lower bound $l(\mathbf{x})$ solution to the transportability task under the given set of selection diagrams $\{\mathcal{G}^{*,i} : i = 1, \ldots, k\}$. ■

Theorem 9 (Thm. 4 restated)**.** *Let M be an arbitrary SCM with graph $\mathcal{G}$ and let $\mathbf{C}$ be any c-component. Then, there exists a corresponding SCM N with finite exogeneous domain compatible with $\mathcal{G}$ such that $Q_M[\mathbf{C}] = Q_N[\mathbf{C}]$, where for every exogenous variable $U \in \mathbf{U}_{\mathbf{C}}$, its cardinality $d_U = \left|\Omega_{Pa(\mathbf{C})}\right|$.*

Proof. The proof follows from Rosset et al. (2017) and Zhang et al. (2021). We include it below for completeness.

We first introduce some necessary notations and concepts. The probability distribution for every exogenous variables $U \subset \mathbf{U}$ is characterized with a probability space. It is frequently designated $\langle \Omega_U, \mathcal{F}_U, P_U \rangle$ where Ω_U is a sample space containing all possible outcomes; $\mathcal{F}_U$ is a σ-algebra containing subsets of Ω_U; P_U is a probability measure on $\mathcal{F}_U$ normalized such that $P_U(\Omega_U) = 1$. Elements of $\mathcal{F}_U$ are called events, which are closed under operations of set complement and unions of countably many sets. By means of P_U a real number $P_U(\mathcal{A}) \in [0, 1]$ is assigned to every event $\mathcal{A} \in \mathcal{F}_U$; it is called the probability of event $\mathcal{A}$. For an arbitrary set of exogenous variables $\mathbf{U}$, its realization $\mathbf{U} = \mathbf{u}$ is an element in the Cartesian product $\times_{U \in \mathbf{U}} \Omega_U$. We may be interested in inferring whether a sequence of events $\mathcal{A}$ for every $U \in \mathbf{U}$ occurs. Such an event is represented by a subset $\times_{U \in \mathbf{U}} \mathcal{A}_U \subseteq \times_{U \in \mathbf{U}} \Omega_U$ which in turn generate a product of σ-algebras $\bigotimes_{U \in \mathbf{U}} \mathcal{F}_U$. Define the product measure $\bigotimes_{U \in \mathbf{U}} P_U$ to satisfy the following mutual independence condition given by the definition of the SCM,

$$P\left(\bigtimes_{U \in \mathbf{U}} \mathcal{A}_U\right) = \prod_{U \in \mathbf{U}} P_U(\mathcal{A}_U). \tag{18}$$

Such P is a probability measure. Moreover,

$$\left\langle \bigtimes_{U \in \mathbf{U})} \Omega_U, \bigotimes_{U \in \mathbf{U}} \mathcal{F}_U, \bigotimes_{U \in \mathbf{U}} P_U \right\rangle, \tag{19}$$

defines a product of probability spaces $\langle \Omega_U, \mathcal{F}_U, P_U \rangle$ that describes measurable events over all exogeneous variables $\mathbf{U}$ partitioned into c-components.

Let $\mathbb{C}$ be the collection of all c-components in $\mathcal{G}$. c-components in $\mathbb{C}$ form a partition $\{\bigcup_{V \in \mathbf{C}} U_V \mid \mathbf{C} \in \mathbb{C}\}$ over exogenous variables U. Therefore, for every $U \in \mathbf{U}$, there must exist a unique c-component denoted by $\mathbf{C}_U$ containing U. For any c-component $\mathbf{C} \in \mathbb{C}$, let $U_{\mathbf{C}} = \bigcup_{V \in \mathbf{C}} U_V$ the set of exogenous variables affecting (at least one of) endogenous variables in $\mathbf{C}$. By the definition of c-components, the exogeneous variables do not overlap between c-components and it holds that,

$$P\left(\bigcap_{U \in \mathbf{U}} \mathcal{A}_U\right) = \prod_{C \in \mathcal{C}(\mathcal{G})} P_U\left(\bigcap_{U \in C} \mathcal{A}_U\right). \tag{20}$$

For any SCM M compatible with the causal graph $\mathcal{G}$ the joint distribution may be factorized into c-components,

$$P(\mathbf{v}) = \prod_{\mathbf{C} \in \mathbb{C}} Q[\mathbf{C}](\mathbf{c}, \boldsymbol{pa}_{\mathbf{C}}). \tag{21}$$

where $Q[\mathbf{C}]$ is a C-factor and is a function of $(\mathbf{c}, \boldsymbol{pa}_{\mathbf{C}})$.

To parameterize this joint distribution it is thus sufficient to look at each C-factor separately. Let $\mathbf{C}$ be a generic c-component in $\mathcal{G}$. Denote by $m = |U_{\mathbf{C}}|$ the number of exogeneous variables related to $\mathbf{C}$. For convenience, we consistently write $\langle \Omega_i, \mathcal{F}_i, P_i \rangle$ as the probability space of i-th exogeneous variable in C. The product of these probability spaces is thus written,

$$\left\langle \bigtimes_{i=1}^{m} \Omega_i, \bigotimes_{i=1}^{m} \mathcal{F}_i, \bigotimes_{i=1}^{m} P_i \right\rangle. \tag{22}$$

Each C-factor may thus be written,

$$Q[\mathbf{C}] = \int_{\times_{i=1}^{m} \Omega_i} \prod_{V \in \mathbf{C}} \mathbb{1}\{f_V(\boldsymbol{pa}_V, \boldsymbol{u}_V) = v\} d \bigotimes_{i=1}^{m} P_i. \tag{23}$$

Our goal is to show that all probabilities $Q[\mathbf{C}]$, induced by exogenous variables described by arbitrary probability spaces could be produced by a "simpler" generative process with discrete exogenous domains. $Q[\mathbf{C}]$ defines a mapping between the space of possible realizations of the variables $Pa(\mathbf{C})$ to the $[0, 1]$ interval. Since $Pa(\mathbf{C})$ are discrete variables with finite domains, the cardinality of the class of probability assignments that must be defined is also finite. It is given at most by the number of possible combinations of realizations of $Pa(\mathbf{C})$ which is given by $\prod_{V \in Pa(\mathbf{C})} |\Omega_V|$.

Let $\bar{P}$ be a vector representing probabilities $Q[\mathbf{C}](\mathbf{c}, \boldsymbol{pa}_{\mathbf{C}})$. Counting all possible combinations of outcomes for all possible conditioning sets, $\bar{P}$ is therefore a vector of at most size $d = \prod_{V \in Pa(\mathbf{C})} |\Omega_V|$. And since $Q[\mathbf{C}](\mathbf{c}, \boldsymbol{pa}_{\mathbf{C}})$ is a probability mass function, it only takes a vector with $d - 1$ dimensions to uniquely determine it. $\bar{P}$ may thus be interpreted as a point in the $(d-1)$-dimensional real space. Similarly, $(P, 1)$ is vector in d-dimensional space where the d-th element is equal to 1.

Now consider sampling a value $U_1 = u_1$ from the underlying SCM and let Q_{u_1} be the probability model with $U_1 = u_1$.

$$Q_{u_1}[\mathbf{C}](\mathbf{c} \mid \boldsymbol{pa}_{\mathbf{C}}) = \left[\int_{\times_{i=2}^{m} \Omega_i} \prod_{V \in \mathbf{C}} \mathbb{1}\{f_V(\boldsymbol{pa}_V, \boldsymbol{u}_V) = v\} d \bigotimes_{i=2}^{m} P_i \right]_{U_1 = u_1}. \tag{24}$$

and $\bar{P}_{u_1}$ is a $(d-1)$-dimensional probability vector representing the probabilities of each one of the combinations $Pa(\mathbf{C})$ given that $U_1 = u_1$. We will show that P_1 may equally well be represented by a discrete distribution. For this, let $\mathcal{U} = \{\bar{P}_{u_1} : u_1 \in \Omega_1\} \subset \mathbb{R}^d$ be the set of probability points that can be constructed as u_1 varies in Ω_1. The average $\int_{\Omega_1} \bar{P}_{u_1} dP_1$ is a convex mixture of points in $\mathcal{U}$ by (Rubin & Wesler, 1958) that equals $\bar{Q}$ since,

$$\bar{P} = \int_{\Omega_1} \left[\int_{\times_{i=2}^{m} \Omega_i} \prod_{V \in \mathbf{C}} \mathbb{1}\{f_V(\boldsymbol{pa}_V, \boldsymbol{u}_V) = v\} d \bigotimes_{i=2}^{m} P_i \right]_{U_1 = u_1} dP_1. \tag{25}$$

By construction, $\bar{P}$ itself is a convex mixture of at most $d + 1$ points in $\mathcal{U}$. That is, by using Carathéodory's theorem (Carathéodory, 1911),

$$\bar{P} = \sum_{k=1}^{d+1} w_k \bar{P}_{u_{1,k}}. \tag{26}$$

Replacing the definition of $\bar{P}_{u_{1,k}}$ we obtain,

$$\bar{P} = \sum_{k=1}^{d+1} w_k \left[\int_{\times_{i=2}^{m} \Omega_i} \prod_{V \in \mathbf{C}} \mathbb{1}\{f_V(\boldsymbol{pa}_V, \boldsymbol{u}_V) = v\} d \bigotimes_{i=2}^{m} P_i \right]_{U_1 = u_{1,k}} \tag{27}$$

This means that we can replace the continuous measure P_1 with a discrete probability set with outcomes $\{u_{1,1}, \ldots, u_{1,d}\}$ and corresponding probabilities $\{w_1, \ldots, w_d\}$ with cardinality d and obtain a probability model that is equivalent to the original $\bar{P}$. This procedure can be repeated for all m exogeneous variables in the c-component $\mathbf{C}$. We are thus left with a model,

$$Q[\mathbf{C}](\mathbf{c}, \boldsymbol{pa}_{\mathbf{C}}) = \int_{\times_{i=1}^{m} \Omega_i} \prod_{V \in \mathbf{C}} \mathbb{1}\{f_V(\boldsymbol{pa}_V, \boldsymbol{u}_V) = v\} d \bigotimes_{i=1}^{m} P_i, \tag{28}$$

equivalent to its discrete counterpart,

$$Q[\mathbf{C}](\mathbf{c}, \boldsymbol{pa}_{\mathbf{C}}) = \sum_{u \in \mathbf{U}_{\mathbf{C}}} \sum_{u=1,\ldots,d} \prod_{V \in \mathbf{C}} \mathbb{1}\{f_V(\boldsymbol{pa}_V, \boldsymbol{u}_V) = v\} \prod_{u \in \boldsymbol{u}_{\mathbf{c}}} P(u), \tag{29}$$

where $d = \prod_{V \in Pa(\mathbf{C})} |\Omega_V|$. ■

Theorem 10 (Thm. 5 restated)**.** *The solution $[l, u]$ to the partial transportability task defined over discrete SCMs is a tight bound over a target query $\mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}]$. The credible interval $[\hat{l}_0, \hat{u}_0]$ coincides with $[l, u]$ as $n_i \to \infty$ in all observable domains π^i, $i = 1, \ldots, k$.*

Proof. The proof strategy follows (Zhang et al., 2021) and shows convergence of the posterior by way of convergence of the likelihood of the data given one SCM $M \in \mathcal{M}(\mathcal{G})$. We look at 'convergence' in a frequentist way, for increasing sample size the posterior will, with increasing probability, be low for any parameter configuration, i.e. for any SCM $M \notin \mathcal{M}(\mathcal{G})$.

By the definition of the optimal bounds $[l, u]$ given by the solution to the partial transportability task,

$$P(\bar{\mathbf{v}} \mid l < \mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}] < u) \rightarrow_p 1. \tag{30}$$

Therefore if the prior on parameters $(\boldsymbol{\xi}, \boldsymbol{\theta})$ defining SCMs is non-zero for any $M \in \mathcal{M}(\mathcal{G})$, also the posterior converges,

$$P(l < \mathbb{E}_{P_{\pi^*}}[y \mid \mathbf{x}] < u \mid \bar{\mathbf{v}}) \rightarrow_p 1, \tag{31}$$

which is the definition credible intervals $[l_0, u_0]$ as the 0th and 100th quantiles of the posterior distribution which coincide with $[l, u]$ asymptotically. ■

B EXPERIMENTAL DETAILS

Using the functional dependencies specified by the selection diagram in Fig. 3, we define the SCMs for domains π^{UK}, π^{FR}, and π^{SW} as follows.

For π^{UK} we generate samples from $P(u_w), P(u_s), P(u_t)$ and $P(u_{sc})$ given by independent Gaussian distributions with mean 0 and variance 1. Each generated (u_w, u_s, u_t, u_{sc}) leads to a sample (w, s, t, c) as follows: $w \leftarrow \mathbb{1}\{u_w > 0\}, s \leftarrow \mathbb{1}\{w + u_{sc} + u_s - 2 > 0\}, t \leftarrow \mathbb{1}\{s - 0.5u_t - 1 > 0\}, c \leftarrow \mathbb{1}\{t - 0.5w + u_{sc} - 1 > 0\}$. For π^{FR} we generate samples from $P(u_w), P(u_s), P(u_t)$ and $P(u_{sc})$ given by independent Gaussian distributions with mean 0 and variance 1. Each generated (u_w, u_s, u_t, u_{sc}) leads to a sample (w, s, t, c) as follows: $w \leftarrow \mathbb{1}\{u_w > 0\}, s \leftarrow \mathbb{1}\{w + u_{sc} + 1.5u_s - 1 > 0\}, t \leftarrow \mathbb{1}\{s - 0.5u_t - 1 > 0\}, c \leftarrow \mathbb{1}\{t - 0.5w + u_{sc} - 1 > 0\}$. Notice that the causal mechanism for S has changed while everything else is unchanged. For π^{SW} we generate samples from $P(u_w), P(u_s), P(u_t)$ and $P(u_{sc})$ given by independent Gaussian distributions with mean 0 and variance 1. Each generated (u_w, u_s, u_t, u_{sc}) leads to a sample (w, s, t, c) as follows: $w \leftarrow \mathbb{1}\{u_w > 0.5\}, s \leftarrow \mathbb{1}\{w + u_{sc} + u_s - 2 > 0\}, t \leftarrow \mathbb{1}\{s - 0.5u_t - 1 > 0\}, c \leftarrow \mathbb{1}\{t - 0.5w + u_{sc} - 1 > 0\}$. $\wedge$ is the xor operator: $x \wedge y$ equal to one if $x = y$ and zero otherwise.

All experiments use 1000 burn-in MCMC samples that are discarded and 5000 MCMC samples considered as independent samples from the posterior distribution and used for the approximation of target queries.