On the Interventional Kullback-Leibler Divergence

Jonas Wildberger WILDBERGER.JONAS@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems, Max-Planck-Ring 4, 72076 Tübingen, Germany

Siyuan Guo SYG26@CANTAB.AC.UK
Max Planck Institute for Intelligent Systems, Max-Planck-Ring 4, 72076 Tübingen, Germany
University of Cambridge, Cambridge

Arnab Bhattacharyya ARNABB@NUS.EDU.SG
School of Computing, National University of Singapore, Singapore

Bernhard Schölkopf BS@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems, Max-Planck-Ring 4, 72076 Tübingen, Germany

Editors: Mihaela van der Shaar, Dominik Janzing and Cheng Zhang

Abstract

Modern machine learning approaches excel in static settings where a large amount of i.i.d. training data are available for a given task. In a dynamic environment, though, an intelligent agent needs to be able to transfer knowledge and re-use learned components across domains. It has been argued that this may be possible through causal models, aiming to mirror the modularity of the real world in terms of independent causal mechanisms. However, the true causal structure underlying a given set of data is generally not identifiable, so it is desirable to have means to quantify differences between models (e.g., between the ground truth and an estimate), on both the observational and interventional level.

In the present work, we introduce the Interventional Kullback-Leibler (IKL) divergence to quantify both structural and distributional differences between models based on a finite set of multi-environment distributions generated by interventions from the ground truth. Since we generally cannot quantify all differences between causal models for every finite set of interventional distributions, we propose a sufficient condition on the intervention targets to identify subsets of observed variables on which the models provably agree or disagree.

Keywords: causal distance, causal discovery, multi-environment learning

1. Introduction

Classical machine learning methods are concerned with learning a distribution P (or properties thereof) from a large i.i.d. sample using a simpler model Q. Recent advances in causality have shown how access to heterogeneous data from multiple environments can help uncover causal relationships within the data and address problems like adversarial robustness or domain generalization (Peters et al., 2017; Zhang et al., 2013; von Kügelgen et al., 2021; Schölkopf et al., 2021; Eastwood et al., 2022; Guo et al., 2022). In this framework, distributional shifts emerge from interventions in some underlying causal model $\mathscr{P}$. Identifying the true causal model, however, is difficult even in the multi-environment regime, requiring many environments and (often unverifiable) assumptions on the type of interventions present. One thus often needs to be satisfied with an approximate model $\mathscr{Q}$, which is *close* to a hypothetical ground truth $\mathscr{P}$.

In the present work, we propose a quantitative notion of observational and interventional *closeness* between two causal models $\mathscr{P}$ and $\mathscr{Q}$. Extending the classical Kullback-Leibler divergence,

© 2023 J. Wildberger, S. Guo, A. Bhattacharyya & B. Schölkopf.

our *Interventional Kullback-Leibler divergence* $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q})$ is computed between a finite set of interventional distributions with known intervention targets $((P^e, \mathcal{I}_e))_{e \in \mathcal{E}}$ from $\mathscr{P}$, denoted as $\mathscr{P}_{\mathcal{E}}$, and our model estimate $\mathscr{Q}$, consisting of a single reference distribution estimate Q and graphical model estimate G_Q. Intuitively, we can understand D_{IKL} as quantifying the average deviation between distributions under interventions and our model's predictions, had the interventions also happened on our model estimate. We thus quantify how good our model estimate is beyond observational.

Minimizing the IKL divergence over our estimate $\mathscr{Q}$ with respect to arbitrary interventions is equivalent to finding the true causal model $\mathscr{P} = \mathscr{Q}$. In practice, having access to only a limited amount of multi-environment distributions presents interesting challenges regarding which structural differences are identifiable. We propose a sufficient condition on the intervention targets trading-off a priori knowledge about $\mathscr{P}$ to establish the equivalence $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = 0 \iff \mathscr{P} = \mathscr{Q}$. If this condition is not known to be satisfied, we further show which differences with respect to subsets of variables can be provably identified.

Previous approaches (Acid and Campos, 2003; Tsamardinos et al., 2006; Peters and Bühlmann, 2013) define interventional distances via differences between two causal graphs. In contrast, our IKL divergence measure (1) does not require access to the true causal graph G_P and (2) also takes into account distributional differences. Further, unlike Acharya et al. (2018); Peyrard and West (2020), there is no need to experimentally perform (hard-)interventions for the evaluation of our distance; instead we allow for arbitrary unknown mechanism changes, as long as the targets of the interventions are discerned.

The remainder of this work is organized as follows: In Section 2, we introduce the problem settings and the assumptions to set the stage for the following theoretical analysis. Section 3 contains a discussion about related work and our main results, studying the Interventional Kullback-Leibler divergence first for the example of a known graph and then in the general setting. Finally, in Section 4, we discuss the operational significance of the IKL divergence and various applications related to the area of multi-environment causal learning. Summarizing, our main contributions are:

- We introduce the Interventional Kullback-Leibler divergence D_{IKL} that quantifies distributional and structural differences between a given set of multi-environment distributions $(P^e)_{e \in \mathcal{E}}$ from the ground truth $\mathscr{P}$ and our inferred model $\mathscr{Q}$.

- We discuss various properties and applications of the IKL divergence related to tasks in causal inference.

- We prove a theorem establishing equivalence between $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = 0$ and equality between causal models $\mathscr{P} = \mathscr{Q}$ under certain conditions on the environments. We thereby trade off (partial) knowledge about $\mathscr{P}$ and access to interventional distributions $(P^e)_{e \in \mathcal{E}}$. This gives rise to a sufficient condition that can be verified using only the inferred model $\mathscr{Q}$.

2. Problem Setting and Preliminaries

We begin by reviewing the causal formalism used below. This is based on interventional distribution shifts, modelled as causal graphical models (Pearl, 2009).

2.1. Causal Terminology

Definition 1 (Causal Graphical Model) *A Causal Graphical Model (CGM) (P, G_P) over a variable set $\mathbf{X} = \{X_1, \dots, X_d\}$ consists of a distribution P and a directed acyclic graph (DAG) $G_P = (\mathbf{V}, \mathbf{E})$ where $\mathbf{V} := \{1, \dots, d\}$ is an index set and $\mathbf{E} \subseteq \mathbf{V}^2$ is an edge set, such that $(i, j) \in \mathbf{E}$ if and only if X_i directly causes X_j. We say that two CGMs $(P, G_P), (Q, G_Q)$ are the same $(P, G_P) = (Q, G_Q)$ if and only if $P = Q$ and $G_P = G_Q$.*

Definition 2 (Markov Factorization) *Given a joint distribution P and a DAG G, we say P satisfies the Markov factorization property*[1] *with respect to G (or is Markovian with respect to G) if*

$$P(X_1, \dots, X_d) = \prod_i P(X_i \mid \mathbf{PA}_i^G), \tag{1}$$

where $\mathbf{PA}_i^G$ denotes the set of parents of X_i in G.

While Definition 2 allows us to deduce properties of the distribution P from the corresponding graph G_P, the following faithfulness property allows us to perform inference in the converse direction, i.e. finding the structure of the graph given the distribution.

Definition 3 (Causal Faithfulness) *A joint distribution P is called faithful with respect to a DAG G if any conditional independence relationship in P is implied by d-separation in G (Spirtes et al., 1993; Peters et al., 2017).*

Assumption 4 *In the following, we consider the set of distributions that are both Markovian and faithful with respect to some causal DAG G, i.e. given an observed distribution P, there exists a compatible graph G_P.*

A justification for the faithfulness assumption is given by Koller and Friedman (2010, Theorem 3.5) stating that almost all distributions P that are Markovian with respect to G are also faithful to G.

2.2. Multi-environment learning

Throughout this work, we take advantage of data from multiple environments. Specifically, we assume that given a set of multi-environment distributions $(P^e)_{e \in \mathcal{E}}$ over some variable set $\mathbf{X}$, there exists an underlying ground truth CGM $\mathscr{P} := (P, G_P)$ giving rise to $(P^e)_{e \in \mathcal{E}}$. The generative process behind these distributions is assumed to satisfy the below (Schölkopf et al., 2012; Peters et al., 2017):

Assumption 5 (Independent Causal Mechanisms (ICM)) *A causal generative process of a system of variables is composed of autonomous modules that do not inform or influence each other. Mathematically speaking, given a Markov factorization of the joint distribution as in (1), the causal conditionals (also called* mechanisms*) $P(X_i \mid \mathbf{PA}_i^{G_P})$ represent autonomous modules in the sense that*

(i) intervening on one $P(X_i \mid \mathbf{PA}_i^{G_P})$ will not change other mechanisms $P(X_j \mid \mathbf{PA}_j^{G_P}), i \neq j$,

1. We assume that all distributions have densities with respect to the Lebesgue measure.

(ii) knowing one mechanism will not provide information about other mechanisms.

Under Assumption 5, environment shifts act independently on each of the mechanisms $P(X_i \mid \mathbf{PA}_i^{G_P})$. Thus, for each environment, $e \in \mathcal{E}$ there exists a well-defined set of mechanisms that are altered by the environment.

Assumption 6 (Multi-Environment Distributions) *For each environment $e \in \mathcal{E}$, 'Nature' first chooses a subset of variables to intervene upon $\mathcal{I}_e \subseteq [d]$, where $[d] := \{1, \dots, d\}$. For each intervened variable, 'Nature' then chooses a mechanism shift $\tilde{P}(X_i \mid \mathbf{PA}_i) \in \mathcal{F}$ to perform, where $\mathcal{F}$ is a family of functions. The transition from P to P^e thus maps $P(X_i \mid \mathbf{PA}_i) \mapsto \tilde{P}(X_i \mid \mathbf{PA}_i)$ if $i \in \mathcal{I}_e$ and $P(X_i \mid \mathbf{PA}_i) \mapsto P(X_i \mid \mathbf{PA}_i)$ otherwise, and we may represent the joint distribution in environment e as:*

$$P^e(X_1, \dots, X_d) = \prod_{i \in \mathcal{I}^e} \tilde{P}(X_i \mid \mathbf{PA}_i) \cdot \prod_{i \in [d] \setminus \mathcal{I}^e} P(X_i \mid \mathbf{PA}_i) \tag{2}$$

Such mechanism shifts arise when the context of the observed distribution changes. This may happen for example as the result of different outer circumstances like climate or time (Mooij et al., 2016) or by explicitly intervening on an experiment. Such shifts are generally called *interventions*. Special cases include *soft interventions* where the structure of the graph is left invariant, i.e., X_i does not become conditionally independent of any of its parents, and *hard interventions* where $\tilde{P}(X_i \mid \mathbf{PA}_i) = \delta(X_i - x)$. In any case, P^e will still be Markovian with respect to the ground truth graph G_P. However, it may no longer be faithful to it, unless the interventions are soft.

Our final assumption concerns the existence of hidden confounder variables. We here relax the causal sufficiency assumption (stating that there are no unobserved confounders) to pseudo causal sufficiency. This ensures in particular that there are no changes in observed distributions that are not explained by the environment and the causal graph G_P.

Assumption 7 (Pseudo causal sufficiency) *The effect of any unobserved confounder can be completely explained by the environment variable, i.e. in any given environment $e \in \mathcal{E}$ potential unobserved confounders are fixed, and causal sufficiency is satisfied (Huang et al., 2019).*

3. The Interventional KL divergence

In this section, we introduce our proposed Interventional KL divergence. We begin by reviewing the problem of defining a distance between two CGMs $\mathscr{P} = (P, G_P)$ and $\mathscr{Q} = (Q, G_Q)$ and discussing related work. We then introduce the IKL divergence first for a known true causal graph G_P. This definition is subsequently generalized to an unknown G_P. Finally, we establish the equivalence $\mathscr{P} = \mathscr{Q} \iff D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = 0$, and derive a corollary in a partial information setting. Thereby, $\mathscr{Q}$ takes the role of an 'estimate' for the unknown $\mathscr{P}$, i.e. we know G_Q and have access to and can evaluate the density of Q. For complete proofs of the results in this section, we refer to Appendix A.

3.1. Related Work

Defining a general distance between two causal graphical models $(P, G_P), (Q, G_Q)$ requires comparing them both with respect to distributional differences between P, Q and structural differences between their underlying causal graphs G_P, G_Q. Early work addressing this problem considered

only structural differences. The Structural Hamming Distance (SHD) (Acid and Campos, 2003; Tsamardinos et al., 2006) counts the number of different edges between two graphs, which has been found to provide limited insights about the effect on interventional distributions (Peters and Bühlmann, 2013). They instead proposed the Structural Intervention Distance (SID) to address this limitation by counting instead the number of different interventional distributions induced by two causal graphs. If the true causal graph G_P is unknown, it is generally unclear how to define a valid distance measure between the causal models. One avenue to solve this problem is leveraging multi-environment distributions that under Assumptions 6 and 7 can provide insights about the structural properties of the true causal graph. Peyrard and West (2020); Acharya et al. (2018) propose a score for comparing general causal graphical models based on their distributional distance after performing hard interventions on both of them. In section 3.3, we show that under these assumptions, the Interventional KL divergence entails their distance as a special case.

Another related area of research is that of multi-environment causal discovery, which seeks to uncover the true causal graph G_P by relaxing the i.i.d. assumption to exchangeable data (Guo et al., 2022) or interventional data (Eaton and Murphy, 2007; Ghassami et al., 2018; He and Geng, 2016). Since these methods usually only provide probabilistic guarantees about recovering the true causal graph, a causal distance metric can be seen as an evaluation tool to track their progress.

Finally, there is related work discussing the identifiability of causal effects and interventional Markov equivalence classes (Tian and Pearl, 2002; Hauser and Bühlmann, 2011, 2015). Our work is tangent to this line of research in that we also seek to understand which kind of interventions are necessary in a general multi-environment setting to distinguish causal models and define a valid distance measure between them.

3.2. The IKL divergence for a known causal graph

Before defining the Interventional KL divergence for general causal models $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ in the next section, we build intuition by first discussing the case of a known ground truth causal graph G_P of $\mathscr{P}$. This means that $G_Q = G_P$, and Q is Markovian with respect to G_P. Furthermore, any differences between $\mathscr{P}, \mathscr{Q}$ are given by distributional differences between P, Q.

Throughout this work, we use the Kullback-Leibler divergence for quantifying distributional differences between P and Q, assuming that P is absolutely continuous with respect to Q:

$$D_{\mathrm{KL}}(P(\mathbf{X}) \| Q(\mathbf{X})) = \int_{\mathcal{X}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}, \tag{3}$$

It is known that $D_{\mathrm{KL}}(P(\mathbf{X}) \| Q(\mathbf{X})) = 0$ if and only if $P = Q$. Now suppose that P^e arises from P via a mechanism shift on the variables $\mathbf{X}_{\mathcal{I}_e}$ for a subset $\mathcal{I}_e \subseteq [d]$, i.e.

$$P^e(X_i | \mathbf{PA}_i^{G_P}) = P(X_i | \mathbf{PA}_i^{G_P}) \quad \text{if and only if} \quad i \in [d] \backslash \mathcal{I}_e. \tag{4}$$

From the chain rule for relative entropy (see e.g. (Cover and Thomas, 2006, Thm 2.5.3), (Budhathoki et al., 2021)), we can deduce the following decomposition to understand how the effect of the mechanism shifts shows up in $D_{\mathrm{KL}}(P^e(\mathbf{X}) \| Q(\mathbf{X}))$. From here on, we denote $\mathbb{E}_{x \sim P(x)}[X]$ as $\mathbb{E}_{P(x)}[X]$.

Lemma 8 *Given causal model $\mathscr{P} = (P, G_P)$ assume that P^e emerges from P via an environment shift according to Assumption 6. If Q is Markovian with respect to G_P, we have the following*

decomposition

$$D_{\mathrm{KL}}(P^e(\mathbf{X})\|Q(\mathbf{X})) = \sum_{i\in\mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_P})}\left[D_{\mathrm{KL}}\left(P^e(X_i \mid \mathbf{pa}_i^{G_P})\|Q(X_i \mid \mathbf{pa}_i^{G_P})\right)\right] + \underbrace{\sum_{i\in[d]\setminus\mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_P})}\left[D_{\mathrm{KL}}\left(P(X_i \mid \mathbf{pa}_i^{G_P})\|Q(X_i \mid \mathbf{pa}_i^{G_P})\right)\right]}_{=:D_{\mathrm{IKL}}(\mathscr{P}_{\{e\}}\|\mathscr{Q})} \tag{5}$$

As a special case we have for $\mathcal{I}_e = \emptyset$:

$$D_{\mathrm{KL}}(P(\mathbf{X})\|Q(\mathbf{X})) = \sum_{i\in[d]} \mathbb{E}_{P(\mathbf{pa}_i^{G_P})}\left[D_{\mathrm{KL}}\left(P(X_i \mid \mathbf{pa}_i^{G_P})\|Q(X_i \mid \mathbf{pa}_i^{G_P})\right)\right] \tag{6}$$

The first sum in equation (5) goes over the intervened mechanisms and therefore fully characterizes the changes that are directly due to environment shifts. Note that it vanishes if we could perfectly replicate the environment shift on our model estimate Q. The second sum, which is a special case of the Interventional KL divergence to be defined more generally below, goes over the non-intervened variables, quantifying the distributions' distance resulting downstream from intervened variables due to the fact that the expectations over parents are taken with respect to intervened distributions.

Now assume that P is unobserved, but we still want to understand whether Q is a valid model. Since, P, Q are Markovian with respect to the same causal DAG G_P, we can express agreement between P and Q equivalently in terms of vanishing Interventional KL divergence $D_{\mathrm{IKL}}(\mathscr{P}_{\{e\}}\|\mathscr{Q})$. First suppose that $P = Q$. The first sum in equation (5) can become arbitrarily large, inflating the KL divergence $D_{\mathrm{KL}}(P^e(\mathbf{X})\|Q(\mathbf{X}))$. However, the second sum in equation (5) vanishes, i.e. $D_{\mathrm{IKL}}(\mathscr{P}_{\{e\}}\|\mathscr{Q}) = 0$, since agreement between the conditional distributions is irrespective of the parent distributions taken in the expectation.[2] The IKL divergence therefore discounts the additional difference between P^e and Q introduced by the environment shift.

Conversely, assume that $D_{\mathrm{IKL}}(\mathscr{P}_{\{e\}}\|\mathscr{Q}) = 0$. It is clear that this does not generally imply that $P = Q$, since we only sum over a subset of the local divergences composing the difference between the joint distribution P and Q, as shown in equation (6). But given multiple interventional distributions $(P^e)_{e\in\mathcal{E}}$, such that each mechanism constituting $P(\mathbf{X})$ remains unchanged in (at least) one environment, we can obtain a valid distance measure between P and Q via averaging. These observations are formalized in the following definition and lemma.

Definition 9 (Interventional KL divergence for shared causal graphs) *Let $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_P)$ be two causal models sharing the same causal graph. Further, assume that we have access to a set of interventional distributions $\mathscr{P}_{\mathcal{E}} = ((P^e, \mathcal{I}_e))_{e\in\mathcal{E}}$ generated from $\mathscr{P}$. We define the Interventional KL divergence as*

$$\begin{aligned} D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}}\|\mathscr{Q}) &= \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\sum_{i\in[d]\setminus\mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_P})}\left[D_{\mathrm{KL}}\left(P^e(X_i \mid \mathbf{pa}_i^{G_P})\|Q(X_i \mid \mathbf{pa}_i^{G_P})\right)\right] \\ &\overset{(4)}{=} \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\sum_{i\in[d]\setminus\mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_P})}\left[D_{\mathrm{KL}}\left(P(X_i \mid \mathbf{pa}_i^{G_P})\|Q(X_i \mid \mathbf{pa}_i^{G_P})\right)\right] \end{aligned} \tag{7}$$

2. Specifically, they need to coincide on the union of support sets $\mathrm{supp}(P(\mathbf{PA}_i^{G_P})) \cup \mathrm{supp}(P^e(\mathbf{PA}_i^{G_P}))$

If $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = 0$, we say that our model $\mathscr{Q}$ is interventionally equivalent with $\mathscr{P}$ with respect to $\mathcal{E}$, denoted as $\mathscr{P} \sim_{\mathcal{E}} \mathscr{Q}$.

Lemma 10 *Under Assumptions 6, 7 and that no variable is always intervened upon, i.e. $\bigcap_{e \in \mathcal{E}} \mathcal{I}_e = \emptyset$, we can conclude that $\mathscr{P}, \mathscr{Q}$ are interventionally equivalent, $\mathscr{P} \sim_{\mathcal{E}} \mathscr{Q}$, if and only if $\mathscr{P} = \mathscr{Q}$.*

Note that if the reference distribution $P(\mathbf{X})$ is included in our set of interventional distributions, i.e. there exists $e \in \mathcal{E}$ such that $\mathcal{I}_e = \emptyset$, the KL divergence $D_{\mathrm{KL}}(P(\mathbf{X}) \| Q(\mathbf{X}))$ makes up one of the terms in the IKL divergence. Thus, the equivalence $\mathscr{P} \sim_{\mathcal{E}} \mathscr{Q} \iff \mathscr{P} = \mathscr{Q}$ is trivially satisfied. We now illustrate an application of the case where we do not have access to the reference distribution P of the CGM (P, G_P).

Example 11 (Partial Observability) *In this example, we illustrate how the ideas presented above formalize the common conception of recovering the joint distribution $P(X_1, X_2, X_3)$ without observing all variables at once — given that we know it factorizes in a non-trivial way. So, let the causal graph G_P be given by the DAG in Figure 1. By the Markov factorization (1), we know that $P(\mathbf{X}) = P(X_1) \cdot P(X_2 \mid X_1) \cdot P(X_3 \mid X_1)$.*

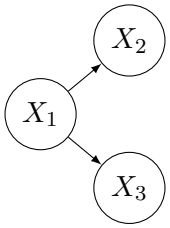

Figure 1: A sample DAG over variables X_1, X_2, X_3.

Now assume that we do not have access to this joint distribution, but instead we are only provided with two marginal distributions $P(X_1, X_2), P(X_1, X_3)$ and the knowledge of the underlying DAG. Even with no interventions, observing the two marginal distributions is sufficient to recover the joint distribution $P(X_1, X_2, X_3)$, since it is possible to separately estimate $P(X_2 \mid X_1)$, $P(X_3 \mid X_1)$ from the marginals and $P(X_1)$ from either of the marginals. This phenomenon is correctly captured in our IKL metric by modelling unobserved variables as interventions, say $P^{e_1}(\mathbf{X}) = P(X_1, X_2)\tilde{P}(X_3)$, $P^{e_2}(\mathbf{X}) = P(X_1, X_3)\tilde{P}(X_2)$:

$$D_{\mathrm{KL}}(P(X_1, X_2) \mid Q(X_1, X_2)) + D_{\mathrm{KL}}(P(X_1, X_3) \mid Q(X_1, X_3)) \tag{8}$$

$$\begin{aligned} &\overset{(6)}{=} D_{\mathrm{KL}}(P(X_1) \| Q(X_1)) + \mathbb{E}_{P(X_1)}[D_{\mathrm{KL}}(P(X_2 \mid X_1) \| Q(X_2 \mid X_1)) \\ &\quad + D_{\mathrm{KL}}(P(X_1) \| Q(X_1)) + \mathbb{E}_{P(X_1)}\left[D_{\mathrm{KL}}(P(X_3 \mid X_1) \| Q(X_3 \mid X_1))\right] \end{aligned} \tag{9}$$

$$\overset{(7)}{=} 2 \cdot D_{\mathrm{IKL}}(\mathscr{P}_{\{e_1, e_2\}} \| \mathscr{Q}) \tag{10}$$

We can now minimize each of the terms in equation (9) with respect to our model distribution Q. By Lemma 10, this will also minimize the distance between the joint distributions P, Q without observing all variables at once. Note that the term $D_{\mathrm{KL}}(P(X_1) \| Q(X_1))$ occurs twice in equation (9). This means that we could still identify the joint distribution in the same way, if an intervention had happened on X_1 in either $P(X_1, X_2)$ or $P(X_1, X_3)$.

3.3. The IKL divergence in the general case

In the previous section, we discussed how the Interventional KL divergence quantifies differences between two causal models $\mathscr{P} = (P, G_P)$ and $\mathscr{Q} = (Q, G_P)$ when G_P is known. Since this is generally not the case, we now generalize Definition 9 to an unknown G_P. To this end, we need to adapt Lemma 8 determining the decomposition of the KL divergence, since the conditioning sets are also unknown. As it turns out, this decomposition still holds when conditioning on $\mathbf{PA}_i^{G_Q}$ as long as P is Markovian with respect to G_Q. Otherwise, there is a residual term quantifying the difference of P from being Markovian, for which we introduce the notion of a *Markov Projection*.

Definition 12 (Markov Projection) *Let $\mathcal{A}$ be the space of probability distributions over the variable set $\mathbf{X}$. The* Markov projection *(or M-projection) of a probability distribution P onto a DAG G is the mapping $\pi_G : \mathcal{A} \to \mathcal{A}$ defined by:*

$$\pi_G(P) = \underset{Q \text{ Markov w.r.t. } G}{\arg\min} D_{\mathrm{KL}}(P \| Q)$$

The following Lemma 13 explicitly describes the Markov projection onto a DAG G.

Lemma 13 *Given a distribution P and a DAG G:*

$$\pi_G(P)(\mathbf{X}) = \prod_{i=1}^{d} P(X_i \mid \mathbf{PA}_i^G). \tag{11}$$

From the Markov factorization property 2 it follows that $P = \pi_G(P)$ if and only if P is Markovian with respect to G. Otherwise, we have $P \neq \pi_G(P)$, since there are variables X_i, X_j such that $X_i \not\perp\!\!\!\perp X_j \mid \mathbf{PA}_i^G$ with respect to P, but $X_i \perp\!\!\!\perp X_j \mid \mathbf{PA}_i^G$ in $\pi_G(P)$. If we now replace G_P by G_Q in Lemma 8 and allow for general (non-Markovian) distributions P, the deviation from being Markovian enters the equation as an additional term:

Lemma 14 *Given two CGMs $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$, assume that P^e emerges from P via an environment shift according to Assumption 6. We then have the following decomposition*

$$\begin{aligned} D_{\mathrm{KL}}(P^e(\mathbf{X}) \| Q(\mathbf{X})) = &\sum_{i \in \mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_Q})} \left[D_{\mathrm{KL}} \left(P^e(X_i \mid \mathbf{pa}_i^{G_Q}) \| Q(X_i \mid \mathbf{pa}_i^{G_Q}) \right) \right] \\ &+ \sum_{i \in [d] \setminus \mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_Q})} \left[D_{\mathrm{KL}} \left(P^e(X_i \mid \mathbf{pa}_i^{G_Q}) \| Q(X_i \mid \mathbf{pa}_i^{G_Q}) \right) \right] \\ &+ D_{\mathrm{KL}} \left(P^e(\mathbf{X}) \| \pi_{G_Q}(P^e)(\mathbf{X}) \right) \end{aligned} \tag{12}$$

As a special case we have for $\mathcal{I}_e = \emptyset$

$$\begin{aligned} D_{\mathrm{KL}}(P(\mathbf{X}) \| Q(\mathbf{X})) = &\sum_{i \in [d]} \mathbb{E}_{P(\mathbf{pa}_i^{G_Q})} \left[D_{\mathrm{KL}} \left(P(X_i \mid \mathbf{pa}_i^{G_Q}) \| Q(X_i \mid \mathbf{pa}_i^{G_Q}) \right) \right] \\ &+ D_{\mathrm{KL}} \left(P(\mathbf{X}) \| \pi_{G_Q}(P)(\mathbf{X}) \right) \end{aligned} \tag{13}$$

With this generalized decomposition, we adapt Definition 9 of the Interventional KL divergence correspondingly to arbitrary CGMs $\mathscr{P}, \mathscr{Q}$.

Definition 15 (Interventional KL divergence) *Let $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ be two causal models. Assume that we have access to a set of interventional distributions $\mathscr{P}_{\mathcal{E}} = ((P^e, \mathcal{I}_e))_{e \in \mathcal{E}}$ generated from $\mathscr{P}$. We then define the Interventional KL divergence as*

$$D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \Bigg[\sum_{i \in [d] \setminus \mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_Q})} \left[D_{\mathrm{KL}} \left(P^e(X_i \mid \mathbf{pa}_i^{G_Q}) \| Q(X_i \mid \mathbf{pa}_i^{G_Q}) \right) \right] + D_{\mathrm{KL}} \left(P^e(\mathbf{X}) \| \pi_{G_Q}(P^e)(\mathbf{X}) \right) \Bigg] \tag{14}$$

Note that this definition is consistent with our previous notion of interventional distance in the case where $G_P = G_Q$, see Definition 9. If $G_P \neq G_Q$, Lemma 14 implies that we can capture differences between the causal models by comparing the Markov factorizations of P^e and Q with respect to G_Q (first line in equation (14)), if we also account for potential deviations from P satisfying this factorization (second line in equation (14)).

The definition of D_{IKL} went under the premise that the effect of the interventions that we consider is unknown, i.e. only the joint distributions $P^e(\mathbf{X})$ and intervention targets $\mathcal{I}_e$ are observed, but we don't know the mapping $P(X_i \mid \mathbf{PA}_i^{G_P}) \mapsto \tilde{P}(X_i \mid \mathbf{PA}_i^{G_P})$. If we know the mechanism shifts that happened between the distribution $P(\mathbf{X})$ and the interventional distributions $P^e(\mathbf{X})$ and we are able to accurately reproduce them in our model $\mathscr{Q} = (Q, G_Q)$, then the D_{IKL} collapses to a simpler form, namely the average of the KL divergences between the joint distributions $P^e(\mathbf{X})$ and $Q(\mathbf{X})$.

Theorem 16 *[Known interventions] Given causal models $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ and a set of interventional distributions $\mathscr{P}_{\mathcal{E}} = ((P^e, \mathcal{I}_e))_{e \in \mathcal{E}}$, define Q^e for each environment $e \in \mathcal{E}$ as follows:*

$$Q^e(\mathbf{X}) = \prod_{i \notin \mathcal{I}_e} Q(X_i \mid \mathbf{PA}_i^{G_Q}) \cdot \prod_{j \in \mathcal{I}_e} P^e(X_j \mid \mathbf{PA}_j^{G_Q}) \tag{15}$$

where $P^e(X_j \mid \mathbf{PA}_j^{G_Q})$ denotes the mechanisms changed as the result of the environment shift. Then we have that

$$D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} D_{\mathrm{KL}}(P^e(\mathbf{X}) \| Q^e(\mathbf{X})) \tag{16}$$

Proof [Sketch] If we apply the same mechanism shifts to Q^e that appear between P and P^e, we can use the decomposition of the KL divergence in Lemma 14 to see that these terms cancel out. The sum over the remaining terms then equals our definition of the Interventional KL divergence. ■

In particular, this Theorem applies to hard interventions on single variables that were considered by Peyrard and West (2020).

3.4. Interventional differences in the IKL divergence

In the previous sections, we have defined the Interventional KL divergence for two general causal graphical models $\mathscr{P}, \mathscr{Q}$ and related it to the (standard) KL divergence. However, it still remains to show that the IKL divergence defines a valid distance between $\mathscr{P}, \mathscr{Q}$, i.e. that $\mathscr{P} \sim_{\mathcal{E}} \mathscr{Q} \iff$

$D_{\text{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = 0 \iff \mathscr{P} = \mathscr{Q}$ under suitable conditions on $\mathcal{E}$. This equivalence will be subject to this section. In particular, we will focus on differences beyond Markov equivalence, since statistically identifiable differences between $\mathscr{P}, \mathscr{Q}$ are already encoded in the purely observational distance $D_{\text{KL}}(P(\mathbf{X}) \| Q(\mathbf{X}))$. For simplicity, we assume in the following that the empty intervention $\mathcal{I} = \emptyset$ is included in the environment. Further, we make use of our assumptions on the multi-environment distributions (Assumptions 6, 7), which imply for Markov equivalent G_P, G_Q that the following statements are equivalent:

(i) For all e with $i \notin \mathcal{I}_e$: $\mathbb{E}_{P^e(\mathbf{PA}_i^{G_Q})} \left[D_{\text{KL}} \left(P^e(X_i \mid \mathbf{PA}_i^{G_Q}) \| P(X_i \mid \mathbf{PA}_i^{G_Q}) \right) \right] = 0$

(i) $\mathbf{PA}_i^{G_Q} = \mathbf{PA}_i^{G_P}$.

As shown in Corollary 20 and Example 21, this equivalence does not hold given only a limited set of interventional distributions. A comprehensive characterization of differences between $P^e(X_i \mid \mathbf{PA}_i^{G_Q})$ and $P(X_i \mid \mathbf{PA}_i^{G_Q})$ can be given based on d-separation in the augmented DAG $G_{P,\mathbf{X} \cup \{e\}}$, assuming faithfulness (Perry et al., 2022). We adopt the faithfulness assumption in the following but, since G_P is assumed to be generally unknown, we here propose a simpler sufficient condition that trades off knowledge about G_P with access to interventional distributions. It is based on the following definition of (directed) unblocked paths between a variable X_i and intervened variables.

Definition 17 *Let $\mathscr{P} = (P, G_P)$ be a CGM and $e \in \mathcal{E}$ an environment. For some (potentially different) DAG G over the same variables $\mathbf{X}$, we say that there exists a (directed) unblocked path $e \overset{G}{\to} X_i$ in G given some variable set $\mathbf{Z} \subseteq \mathbf{X}$, if there exists $k \in \mathcal{I}_e$ and a directed path $X_k \to \ldots \to X_i$ in G such that no element of $\mathbf{Z}$ intersects with the path. If $k = i$, this condition is trivially satisfied.*

So, in particular, $e \overset{G_P}{\to} X_i$ implies that the variable X_i is d-connected to an intervened variable given conditioning set $\mathbf{Z}$. Assuming faithfulness, we can conclude that the environment has an influence on X_i given $\mathbf{Z}$ and thus $P^e(X_i|\mathbf{Z}) \neq P(X_i|\mathbf{Z})$. If G_Q shares the same skeleton with G_P, this definition yields a sufficient condition for the mechanism $P^e(X_i \mid \mathbf{PA}_i^{G_Q})$ to be different from $P(X_i \mid \mathbf{PA}_i^{G_Q})$.

Lemma 18 *Let $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ be CGMs, such that G_P and G_Q share the same skeleton and $X_i \overset{G_Q}{\to} X_j$, $X_i \overset{G_P}{\leftarrow} X_j$ be a flipped edge between G_P and G_Q. Further, let P^e be an interventional distribution generated from $\mathscr{P}$.*

(i) If there exists an unblocked path $e \overset{G_P}{\to} X_i$ given $\mathbf{PA}_j^{G_Q} \setminus X_i$, then $P^e(X_j \mid \mathbf{PA}_j^{G_Q}) \neq P(X_j \mid \mathbf{PA}_j^{G_Q})$.

(ii) If there exists an unblocked path $e \overset{G_P}{\to} X_j$ given $\mathbf{PA}_i^{G_Q}$, then $P^e(X_i \mid \mathbf{PA}_i^{G_Q}) \neq P(X_i \mid \mathbf{PA}_i^{G_Q})$.

This Lemma yields a sufficient condition for structural differences beyond Markov-equivalence to show up in the IKL-divergence. Even if $P = Q$ coincide observationally, the interventional terms

$\mathbb{E}_{P^e(\mathbf{PA}_k^{G_Q})}\left[D_{\mathrm{KL}}\left(P^e(X_k \mid \mathbf{PA}_k^{G_Q})\|Q(X_k \mid \mathbf{PA}_k^{G_Q})\right)\right] > 0$ for $k \in \{i,j\}$ under the conditions of this Lemma. Since these are conditions on the graphical structure of G_P, they cannot be verified if G_P is unknown. But as it turns out, it is sufficient if such paths exist in G_Q. We formalize these observations in the following Theorem:

Theorem 19 *Let $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ be CGMs and $\mathcal{E}$ a set of environments. Now assume that for all (unoriented) edges $X_i \overset{G_Q}{\to} X_j$ there exists an environment $e \in \mathcal{E}$ such that one of the following conditions holds:*

(i) There exists an unblocked path $e \overset{G_Q}{\to} X_i$ given $\mathbf{PA}_j^{G_Q} \setminus X_i$ and $j \notin \mathcal{I}_e$

(ii) There exists an unblocked path $e \overset{G_Q}{\to} X_j$ given $\mathbf{PA}_i^{G_Q}$ and $i \notin \mathcal{I}_e$

Then, $\mathscr{P}, \mathscr{Q}$ are interventionally equivalent, $\mathscr{P} \sim_{\mathcal{E}} \mathscr{Q}$, if and only if $\mathscr{P} = \mathscr{Q}$.

Proof [sketch] If $\mathscr{P} = \mathscr{Q}$, we can deduce $D_{\mathrm{IKL}}(\mathscr{P}\|\mathscr{Q}) = 0$ from our Assumptions on the environment 6 and pseudo-causal faithfulness 7. Conversely, if $P \neq Q$, $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}}\|\mathscr{Q}) \geq 1/|\mathcal{E}|$ $D_{\mathrm{KL}}(P(\mathbf{X})\|Q(\mathbf{X})) > 0$. So it suffices to show that $P = Q$, but $G_P \neq G_Q$ implies that $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}}\|\mathscr{Q}) > 0$. This case then follows from Lemma 18, since an unblocked path in G_Q gives rise to an unblocked path to a mis-oriented edge in G_P and thus one of the conditions of Lemma 18 is satisfied, leading to a nonzero term in the IKL divergence. ■

The less we know about the structure of G_P a priori, the more interventional distributions we need to identify all structural differences between the causal models. Since interventional data can be expensive to collect, we will now show how to identify partial structural differences between $\mathscr{P}$ and $\mathscr{Q}$, given insufficient multi-environmental information to satisfy the conditions of Theorem 19. In particular, we also allow for environments where only marginal information about a subset of all variables is observed.

Corollary 20 *Let $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ be CGMs with $P = Q$ and $\mathcal{E}$ a set of environments. Assume that we only have partial multi-environmental information in the following ways:*

1) We only observe a subset of all variables in the causal graph $\mathbf{X}_S \subseteq \mathbf{X}$, i.e. we can only distinguish the graphical sub-models $\mathscr{P}\big|_S, \mathscr{Q}\big|_S$, induced by removing the unobserved variables from distributions and causal graphs.

2) We know that the conditions of Theorem 19 are only satisfied for a subset $E_{\mathcal{E}}$ of all edges in $G_Q\big|_S$.

We define the restricted IKL divergence via:

$$D_{\mathrm{IKL}}^{res}(\mathscr{P}_{\mathcal{E}}\|\mathscr{Q}) = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\sum_{i\in\mathbf{X}_{E_{\mathcal{E}}}\cap\mathcal{I}_e^C}\mathbb{E}_{P^e(\mathbf{pa}_i^{G_Q|_S})}\left[D_{\mathrm{KL}}\left(P^e(X_i \mid \mathbf{pa}_i^{G_Q|_S})\|Q(X_i \mid \mathbf{pa}_i^{G_Q|_S})\right)\right] \tag{17}$$

summing only over un-intervened variables X such that $(X,\cdot)$ or $(\cdot,X)$ is contained in $E_{\mathcal{E}}$. Then, $D_{\mathrm{IKL}}^{res}(\mathscr{P}_{\mathcal{E}}\|\mathscr{Q}) = 0$ implies that the graphs G_P, G_Q coinicde on the edge set $E_{\mathcal{E}}$. Conversely, if $D_{\mathrm{IKL}}^{res}(\mathscr{P}_{\mathcal{E}}\|\mathscr{Q}) > 0$, there exists a variable $X_i \in \mathbf{X}_{E_{\mathcal{E}}}$ such that $\mathbf{PA}_i^{G_P} \neq \mathbf{PA}_i^{G_Q|_S}$.

Example 21 *Let $(P, G_P), (Q, G_Q)$ be two CGMs with $P = Q$ and graphs given in Figure 2. We want to compute the interventional KL divergence to evaluate the fit of (Q, G_Q) to (P, G_P) using interventional distributions P^{e_1}, P^{e_2}. Assume that in environment e_1 only variables X_2, X_3, X_5 are*

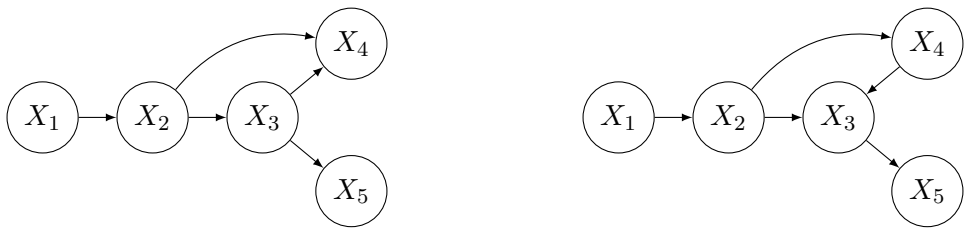

Figure 2: Two Markov equivalent DAGs G_P (left) and G_Q (right) over variables $X_1, \dots, X_5$.

observed with a mechanism shift $P(X_5|X_3) \mapsto \tilde{P}(X_5|X_3)$, i.e. $E_{e_1} = \{(X_3, X_5)\}$ and $\mathbf{X}_{e_1} = \{X_3, X_5\}$ in Corollary 20. We can then compute the restricted IKL divergence using only the observed variables:

$$D^{res}_{\mathrm{IKL}}(\mathscr{P}_{\{e_1\}} \| \mathscr{Q}) = \mathbb{E}_{P^{e_1}(X_2)} D_{\mathrm{KL}}(P^{e_1}(X_3|X_2) \| Q(X_3|X_2)) = 0 \tag{18}$$

indicating that the edge (X_3, X_5) is oriented correctly in G_Q. Note that this equality holds even if the unobserved variable X_4 also happened to be intervened, since it is not conditioned upon in the restricted IKL divergence. Next, assume that in another environment e_2, variables X_2, X_3, X_4 are observed with a mechanism shift affecting $P(X_3|X_2) \mapsto \tilde{P}(X_3|X_2)$. Thus, $E_{e_2} = \{(X_4, X_3), (X_2, X_3)\}, \mathbf{X}_{e_2} = \{X_2, X_3, X_4\}$ and the restricted IKL divergence equals

$$D^{\mathrm{res}}_{\mathrm{IKL}}(\mathscr{P}_{\{e_2\}} \| \mathscr{Q}) = \mathbb{E}_{P^{e_2}(X_2)} D_{\mathrm{KL}}(P^{e_1}(X_4|X_2) \| Q(X_4|X_2)) + D_{\mathrm{KL}}(P^{e_2}(X_2) \| Q(X_2)). \tag{19}$$

The first term is strictly greater than zero due to the mis-oriented edge between X_3 and X_4, i.e. case (ii) in Lemma 18 is satisfied. Despite the correct orientation of the edge $X_2 \to X_3$, the second term may also be greater than zero. This happens precisely if the unobserved variable X_1 also happened to be intervened (case (ii) of Lemma 18 would again be satisfied). Otherwise, this term would vanish, assuring us that the edge is oriented correctly. Thus, in the presence of partial observability, nonzero terms in the interventional KL divergence are not necessarily caused by incorrect (observed) edges. Conversely, however, a vanishing restricted IKL divergence guarantees all identifiable edges $E_{\mathcal{E}}$ to be oriented correctly.

4. Discussion

Multi-environment distributions provide a natural setting for machine learning algorithms to both learn causal structures and leverage them to achieve robustness and out-of-distribution generalization. We have assayed how knowledge about the true causal graph can be used to facilitate learning under distribution shifts and generalize from marginal observations to properties of the joint distribution. Both of these are significant problems for learning in real-world settings, where we cannot always afford to collect a large dataset suited to a task. Conversely, multi-domain data can provide a useful learning signal to drive causal discovery.

The focus of the current paper has been a theoretical exploration of properties of the Interventional KL divergence, and applications are thus beyond our scope. Nevertheless, we would like to discuss some possible use cases to point out directions for future work.

Interventional KL Divergence and Estimation The following result captures the operational significance of low IKL divergence.

Theorem 22 *Given causal models $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ and a set of interventional distributions $\mathscr{P}_{\mathcal{E}} = ((P^e, \mathcal{I}_e))_{e \in \mathcal{E}}$, define Q^e for each environment $e \in \mathcal{E}$ as in Theorem 16.*

Suppose $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) \leq \epsilon$. Then, for any bounded function f mapping to $[-B, B]$ and any parameter $\rho > 0$, for at least $1 - \rho$ fraction of the environments $e \in \mathcal{E}$:

$$\left| \mathbb{E}_{P^e(\mathbf{X})}[f(\mathbf{X})] - \mathbb{E}_{Q^e(\mathbf{X})}[f(\mathbf{X})] \right| \leq \frac{B\sqrt{\epsilon}}{\rho}$$

In other words, if the IKL divergence is small, then for most environments, executing the same mechanism changes in 'Nature' and in our model $\mathscr{Q}$ would yield similar statistics. Thus, the more heterogeneous the set of environments, the more similar $\mathscr{P}$ and $\mathscr{Q}$ are in terms of their causal implications.

Evaluation tool for causal discovery methods If data from multi-environment distributions are available, there is a range of methods to recover the true causal structure asymptotically (Perry et al., 2022; Janzing et al., 2012; Huang et al., 2019; He and Geng, 2016; Rojas-Carulla et al., 2015; Arjovsky et al., 2019; Krueger et al., 2021). Here, the IKL divergence could help develop methods to verify a learned causal model, provided that intervention targets are known for some of the environments. This can then take into account both the interventional differences, as encoded by the learned DAG, and the observational differences encoded in the learned conditional distributions.

Online Learning of Causal Structures If we do not have a sufficient number of environment distributions a priori to recover the true causal graph, but intervention targets are more broadly available, we have shown in Corollary 20 that one could evaluate the fit of a model to the true causal model with respect to subsets of variables. These partially learned and verified models could then be used for inference until more interventional data become available to uncover further causal relationships. Indeed, our metric can even serve for orienting edges in the graph: If we compute $D_{\mathrm{IKL}}(\mathscr{P}_{\{e_1\}} \| (P, G_Q))$, then any non-zero term in the IKL divergence is necessarily caused by mis-oriented edges. Specifically, if for some variable $i \in [d] \backslash \mathcal{I}_e$, we have

$$\mathbb{E}_{P^e(\mathbf{PA}_i^{G_Q})} \left[D_{\mathrm{KL}} \left(P^e(X_i \mid \mathbf{PA}_i^{G_Q}) \| P(X_i \mid \mathbf{PA}_i^{G_Q}) \right) \right] > 0,$$

then $\mathbf{PA}_i^{G_Q} \neq \mathbf{PA}_i^{G_P}$. Thus, we can identify the true parent set by re-computing this term with respect to parent sets $\mathbf{PA}_i^{G}$ for all graphs G in the Markov equivalence class of G_Q. In this way, the IKL divergence is monotonic w.r.t. both distributional differences and structural differences represented by the number of correctly identified edges.

Acknowledgments

We thank Armin Kekić, Julius von Kügelgen, Nasim Rahaman and the Tübingen Causality Team for helpful discussions and comments.

References

Jayadev Acharya, Arnab Bhattacharyya, Constantinos Daskalakis, and Saravanan Kandasamy. Learning and testing causal models with interventions, 2018. URL https://arxiv.org/abs/1805.09697.

S. Acid and L. M. De Campos. Searching for bayesian network structures in the space of restricted acyclic partially directed graphs. *Journal of Artificial Intelligence Research*, 18:445–490, may 2003. doi: 10.1613/jair.1061. URL https://doi.org/10.1613%2Fjair.1061.

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

Kailash Budhathoki, Dominik Janzing, Patrick Bloebaum, and Hoiyi Ng. Why did the distribution change? In *AISTATS 2021*, 2021. URL https://www.amazon.science/publications/why-did-the-distribution-change.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

Cian Eastwood, Alexander Robey, Shashank Singh, Julius von Kügelgen, Hamed Hassani, George J. Pappas, and Bernhard Schölkopf. Probable domain generalization via quantile risk minimization, 2022. URL https://arxiv.org/abs/2207.09944.

Daniel Eaton and Kevin Murphy. Exact bayesian structure learning from uncertain interventions. In Marina Meila and Xiaotong Shen, editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 107–114, San Juan, Puerto Rico, 21–24 Mar 2007. PMLR. URL https://proceedings.mlr.press/v2/eaton07a.html.

AmirEmad Ghassami, Negar Kiyavash, Biwei Huang, and Kun Zhang. Multi-domain causal structure learning in linear systems. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/6ad4174eba19ecb5fed17411a34ff5e6-Paper.pdf.

Siyuan Guo, Viktor Tóth, Bernhard Schölkopf, and Ferenc Huszár. Causal de Finetti: On the Identification of Invariant Causal Structure in Exchangeable Data. *arXiv e-prints*, art. arXiv:2203.15756, March 2022.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. 2011. doi: 10.48550/ARXIV.1104.2808. URL https://arxiv.org/abs/1104.2808.

Alain Hauser and Peter Bühlmann. Jointly interventional and observational data: estimation of interventional markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 77(1):291–318, 2015. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/24774735.

Yango He and Zhi Geng. Causal network learning from multiple interventions of unknown manipulated targets, 2016. URL https://arxiv.org/abs/1610.08611.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data with independent changes. 2019. doi: 10.48550/ARXIV.1903.01672. URL https://arxiv.org/abs/1903.01672.

Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182-183:1–31, 2012. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2012.01.002. URL https://www.sciencedirect.com/science/article/pii/S0004370212000045.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. 2010.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation. In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

Joris M. Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. 2016. doi: 10.48550/ARXIV.1611.10351. URL https://arxiv.org/abs/1611.10351.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

Ronan Perry, Julius von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis, 2022. URL https://arxiv.org/abs/2206.02013.

Jonas Peters and Peter Bühlmann. Structural intervention distance (sid) for evaluating causal graphs. 2013. doi: 10.48550/ARXIV.1306.1043. URL https://arxiv.org/abs/1306.1043.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319.

Maxime Peyrard and Robert West. A ladder of causal distances, 2020. URL https://arxiv.org/abs/2005.02480.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. 2015. doi: 10.48550/ARXIV.1507.05333. URL https://arxiv.org/abs/1507.05333.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning, 2012. URL https://arxiv.org/abs/1206.6471.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning, 2021. URL https://arxiv.org/abs/2102.11107.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*, volume 81. 01 1993. ISBN 978-1-4612-7650-0. doi: 10.1007/978-1-4612-2748-9.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Eighteenth National Conference on Artificial Intelligence*, page 567–573, USA, 2002. American Association for Artificial Intelligence. ISBN 0262511290.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style, 2021. URL https://arxiv.org/abs/2106.04619.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page III–819–III–827. JMLR.org, 2013.

Appendix A. Full proofs

A.1. Proof of Lemmas 8 and 14

Lemma 8 *Given causal model $\mathscr{P} = (P, G_P)$ assume that P^e emerges from P via an environment shift according to Assumption 6. If Q is Markovian with respect to G_P, we have the following decomposition*

$$\begin{aligned} D_{\mathrm{KL}}(P^e(\mathbf{X})\|Q(\mathbf{X})) = &\sum_{i\in\mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_P})}\left[D_{\mathrm{KL}}\left(P^e(X_i \mid \mathbf{pa}_i^{G_P})\|Q(X_i \mid \mathbf{pa}_i^{G_P})\right)\right] \\ &+ \underbrace{\sum_{i\in[d]\setminus\mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_P})}\left[D_{\mathrm{KL}}\left(P(X_i \mid \mathbf{pa}_i^{G_P})\|Q(X_i \mid \mathbf{pa}_i^{G_P})\right)\right]}_{=:D_{\mathrm{IKL}}(\mathscr{P}_{\{e\}}\|\mathscr{Q})} \end{aligned} \tag{5}$$

As a special case we have for $\mathcal{I}_e = \emptyset$:

$$D_{\mathrm{KL}}(P(\mathbf{X})\|Q(\mathbf{X})) = \sum_{i\in[d]} \mathbb{E}_{P(\mathbf{pa}_i^{G_P})}\left[D_{\mathrm{KL}}\left(P(X_i \mid \mathbf{pa}_i^{G_P})\|Q(X_i \mid \mathbf{pa}_i^{G_P})\right)\right] \tag{6}$$

Lemma 14 *Given two CGMs $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$, assume that P^e emerges from P via an environment shift according to Assumption 6. We then have the following decomposition*

$$\begin{aligned} D_{\mathrm{KL}}(P^e(\mathbf{X})\|Q(\mathbf{X})) = &\sum_{i\in\mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_Q})}\left[D_{\mathrm{KL}}\left(P^e(X_i \mid \mathbf{pa}_i^{G_Q})\|Q(X_i \mid \mathbf{pa}_i^{G_Q})\right)\right] \\ &+ \sum_{i\in[d]\setminus\mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_Q})}\left[D_{\mathrm{KL}}\left(P^e(X_i \mid \mathbf{pa}_i^{G_Q})\|Q(X_i \mid \mathbf{pa}_i^{G_Q})\right)\right] \\ &+ D_{\mathrm{KL}}\left(P^e(\mathbf{X})\|\pi_{G_Q}(P^e)(\mathbf{X})\right) \end{aligned} \tag{12}$$

As a special case we have for $\mathcal{I}_e = \emptyset$

$$D_{\mathrm{KL}}(P(\mathbf{X})\|Q(\mathbf{X})) = \sum_{i\in[d]} \mathbb{E}_{P(\mathbf{pa}_i^{G_Q})}\left[D_{\mathrm{KL}}\left(P(X_i \mid \mathbf{pa}_i^{G_Q})\|Q(X_i \mid \mathbf{pa}_i^{G_Q})\right)\right] + D_{\mathrm{KL}}\left(P(\mathbf{X})\|\pi_{G_Q}(P)(\mathbf{X})\right) \tag{13}$$

Proof Both of these Lemmas are immediate consequences of Lemma 23, noting that $\pi_{G_Q}(P)$ satisfies the condition of $\hat{P}$ by Lemma 13. The decomposition for $D_{\mathrm{KL}}(P^e(\mathbf{X})\|Q(\mathbf{X}))$ then just follows by splitting up the sums.

■

A.2. Proof of Theorem 10

Lemma 10 *Under Assumptions 6, 7 and that no variable is always intervened upon, i.e. $\bigcap_{e\in\mathcal{E}} \mathcal{I}_e = \emptyset$, we can conclude that $\mathscr{P}, \mathscr{Q}$ are interventionally equivalent, $\mathscr{P} \sim_{\mathcal{E}} \mathscr{Q}$, if and only if $\mathscr{P} = \mathscr{Q}$.*

Proof Since P, Q are both Markovian with respect to G_P, it follows by Lemma 8 that

$$P = Q \Leftrightarrow D_{\mathrm{KL}}(P(\mathbf{X})\|Q(\mathbf{X})) = 0 \tag{20}$$

$$\overset{(6)}{\Leftrightarrow} \mathbb{E}_{P(\mathbf{pa}_i^{G_P})}\left[D_{\mathrm{KL}}(P(X_i \mid \mathbf{pa}_i^{G_P})\|Q(X_i \mid \mathbf{pa}_i^{G_P}))\right] = 0 \quad \text{for any } i \in [d] \tag{21}$$

$$\Leftrightarrow P(X_i \mid \mathbf{pa}_i^{G_P}) = Q(X_i \mid \mathbf{pa}_i^{G_P}) \quad \text{for any } i \in [d] \tag{22}$$

By our assumption on the multi-environment distributions, we further have that

$$P^e(X_i \mid \mathbf{PA}_i^{G_P}) = P(X_i \mid \mathbf{PA}_i^{G_P}) \quad \text{if and only if} \quad i \in [d]\backslash\mathcal{I}_e \tag{23}$$

Concluding, $P = Q$ implies that $P^e(X_i \mid \mathbf{PA}_i^{G_P}) = Q(X_i \mid \mathbf{PA}_i^{G_P})$ for all $i \in [d]\backslash\mathcal{I}_e$ leading to $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}}\|\mathscr{Q}) = 0$. Conversely, assume that $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}}\|\mathscr{Q}) = 0$. Since $\bigcap_{e\in\mathcal{E}} \mathcal{I}_e = \emptyset$, it follows that $\bigcup_e [d]\backslash\mathcal{I}_e = [d]$, i.e. each term $\mathbb{E}_{P^e(\mathbf{pa}_i^{G_P})}\left[D_{\mathrm{KL}}(P(X_i \mid \mathbf{pa}_i^{G_P})\|Q(X_i \mid \mathbf{pa}_i^{G_P}))\right]$ is included in $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}}\|\mathscr{Q})$ for some environment $e \in \mathcal{E}$. From $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}}\|\mathscr{Q}) = 0$, we can thus deduce that each local divergence vanishes, and $P(X_i \mid \mathbf{pa}_i^{G_P}) = Q(X_i \mid \mathbf{pa}_i^{G_P})$ for any $i \in [d]$ implying that $P = Q$. ■

A.3. Proof of Lemma 13

Lemma 13 *Given a distribution P and a DAG G:*

$$\pi_G(P)(\mathbf{X}) = \prod_{i=1}^{d} P(X_i \mid \mathbf{PA}_i^G). \tag{11}$$

Proof To prove the identity, we make use of the following technical Lemma:

Lemma 23 *For distributions P, Q such that Q is Markovian w.r.t. G, we have the following decomposition:*

$$D_{\mathrm{KL}}(P(\mathbf{X})\|Q(\mathbf{X})) = \sum_{i=1}^{d} \mathbb{E}_{P(\mathbf{pa}_i^G)}\left[D_{\mathrm{KL}}\left(P(X_i \mid \mathbf{pa}_i^G)\|Q(X_i \mid \mathbf{pa}_i^G)\right)\right] + D_{\mathrm{KL}}(P(\mathbf{X})\|\hat{P}(\mathbf{X})) \tag{24}$$

where $\hat{P}$ is a distribution satisfying

$$\hat{P}(\mathbf{X}) = \prod_{i=1}^{d} P(X_i \mid \mathbf{PA}_i^G) \tag{25}$$

Proof [of Lemma 23]

$$\begin{aligned}
D_{\mathrm{KL}}(P(\mathbf{X})\|Q(\mathbf{X})) &= \mathbb{E}_{P(\mathbf{X})}\left[\log \frac{P\hat{P}}{Q\hat{P}}\right] \\
&= \mathbb{E}_{P(\mathbf{X})}\left[\log \frac{\hat{P}}{Q}\right] + \mathbb{E}_{P(\mathbf{X})}\left[\log \frac{P}{\hat{P}}\right] \\
&= \mathbb{E}_{P(\mathbf{X})}\left[\log \frac{\hat{P}}{Q}\right] + D_{\mathrm{KL}}(P(\mathbf{X})\|\hat{P}(\mathbf{X}))
\end{aligned}$$

We now proceed to analyze the remaining term using the definition of $\hat{P}$.

$$\begin{aligned}
\mathbb{E}_{P(\mathbf{x})}\left[\log \frac{\hat{P}}{Q}\right] &= \mathbb{E}_{P(\mathbf{x})}\left[\log \prod_{i=1}^{d} \frac{\hat{P}(X_i \mid \mathbf{PA}_i)}{Q(X_i \mid \mathbf{PA}_i)}\right] \\
&= \sum_{i=1}^{d} \mathbb{E}_{P(\mathbf{x})}\left[\log \frac{P(X_i \mid \mathbf{PA}_i)}{Q(X_i \mid \mathbf{PA}_i)}\right] \\
&= \sum_{i=1}^{d} \sum_{x_i, \mathbf{pa}_i} \log \frac{P(x_i \mid \mathbf{pa}_i)}{Q(x_i \mid \mathbf{pa}_i)} \\
&\quad \cdot \sum_{\mathbf{x}_{[d]} \setminus \{x_i, \mathbf{pa}_i\}} P(x_i, \mathbf{pa}_i) \cdot P(\mathbf{x}_{[d]} \setminus \{x_i, \mathbf{pa}_i\} \mid x_i, \mathbf{pa}_i) \\
&= \sum_{i=1}^{d} \sum_{x_i, \mathbf{pa}_i} \log \frac{P(x_i \mid \mathbf{pa}_i)}{Q(x_i \mid \mathbf{pa}_i)} \cdot P(x_i \mid \mathbf{pa}_i) \cdot P(\mathbf{pa}_i) \\
&\quad \underbrace{\cdot \sum_{\mathbf{x}_{[d]} \setminus \{x_i, \mathbf{pa}_i\}} P(\mathbf{x}_{[d]} \setminus \{x_i, \mathbf{pa}_i\} \mid x_i, \mathbf{pa}_i)}_{=1 \text{ for every choice of } (x_i, \mathbf{pa}_i)} \\
&= \sum_{i=1}^{d} \mathbb{E}_{P(\mathbf{pa}_i^G)}\left[D_{\mathrm{KL}}\left(P(X_i \mid \mathbf{pa}_i^G)\|Q(X_i \mid \mathbf{pa}_i^G)\right)\right]
\end{aligned}$$

To simplify notation, we here leave away the supscript G for the parents and use sums for the expected values, but the continuous case follows with integrals analogously. ■

We can now apply the decomposition above to obtain the result:

$$\begin{aligned}
\pi_G(P)(\mathbf{X}) &= \underset{\tilde{P} \text{ Markovian w.r.t. } G}{\arg\min} \; D_{\mathrm{KL}}(P(\mathbf{X}) \| \tilde{P}(\mathbf{X})) \\
&= \underset{\tilde{P} \text{ Markovian w.r.t. } G}{\arg\min} \; \sum_{i=1}^{d} \mathbb{E}_{P(\mathbf{pa}_i^G)} \left[D_{\mathrm{KL}} \left(P(X_i \mid \mathbf{pa}_i^G) \| \tilde{P}(X_i \mid \mathbf{pa}_i^G) \right) \right] + D_{\mathrm{KL}}(P(\mathbf{X}) \| \hat{P}(\mathbf{X})) \\
&= \underset{\tilde{P} \text{ Markovian w.r.t. } G}{\arg\min} \; \sum_{i=1}^{d} \mathbb{E}_{P(\mathbf{pa}_i^G)} \left[D_{\mathrm{KL}} \left(P(X_i \mid \mathbf{pa}_i^G) \| \tilde{P}(X_i \mid \mathbf{pa}_i^G) \right) \right]
\end{aligned}$$

Note that the second term in line 2 does not depend on $\tilde{P}$. The claim can then be deduced from noting that $\tilde{P}(\mathbf{X}) = \prod_{i=1}^{d} P(X_i | \mathbf{PA}_i^G)$ sets the sum to zero. ■

A.4. Proof of Theorem 16

Theorem 16 *[Known interventions] Given causal models $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ and a set of interventional distributions $\mathscr{P}_{\mathcal{E}} = ((P^e, \mathcal{I}_e))_{e \in \mathcal{E}}$, define Q^e for each environment $e \in \mathcal{E}$ as follows:*

$$Q^e(\mathbf{X}) = \prod_{i \notin \mathcal{I}_e} Q(X_i \mid \mathbf{PA}_i^{G_Q}) \cdot \prod_{j \in \mathcal{I}_e} P^e(X_j \mid \mathbf{PA}_j^{G_Q}) \tag{15}$$

where $P^e(X_j \mid \mathbf{PA}_j^{G_Q})$ denotes the mechanisms changed as the result of the environment shift. Then we have that

$$D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} D_{\mathrm{KL}}(P^e(\mathbf{X}) \| Q^e(\mathbf{X})) \tag{16}$$

Proof We show the identity using Lemma 14 noting that Q^e is still Markovian with respect to G_Q (although it may no longer be faithful to G_Q, e.g. if an intervention was hard). Thus, we get the following decomposition

$$\begin{aligned}
&\frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} D_{\mathrm{KL}}(P^e(\mathbf{X}) \| Q^e(\mathbf{X})) \\
&= \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \left[\sum_{i=1}^{d} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_Q})} \left[D_{\mathrm{KL}} \left(P^e(X_i \mid \mathbf{pa}_i^{G_Q}) \| Q^e(X_i \mid \mathbf{pa}_i^{G_Q}) \right) \right] + D_{\mathrm{KL}}(P^e(\mathbf{X}) \| \pi_{G_Q}(P^e)) \right]
\end{aligned}$$

By assumption, we have that

$$Q^e(X_i \mid \mathbf{PA}_i^{G_Q}) = \begin{cases} P^e(X_i \mid \mathbf{PA}_i^{G_Q}), & \text{if } i \in \mathcal{I}_e \\ Q(X_i \mid \mathbf{PA}_i^{G_Q}) & \text{if } i \in [d] \backslash \mathcal{I}_e \end{cases} \tag{26}$$

Thus, the terms corresponding to $i \in \mathcal{I}_e$ vanish yielding

$$\begin{aligned}
&= \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \left[\sum_{i \in [d] \backslash \mathcal{I}_e} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_Q})} \left[D_{\mathrm{KL}} \left(P^e(X_i \mid \mathbf{pa}_i^{G_Q}) \| Q(X_i \mid \mathbf{pa}_i^{G_Q}) \right) \right] + D_{\mathrm{KL}} \left(P^e(\mathbf{X}) \| \pi_{G_Q}(P^e) \right) \right] \\
&= D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q})
\end{aligned}$$

■

A.5. Proof of Lemma 18

Lemma 18 *Let $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ be CGMs, such that G_P and G_Q share the same skeleton and $X_i \overset{G_Q}{\to} X_j$, $X_i \overset{G_P}{\leftarrow} X_j$ be a flipped edge between G_P and G_Q. Further, let P^e be an interventional distribution generated from $\mathscr{P}$.*

(i) If there exists an unblocked path $e \overset{G_P}{\to} X_i$ given $\mathbf{PA}_j^{G_Q} \backslash X_i$, then $P^e(X_j \mid \mathbf{PA}_j^{G_Q}) \neq P(X_j \mid \mathbf{PA}_j^{G_Q})$.

(ii) If there exists an unblocked path $e \overset{G_P}{\to} X_j$ given $\mathbf{PA}_i^{G_Q}$, then $P^e(X_i \mid \mathbf{PA}_i^{G_Q}) \neq P(X_i \mid \mathbf{PA}_i^{G_Q})$.

Proof This Lemma follows from Corollary 4.4. by Perry et al. (2022) using $\mathbf{Z} = \mathbf{PA}_i^{G_Q}$. If case (i), then there exists an unblocked path to a conditioned child of X_j and thus e is d-connected to X_j resulting in a change in the corresponding mechanism by faithfulness. Otherwise, if case (ii), there is an unblocked path to an unconditioned parent of X_i and therefore e is d-connected to X_i again resulting in a change in the corresponding mechanism.

■

A.6. Proof of Theorem 19

Theorem 19 *Let $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ be CGMs and $\mathcal{E}$ a set of environments. Now assume that for all (unoriented) edges $X_i \overset{G_Q}{\to} X_j$ there exists an environment $e \in \mathcal{E}$ such that one of the following conditions holds:*

(i) There exists an unblocked path $e \overset{G_Q}{\to} X_i$ given $\mathbf{PA}_j^{G_Q} \backslash X_i$ and $j \notin \mathcal{I}_e$

(ii) There exists an unblocked path $e \overset{G_Q}{\to} X_j$ given $\mathbf{PA}_i^{G_Q}$ and $i \notin \mathcal{I}_e$

Then, $\mathscr{P}, \mathscr{Q}$ are interventionally equivalent, $\mathscr{P} \sim_{\mathcal{E}} \mathscr{Q}$, if and only if $\mathscr{P} = \mathscr{Q}$.

Proof Let's first assume that $(P, G_P) = (Q, G_Q)$. By our assumption on the multi-environment distributions 6 and 7, we have for all $e \in \mathcal{E}$

$$P^e(X_i \mid \mathbf{PA}_i^{G_Q}) = P^e(X_i \mid \mathbf{PA}_i^{G_P}) = P(X_i \mid \mathbf{PA}_i^{G_P}) = Q(X_i \mid \mathbf{PA}_i^{G_Q}) \tag{27}$$

if and only if $i \in [d] \backslash \mathcal{I}_e$. Thus, all the conditional divergences in the IKL divergence vanish. Since $G_P = G_Q$, P is Markovian w.r.t. G_Q and so are all interventional distributions. Thus, all the residual terms $D_{\mathrm{KL}}(P^e(\mathbf{X}) \| \pi_{G_Q}(P^e)(\mathbf{X})) = 0$ and therefore $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = 0$.

Conversely, assume that $(P, G_P) \neq (Q, G_Q)$. If $P \neq Q$, we have that $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) \geq 1/|\mathcal{E}| D_{\mathrm{KL}}(P(\mathbf{X}) \| Q(\mathbf{X})) > 0$. So it suffices to show that $P = Q$, but $G_P \neq G_Q$ implies that $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) > 0$. If $P = Q$, then G_Q is necessarily Markov equivalent to G_P and thus G_Q, G_P share the same skeleton and v-structures. Since $G_P \neq G_Q$, there exist adjacent variables X_i, X_j such that $X_j \overset{G_P}{\to} X_i$, but $X_i \overset{G_Q}{\to} X_j$. If condition (i) is satisfied and there is also an unblocked path in

G_P $e \overset{G_P}{\to} X_i$ given $\mathbf{PA}_j^{G_Q} \backslash X_i$, we have by Lemma 18 that $P^e(X_j \mid \mathbf{PA}_j^{G_Q}) \neq P(X_j \mid \mathbf{PA}_j^{G_Q}) = Q(X_j \mid \mathbf{PA}_j^{G_Q})$ which gives us

$$\mathbb{E}_{P^e(\mathbf{PA}_j^{G_Q})}\left[D_{\mathrm{KL}}\left(P^e(X_j \mid \mathbf{PA}_j^{G_Q}) \| Q(X_j \mid \mathbf{PA}_j^{G_Q})\right)\right] > 0$$

so that $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) > 0$, since $j \notin \mathcal{I}_e$. If otherwise, every such path is blocked in G_P, let $(e, X_0, \ldots, X_i)$ be the shortest of such paths in G_Q, which implies that

- None of the intermediate variables is intervened
- Each variable on the path has exactly one parent in G_Q on the path.

If $X_0 = X_i$, the path is also unblocked in G_P, since X_i is directly intervened upon. So, the length of the path is at least 2. Since this path is blocked in G_P, there has to be a mis-oriented edge on the path in G_P compared to G_Q. Let $X_k \overset{G_Q}{\to} X_l$ and $X_l \overset{G_P}{\to} X_k$ denote the first of such mis-oriented edges on the path. But then $(e, X_0, \ldots, X_k)$ yields an unblocked path $e \overset{G_P}{\to} X_k$ given $\mathbf{PA}_l^{G_Q} \backslash X_k$ and $X_l \notin \mathcal{I}_e$. Therefore, Lemma 18 applies, leading to a nonzero term in the IKL divergence. If condition (ii) is satisfied, and there is an unblocked path in G_p $e \overset{G_P}{\to} X_j$ given $\mathbf{PA}_i^{G_Q}$, we have

$$\mathbb{E}_{P^e(\mathbf{PA}_i^{G_Q})}\left[D_{\mathrm{KL}}\left(P^e(X_i \mid \mathbf{PA}_i^{G_Q}) \| Q(X_i \mid \mathbf{PA}_i^{G_Q})\right)\right] > 0$$

and consequently $D_{\mathrm{IKL}}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) > 0$. Otherwise, we get an unblocked path to an intermediate node in the same way as for case (i).

■

A.7. Proof of Corollary 20

Corollary 20 *Let $\mathscr{P} = (P, G_P), \mathscr{Q} = (Q, G_Q)$ be CGMs with $P = Q$ and $\mathcal{E}$ a set of environments. Assume that we only have partial multi-environmental information in the following ways:*

1) *We only observe a subset of all variables in the causal graph $\mathbf{X}_S \subseteq \mathbf{X}$, i.e. we can only distinguish the graphical sub-models $\mathscr{P}\big|_S, \mathscr{Q}\big|_S$, induced by removing the unobserved variables from distributions and causal graphs.*

2) *We know that the conditions of Theorem 19 are only satisfied for a subset $E_{\mathcal{E}}$ of all edges in $G_Q\big|_S$.*

We define the restricted IKL divergence via:

$$D_{\mathrm{IKL}}^{res}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \sum_{i \in \mathbf{X}_{E_{\mathcal{E}}} \cap \mathcal{I}_e^C} \mathbb{E}_{P^e(\mathbf{pa}_i^{G_Q|_S})}\left[D_{\mathrm{KL}}\left(P^e(X_i \mid \mathbf{pa}_i^{G_Q|_S}) \| Q(X_i \mid \mathbf{pa}_i^{G_Q|_S})\right)\right] \tag{17}$$

summing only over un-intervened variables X such that $(X, \cdot)$ or $(\cdot, X)$ is contained in $E_{\mathcal{E}}$. Then, $D_{\mathrm{IKL}}^{res}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) = 0$ implies that the graphs G_P, G_Q coinicde on the edge set $E_{\mathcal{E}}$. Conversely, if $D_{\mathrm{IKL}}^{res}(\mathscr{P}_{\mathcal{E}} \| \mathscr{Q}) > 0$, there exists a variable $X_i \in \mathbf{X}_{E_{\mathcal{E}}}$ such that $\mathbf{PA}_i^{G_P} \neq \mathbf{PA}_i^{G_Q|_S}$.

Proof Since $P = Q$, we have that G_P, G_Q are Markov equivalent, so they share the same skeleton. In particular, the subgraphs $G_P|_S, G_Q|_S$ have the same skeleton. By definition, $E_{\mathcal{E}}$ contains exactly those edges for which one of the conditions in Theorem 19 is satisfied. If one of the edges in $E_{\mathcal{E}}$ was mis-oriented between $G_P|_S, G_Q|_S$ (and therefore G_P, G_Q), this would lead to a nonzero term in the restricted IKL divergence by the same arguments as in proof of Theorem 19.

Conversely, if we had $\mathbf{PA}_i^{G_P} = \mathbf{PA}_i^{G_Q|_S}$ for all variables X_i that occur in the sum of the IKL divergence, by assumptions 6 and 7, all terms in the IKL would vanish.

■

A.8. Proof of Theorem 22

Proof By Theorem 16:

$$\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} D_{\mathrm{KL}}(P^e(\mathbf{X})\|Q^e(\mathbf{X})) \leq \epsilon$$

Using Pinsker's inequality and the concavity of the square-root function:

$$\begin{aligned}\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} D_{\mathrm{TV}}(P^e(\mathbf{X}), Q^e(\mathbf{X})) &\leq \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \sqrt{D_{\mathrm{KL}}(P^e(\mathbf{X}), Q^e(\mathbf{X}))/2} \\ &\leq \sqrt{\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} D_{\mathrm{KL}}(P^e(\mathbf{X}), Q^e(\mathbf{X}))} \\ &\leq \sqrt{\epsilon}\end{aligned}$$

By Markov's inequality, for at least $1-\rho$ fraction of e's,

$$D_{\mathrm{TV}}(P^e(\mathbf{X}), Q^e(\mathbf{X})) \leq \frac{\sqrt{\epsilon}}{\rho}.$$

The claim now follows from the boundedness of f. ■