Explaining Distribution Shifts from Changing Causal Mechanisms

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

A solar-powered weather-station, one day, detects anomalous power supply. Comparing the data of the last hours to normal data, we notice, that the commanded orientation of the solar-panels normally – but no longer – affects available power. We go out and fix the pointing mechanism. Another time, same problem, same result from data, but the anomalous data is from during the night. Why do we draw a different conclusion? How can the distinction be formally captured and automatically detected? To this end, we define and explore the properties of graphical objects, arising from causal models for multi-context settings – settings where the underlying model varies in response to the value of a "context-indicator" variable – that capture qualitative relations *and observational access*. These not only describe relevant mechanisms within a specific context, but also capture where physical changes must have occurred compared to other observed contexts. We then focus on the identifiability of these objects from data, by connecting them to context-specific independences as well as joint-causal-inference- and transfer-arguments. Potential applications include improvements in the understanding of anomalies or extremes from a causal perspective.

1 Introduction

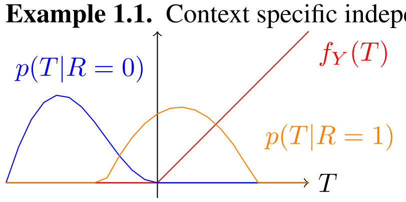
The combination of data from multiple datasets obtained from similar generating processes (contexts), e. g. the transfer of knowledge between contexts, is an important topic of study. Especially for – presumably robust [18; 19] – causal models [12; 9]. Data from multiple contexts has both shared (between contexts) and individual (per context) properties. In order to capture as much information about the underlying system(s) as possible, it seems natural to consider understanding qualitative aspects, for example causal graphs, of both the shared and the individual contexts [4]. We focus on representing qualitative information about the individual contexts, but enriched with information from the joint model. More precisely, we are interested in the following problem: One cannot infer properties of mechanisms outside the range of values that are actually observed, without prior knowledge about extrapolation. But when combining data from multiple contexts, the other contexts do provide knowledge about extrapolation for an individual context. Indeed it turns out, that combining support-properties and causal dependencies in a single graphical objects allows for qualitative statements (like the distinction in the abstract) by tracking few qualitative properties.

This has interesting implications e. g. for understanding anomalies or extreme events. It provides a possible explanation why it seems to often be the case that (presumably robust) causal mechanisms apparently change under extreme conditions (§3.3). Intuitively, per-context information, from the SCM perspective, should be a qualitative change. For example $Y = \mathbb{1}(R) \times X + \eta_Y$. Such a structure induces a context-specific independence (CSI), e. g. $X \perp\!\!\!\perp Y | Z, R = 0$. Intriguingly, the

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

other direction from CSI-structure [13; 4; 8] to SCM-structure is more subtle, as the example below (extending on observations of [1]) illustrates[1]:

Example 1.1. Context specific independence from non-observation:



Consider the following model with dependencies $R \to T \to Y$, think e. g. of $R = 0$ as indicating samples taken in winter, $R = 1$ samples taken in summer, T is the temperature and Y depends on temperature but only if $T > 0$°C (above freezing).

From this example we notice:

(a) f_Y depends on T, but this dependence becomes *within our observations* invisible for $R = 0$ ("system states" with $T > 0$ where also $R = 0$ are never reached).

(b) f_Y itself does not actually change (it doesn't even depend on R).

(c) Given any sort of independence model represented by a graph (e. g. an LDAG [13]), does it agree with absence (a) or presence (b) of the edge $T \to Y$ for $R = 0$?

The point of view (a) prefers a "simpler" model for regime $R = 0$, in an Occam's razor sense for *this regime*, i. e. following the philosophy that a model for this regime should only be as complicated as it has to be to describe this regime relative to no prior knowledge. We will call this concept the "descriptive" graph, for the example above, it should *not* include the edge $T \to Y$. For example there would be no reason to fit a regressor of Y to T in this regime.

The point of view (b) prefers a "simpler" model for regime $R = 0$, in an Occam's razor sense relative to *all* the data. It follows the philosophy that assuming causal models are robust, we should consider validity of transfer the norm and only claim a regime-specific model to be different from the union-model, if there is evidence for this difference. A model for this regime should only be as complicated as it has to be to describe this regime relative to knowledge of the union-model. We will call this concept the "transfer"[2] or "physical" graph. For the example above, it should include the edge $T \to Y$. The intuitive answer to the question from the abstract, takes this perspective as well: There is no evidence for a change of mechanism, so that alone cannot explain the distribution-shift.

Finally, concerning point (c), which is intimately linked to the possibility of constraint-based discovery of these graphical objects – we discuss the identifiability of these structures from data in detail in §4 and §5 – we find that the SCM-centered perspective here includes slightly different information than many commonly used independence models (see §A.3).

Contributions

- We motivate, define and study graphical objects, in part encountered in [1], that capture qualitative information about the causal structure plus availability of observations, with particular interest in their differences.
- We show, that these objects are empirical, i. e. can be identified from data at least in part. In doing so, we focus on the graphs' skeleta (that is, on their adjacencies only).
- We provide a mathematical framework based on solution-functions, that captures implications of CSI in terms of SCM-properties. We focus on a single context-indicator and skeleta, but the framework should allow for the derivation of more general results.

2 Related Literature

Combining datasets for causal modeling, in particular using a context-indicator variable, has been studied extensively to gain insights (e. g. orientations) about the joint- / "union"-model [2; 12; 6; 9]. E g. [2; 12] in particular discuss transportability between contexts, but concerning identifiability

[1] We do not discuss finite-sample properties, but these effects also occur, e. g. if observational support becomes narrow on the source compared to the derivative of the mechanism and noise on the target (Rmk. 3.5).

[2] If we "learn" the, in this example identifiable. f_Y on the pool and transfer it to the regime $R = 0$, the form of f_Y together with the observational support $\mathrm{supp}(P(T|R = 0))$ already explains the absence of the edge.

(structure of hidden confounders), not available observational support. Per-context causal models have been considered e. g. by [20], but without the descriptive vs. physical distinction made here and without the connection to context-specific independence (CSI). Graphical models for CSI in turn have been studied e. g. by [10], or as labeled directed acyclic graphs (LDAGs) [13; 4], but from the independence-model perspective, i. e. without the connection to SCMs (and thus causal modeling). Our approach opens interesting possibilities of connecting the causal and the independence-model world (§A.3). We can treat certain types of cyclic models, much less general than [3], but by adding CSI-information, we show that for these specific cyclic models (away from R) the causal graph of the union-model can be recovered (lemma 5.1), not just its acyclification, causal discovery with cyclic union-graphs is also discussed e. g. in [7; 25]. The lack of observational support we study has certainly been noticed in effect-estimation, where statements can only be made where the fit has support – at least for single-step adjustment [22; 14], for multi-step procedures e. g. the ID-Algo [26; 23] or counter-factual queries like natural direct effects [22] the issue might be more subtle. For counter-factuals more generally similar issues have been observed [17, §5.1], but not treatment from the perspective given here seems to be available. There is also a close connection to minimality conditions affecting graph-definitions (§E), e. g. on parent-sets directly [3, Def. 2.6], "causally minimal" [15, §6.5.3] – replacing faithfulness by using the minimal edge-set for which a Markov-condition holds (see §4.2) – or asking for adjacency-faithfulness [16] only. Finally, lack of observational support may be seen as a missing data problem. The literature combining missing data with causal models typically considers latent variables [24; 30], missing datasets for certain interventions [29; 27], or robustness of causal models [18; 19], which are quite different from the problem we study. See also §A.

Our novel contribution is, that we study multiple meaningful graphs associated to a single context, beyond [1] (see A.2), that can be distinguished from data. These capture qualitative and relevant aspects of the support problem, which seems important to make the problem tractable in practice.

3 Causal Graphical Models

3.1 Structural Causal Models (SCM)

We work within the framework of "structural causal models" (SCM) [11; 15]: We fix a set of **endogenous variables** (observable) $\{X_i\}_{i\in I}$, for some finite I, taking values in $\mathcal{X}_i$, and **exogenous noises** (hidden) $\{\eta_i\}_{i\in I}$, taking values in $\mathcal{N}_i$. We write $V := \{X_i | i \in I\}$ for the set of all endogenous variables, $U := \{\eta_i | i \in I\}$ for the set of all exogenous noises, and for $A \subset V$, let $\mathcal{X}_A := \prod_{j\in A} \mathcal{X}_j$, further $\mathcal{X} := \mathcal{X}_V$ and $\mathcal{N} := \mathcal{N}_U$.

Definition 3.1. A set of **structural equations** (mechanisms) $\mathcal{F} := \{f_i | i \in I\}$ is an assignment of parent-sets $\mathrm{Pa}(i) \subset V$ together with mappings $f_i : \mathcal{X}_{\mathrm{Pa}(i)} \times \mathcal{N}_i \to \mathcal{X}_i$ for all i. An **intervention** $\mathcal{F}_{\mathrm{do}(A=g)}$ on a subset $A \subset V$ is a replacement of $f_j \mapsto g_j$ for $j \in A$. We will only consider "hard" interventions $g_j \equiv x_j = \mathrm{const}$.

Given a distribution P_η of the noises U, if a set of random variables V solving the equations in $\mathcal{F}$ exists, we call their distribution $P_{\mathcal{F},P_\eta}(V)$ an **observable distribution**. For the models we consider, solutions are always unique and are solutions in terms of the noise-values in the intuitive sense, §C.

Definition 3.2. An **SCM** M is a triple $M = (V, U, \mathcal{F})$, with V distributed according to an observable distribution $P_{\mathcal{F},P(U)}(V)$. The **intervened model** $M_{\mathrm{do}(A=g)}$ is an SCM with $M_{\mathrm{do}(A=g)} = (V', U, \mathcal{F}_{\mathrm{do}(A=g)})$ i. e. U is distributed according to the same P_η as for M and the structural equations are replaced according to the intervention.

3.2 Induced Graphical Objects

An important concept in causal inference is to capture qualitative relations between variables as described by (suitably minimal, see next sections) parent-sets $\mathrm{Pa}_i \subset V$ in a **causal graph**, constructed with nodes V and a directed edge from each $p \in \mathrm{Pa}_i$ to X_i. In multi-context settings, there is additional qualitative information available "per context", but as explained in the introduction, multiple meaningful definitions of parent-sets (hence graphs) exist. The simplest way to describe qualitative properties of an SCM is via the mechanisms only:

Definition 3.3. Given mechanisms $\mathcal{F}$, the mechanism-graph $G[\mathcal{F}]$ is constructed with parent-sets Pa such that:

$$X \in \mathrm{Pa}(Y) \quad \Leftrightarrow \quad f_Y \text{ is a non-constant (in } X\text{) function of } X.$$

If one actually knows the underlying SCM, this is well-defined. However, in most applications, one has limited knowledge about the "real" SCM (approximating) a physical system and, thus, uses observed data generated by the SCM to draw conclusions. The choice of suitable (empirically accessible) graphical objects is intimately linked to minimality and faithfulness assumptions (§4.2, §E). To capture "accessible states" we need to build information about observational support into our graphical objects.

Definition 3.4. Given a set of mechanisms $\mathcal{F}$, and a (as of now arbitrary / unrelated to M or $\mathcal{F}$) distribution $Q(V)$ of the observables V, the **observable graph** $G[\mathcal{F}, Q]$ is constructed by defining parent-sets $\mathrm{Pa}' \subset \mathrm{Pa}$ such that:

$$X \in \mathrm{Pa}'(Y) \quad \Leftrightarrow \quad f_Y|_{\mathrm{supp}(Q(\mathrm{Pa}(Y)))} \text{ is non-constant (§B.2) in } X.$$

Note, that this depends qualitatively on Q, in the sense that it *only* depends on the essential support $\mathrm{supp}(Q_A)$ of marginalizations of Q to $A \subset V$.

Remark 3.5. One may replace the above definition by one that also includes a dependence-measure d (or rather its estimator) used to test independences, see §B.3. This seems to feature the same distinction of descriptive / "detectable" vs. physical changes. But it inherently depends on finite-sample-properties, putting it outside the scope of this paper. We focus on the support instead.

This graph moves the problem of observational support from the faithfulness assumption into the graph-definition (§E) in the following sense: If the model M is *not* faithful to its "visible" graph:

Definition 3.6. Given an SCM $M = (V, U, \mathcal{F})$, with observable variables distributed by $P_M(V)$, then the **visible graph** $G^{\text{visible}}[M]$ is the observable graph $G[\mathcal{F}, P_M]$.

Then this failure of faithfulness must arise from a property *other than observational support*. So $G^{\text{visible}}[M]$ is what would commonly be defined as "the" causal graph. This is, in the single context case, the same as the mechanism graph after enforcing a suitable minimality condition (like [3, Def. 2.6]) on $\mathcal{F}$. The visible graph is explicitly constructed as a graph "$G[M]$" associated to an SCM. Other than before (for $G[\mathcal{F}]$ and $G[\mathcal{F}, Q]$), there is more than one meaningful choice here! We fix a (finite, $P(R = r) > 0$) categorical context / "regime-indicator" variable R and want to understand qualitative changes in the model between different values of R (cf. also [10; 13]).

Definition 3.7. Given an SCM $M = (V, U, \mathcal{F})$ and $R \in V$, the **regime graph** (see [1]) is

$$\bar{G}^{\text{descr}}_{R=r}[M] := G[\mathcal{F}_{\mathrm{do}(R=r)}, P_M(V|R=r)].$$

Fixing R to a value, removes dependencies involving R, so we add this information back in by defining $G^{\text{descr}}_{R=r}[M]$ as $\bar{G}^{\text{descr}}_{R=r}[M]$ plus edges involving R in $G^{\text{visible}}[M]$.

This object describes the qualitative relations between variables of the regime-"enforced" model $M_{\mathrm{do}(R=r)}$ that can be learned from the observed distribution P_M (via conditioning) and contains the "descriptive" information about (in)dependences we want to learn (see §1, point (a)).

Remark 3.8. This graph is *very* different from a "conditioned" model: For example there are no spurious links from selection-bias. This is, because this graph describes *properties of the underlying SCM* under constraints on observable "system-states", and makes no reference to e. g. independencies. It is however closely connected to independence properties (cf. §4).

For edges not involving R, and thus for $\bar{G}^{\text{descr}}_{R=r}[M]$, we could replace $\mathcal{F}_{\mathrm{do}(R=r)}$ by $\mathcal{F}$ in definition 3.7, which underlines the idea of describing an object that can be inferred from observations, but contains information about the interventional model. To capture §1, point (b), we use:

Definition 3.9. Given an SCM $M = (V, U, \mathcal{F})$, and $R \in V$, the **transfer / physical graph** is $\bar{G}^{\text{phys}}_{R=r}[M] := G[\mathcal{F}_{\mathrm{do}(R=r)}, P_M(V)]$. and again $G^{\text{phys}}_{R=r}[M]$ adds edges involving R.

As illustrated in the introduction, this keeps links that vanish through changing state-accessibility of the system (it keeps information available on the pool), but deletes those that "explicitly" change via $\mathrm{do}(R = r)$, e. g. if $Y = \mathbb{1}(R) \times X + \eta_Y$ (so it captures a very intuitive notion of "per-regime" changes). Finally interventional models – note, that Def. 3.2 keeps the exogenous noises in the definition of the intervened model, hence it has a "counter-factual" character (§A.4) – correspond to

Definition 3.10. Given an SCM $M = (V, U, \mathcal{F})$, and $R \in V$, the **counter-factual graph** is $G^{\text{CF}}_{R=r}[M] := G[\mathcal{F}_{\text{do}(R=r)}, P_M(V \mid \text{do}(R=r))] = G^{\text{visible}}[M_{\text{do}(R=r)}]$.

See also §A.4, where it is quickly explained why $G^{\text{CF}}_{R=r}$ seems more relevant with experimental data, we focus on observational data here. Finally, some properties of these graphs (proofs are in §B):

Lemma 3.11. *Union Properties, for $G^{\text{union}}[M] := G^{\text{visible}}[M]$:*

(i) $G^{\text{union}}[M]$ is the "union graph" in the sense of [20]

(ii) $G^{\text{union}}[M] = \cup_r G^{\text{phys}}_{R=r}[M]$

(iii) $G^{\text{union}}[M] = \cup_r G^{\text{descr}}_{R=r}[M]$, if M is strongly R-faithful (Def. 4.6)

Point (ii) is of course the motivation of (i) in [20], but here we can explicitly see that in this case (for the union), the specific choice of graph ($G^{\text{phys}}_{R=r}$ vs. $G^{\text{descr}}_{R=r}$) is (mostly) unimportant.

Lemma 3.12. *Relations of edge-sets:*

$$G^{\text{descr}}_{R=r}[M] \quad \subset \quad G^{\text{phys}}_{R=r}[M] \quad \subset \quad G^{\text{union}}[M]$$

writing "$G' \subset G$" if both G and G' are defined on the same nodes, and the subset-relation holds for the edge-sets.

Lemma 3.13. *Physical changes are in regime-children:*
If $Y \neq R$ with $R \notin \text{Pa}^{\text{union}}(Y)$, then $\text{Pa}^{\text{phys}}_{R=r}(Y) = \text{Pa}^{\text{union}}(Y)$.

3.3 Potential Applications

Where are these graphs relevant? For applications like earth-sciences, the problem of restricted support seems to exist at least from a finite sample perspective. Further many important applications here involve the study of extreme events, where a restriction to small regions of the state-space is believed to occur [5; 28] – one possible intuition is, that extremes occur from the coincidence / synergy of different pathways, for example many time-steps with little precipitation followed by high temperatures putting drought extremes in a "corner" of the state-space. It is often somewhat unclear why (presumably robust) causal mechanisms seem to change under extreme conditions. Our approach provides a possible explanation, as causal discovery (e. g. with masking, rmk. D.10) should typically (see §4) recover $G^{\text{descr}}_{R=r}$ at best, thus is very sensitive to observational support. Extreme states (like droughts) are often endogenous, i. e. themselves driven by system-variables (e. g. by soil-moisture feed-backs).

The setup also relates naturally to "technological" data like satellite-telemetry or IT-safety applications, where systems behave much more like state-machines (or actors) with many actions only available in certain states. Note, that here the state typically changes in response to sensory input, so when modeling data about system *and* environment (e. g. by including data for sensory input), the resulting contexts are typically endogenous. While our approach is still very far from systematically recovering a state-machine as a causal model, an understanding of the observations-support properties studied here seems to be an important building-block when approaching this problem. It seems noteworthy, that also a causal perspective on explainable AI (XAI), treating neural network (layers) and their inputs as SCMs, typically have such qualitative structure, e. g. from ReLU activation-functions.

What are these graphs good for? A common problem in practice is, given two (or more) contexts – e. g. normal data and anomalous data – to "explain" (for some meaningful notion of "explain") the difference. If, between the two contexts, a mechanism f_i changes its parents *physically*, then this change at f_i probably should be part of the explanation for changed observations. If, on the other hand, the changes (addition, removal or combinations) of the parents in f_i do not require any explanation beyond the change in support, i. e. if they are purely descriptive (non-physical), then the explanation for the changing observations should be found in the ancestors, not at f_i. E. g. for example 1.1 in the introduction, assuming we observe additional nodes that provide orientation-information (or if there is a mediator $R \to M \to T$), we note, that $T \to Y$ cannot be a physical change because $R \notin \text{Pa}^{\text{union}}(Y)$ (see corollary 5.3 below). So, instead of claiming the two contexts to differ by a change in f_Y (which is indeed *not* the case), we should look further upstream in the graph, which, here, leaves f_T. Given R the only "real" remaining degree of freedom is $P(T|R=0) \neq P(T|R=1)$, which is a surprisingly accurate diagnosis.

Further, also interventional and counterfactual queries happen in a different (non-union) context in terms of knowledge about mechanisms in certain value-ranges. We consider this to be a known problem and especially for counterfactuals it has been discussed from slightly different points of view (see §A.4). Our treatment certainly does not suffice to "solve" this problem, but we show, that including information about knowledge and observability into causal inference – for multi-context data – can (and by the motivations above, maybe should) be systematically approached.

4 Context-Specific Independence

Note, that while changes in $\mathrm{Pa}^{\text{phys}}(Y)$ from mechanism-changes only occur if $R \in \mathrm{Pa}^{\text{union}}(Y)$ (lemma 3.13), "state-access" induced changes can even (also if R is *not* on any cycle in G^{union}), remove links in G^{descr} between ancestors of R. This can be undetectable even from context-specific independencies, if the same link "should" be removed for a specific regime r – requiring us to conditioning on $R=r$ – but gets "reinstated" by selection-bias – because we are conditioning on R. For a concrete example, see D.13 in the appendix. This section shows, that – up to this issue (links vanishing in ground-truth between ancestors of R due to state-access restrictions) – the descriptive graph $G^{\text{descr}}_{R=r}$ can be recovered from combining context-specific and non-context-specific independence-tests, if a suitable faithfulness property is is satisfied (§4.2).

4.1 Markov Properties

Here we study the following question: When can the absence of an edge in the graphical objects – in particular $G^{\text{descr}}_{R=r}$ – defined above, be detected directly from independence constraints? I. e. by Markov properties we refer to factorization properties of the joint probability density and the question of "completeness" of constraint-based causal discovery. For DAGs these properties coincide with the question "does d-separation in the graph imply independence?", and sometimes this is taken as the definition a "Markov property" (e. g. [3]). We are primarily interested in discovering properties of the SCM as described by the graphical objects defined above from data, while the "d-separation criterion", for cyclic models, only recovers "acyclifications" of such graphs.

The Proof-Idea: A more detailed description and proofs can be found in §D. Here we only sketch the proof idea. Commonly one starts from the *local* Markov-property: The idea is that knowing Z containing the parents of Y renders Y independent of all noises other than η_Y, because $Y = f_Y(\mathrm{Pa}_y = \mathrm{pa}_Y, \eta_Y)$. The subtle problem here is to (a) understand this not only for union-parents, but also for parents in $G^{\text{descr}}_{R=r}$ and (b) to then combine both. The issue is, that $G^{\text{descr}}_{R=r}$ is not a causal graph associated to a SCM in the standard sense[3] (i. e. there need not exist an SCM M' with $G^{\text{visible}}[M'] = G^{\text{descr}}_{R=r}$), so the local Markov-property is not obvious anymore. We come back to this momentarily. Seemingly this problem (a) can be resolved by considering a "conditioned" SCM (replace $P(\eta)$ by $P(\eta|R=r)$ and keep $\mathcal{F}$ to obtain M', which confounds ancestors of R, see lemma D.2), but than point (b) becomes even harder – intuitively, since information in causal discovery is in the missing links, one wants to combine information of link-removals from CSI (the "conditional" graph) with link-removals from the union-model, so one is inclined to intersect the respective edge-sets. But problem (b) essentially asks about the connection of the resulting object to the underlying SCM (and the regime-specific SCM $M_{\mathrm{do}(R=r)}$). An important contribution of our proof is, that it shows, how this information ("intersect two graphs") is related to the SCM via $G^{\text{descr}}_{R=r}$. This connection allows then for further inferences in §5 and §5.3.

The way we approach the problem, is by first facing yet another subtlety: The "propagation" of the local information from the local Markov-property to obtain global statements about the graph is normally done via a separation-criterion, that analyzes individual paths in the graph. But what does blocking a path in $G^{\text{descr}}_{R=r}$ mean? The idea we propose is to go from graphical properties to conditional independencies not via a separation-criterion (when blocking Z) and paths as an intermediate step,

[3] In [1], it is shown that there are meaningful "sufficiency" assumptions, s. t. $G^{\text{detect}}_{R=r} = G^{\text{phys}}_{R=r} = G^{\text{CF}}_{R=r} = G^{\text{visible}}[M_{\mathrm{do}(R=r)}]$, in which case the problem (mostly) is reduced to (b). Here we are interested in the differences between those graphical objects in particular, so we need identifiability-results that hold more generally.

but via changes in the noise-distribution (when conditioning on Z) and the form of solution-functions. Here the "form of solution-functions" captures graphical properties, because the system of structural equations can be solved "downstream" starting from root-nodes, successively working down their descendants – as follows (see §C):

Cor. C.5. Given a solvable, weakly regime-acyclic model, then, for any set of variables X:

(a) F_X depends only on noise-terms of ancestors of X in G^{union}.

(b) $F_X^{R=r} := F_X|_{F_R^{-1}(\{r\})}$ depends only on noise-terms of ancestors of X in $G^{\text{descr}}_{R=r}$.

As the reader may have noticed we phrased the local Markov-property above via dependence of noise-terms (rather than independence of non-descendants). Next, consider a conditioning set $Z \supset \mathrm{Pa}(Y)$. The essential trick is, that since knowledge via Z of the parents of Y renders Y independent of all noises other than η_Y, another variable $X = F_X(\eta_A)$ is independent of Y given Z as long as $\eta_A \perp\!\!\!\perp \eta_Y | Z$. But again by the form of solution-functions, this time of F_Z, we know which η_i will be "mixed" (become dependent, see lemma D.2) when conditioning on Z.

Formulating the local Markov-property directly through "dependence on η_Y only" works for our setup immediately. Further it makes problem (a) solvable since Cor. C.5b is applicable for $G^{\text{descr}}_{R=r}$! Finally, these constraints obtained through solution-functions are uniform, in the sense that it does not matter if we used Cor. C.5a or Cor. C.5b to obtain an intermediate result. The obtained statements can thus be easily combined, which eventually allows to resolve problem (b).

The Result: As illustrated in the introduction to this section, we will have to exclude relations between ancestors (beyond the union-graph) from our formal claims (see example D.13), as they are not generally accessible:

Definition 4.1. Define the (identifiable) ancestor–ancestor correction $G^{\text{ident}}_{R=r}$ as follows: Start with $G^{\text{ident}}_{R=r} = G^{\text{descr}}_{R=r}$, then add all edges in G^{union}, between any two ancestors in G^{union} of R to $G^{\text{ident}}_{R=r}$.

Lemma 4.2. *There are no physical ancestor–ancestor problems (proof in §B):* $G^{\text{descr}}_{R=r} \subset G^{\text{ident}}_{R=r} \subset G^{\text{union}}$ *and if M is strongly regime-acyclic, then* $G^{\text{ident}}_{R=r} \subset G^{\text{phys}}_{R=r}$.

Finally, the Markov-property we obtain reads – note the restriction on where to search for Z, which is relevant for causal discovery in practice, is a bit subtle here (see cor. D.9, rmk. D.10):

Proposition 4.3. *Assume the model is strongly regime-acyclic and causally sufficient. If X and Y are non-adjacent in $G^{\text{ident}}_{R=r}$ and both $X, Y \neq R$, then either*

(a) for $Z = \mathrm{Pa}^{\text{union}}(X)$ *or* $Z = \mathrm{Pa}^{\text{union}}(Y)$ *it holds* $X \perp\!\!\!\perp Y | Z$*, or*

(b) for $Z = \mathrm{Pa}^{\text{descr}}_{R=r}(X) - \{R\}$ *or* $Z = \mathrm{Pa}^{\text{descr}}_{R=r}(Y) - \{R\}$ *it holds* $X \perp\!\!\!\perp Y | Z, R = r$.

Further, if either $X \notin \mathrm{Anc}^{\text{union}}(R)$ *or* $Y \notin \mathrm{Anc}^{\text{union}}(R)$*, then (b) applies, otherwise (a) applies.*

Remark 4.4. If one of the variables is R then (for univariate R) no regime-specific tests are available and we have to fall back to the "standard" result (see e. g. [3]): Assume the model is causally sufficient. If R and Y are non-adjacent in $\mathrm{Acycl}(G^{\text{union}})$, then there is $Z = \mathrm{Pa}^{\text{union}}(R)$ or $Z = \mathrm{Pa}^{\text{union}}(Y)$ with $R \perp\!\!\!\perp Y | Z$. If Y is an ancestor of R this does not change the result if the model is strongly regime acyclic. However, if Y is part of a directed cycle involving at least one child of R, then the edge $R \to Y$ in $\mathrm{Acycl}(G^{\text{union}})$ cannot be deleted from our independence-constraints, even if it is not in G^{union}. By the above, together with prop. 4.3, this is the only such issue, that can occur.

4.2 Faithfulness Properties

As is shown in [1] (and repeated in E) standard faithfulness assumptions by a (short) argument justify the following

Assumption 4.5. We assume the model to be R-adjacency-faithful in the sense that for all r:

$$\exists Z \text{ s.t.} \left\{ \begin{array}{c} X \perp\!\!\!\perp Y | Z \quad \text{or} \\ X, Y \neq R \text{ and } X \perp\!\!\!\perp Y | Z, R = r \end{array} \right\} \quad \Rightarrow \quad X \text{ and } Y \text{ are not adjacent in } G^{\text{descr}}_{R=r}$$

The from a theoretical point of view potentially more interesting observations is: The CSI-Markov-property (§4.1) guarantees independences for edges not in $G^{\text{ident}}_{R=r}$, while the R-faithfulness argument only provides dependence-guarantees for edges in $G^{\text{descr}}_{R=r}$. As the counter-example D.13 shows, the Markov-property in general cannot hold for $G^{\text{descr}}_{R=r}$, but it might of course still hold for a graph G^{CSI} "in-between" $G^{\text{descr}}_{R=r} \subset G^{\text{CSI}}_{R=r} \subset G^{\text{ident}}_{R=r}$. It is unclear if such a $G^{\text{CSI}}_{R=r}$ for which both meaningful faithfulness- and Markov-properties hold exists (see §E, §D.4). For the moment, we are primarily interested in relating CSI-information to SCM-information, so we leave details on the CSI-independence-structure of distributions induced by (e. g. regime-acyclic) SCMs to future research.

Further, to recover a union-graph, we will need (see §E.2, §B):

Definition 4.6. We say M is strongly R-faithful, if it is R-faithful and the mechanisms of the union-model are non-deterministic, in the sense, that there is no set of mechanisms $\mathcal{F}'$ which almost always produces the same observations as $\mathcal{F}$, but has different minimal parent-sets.

5 Joint Causal Inference and Transfer

The previous section explained, how (most of) the information of $G^{\text{descr}}_{R=r}[M]$ can be recovered from (testable) independence-constraints (prop. 4.3 and ass. 4.5), leading to a graph (see §D.4) $G^{\text{detect}}_{R=r}$ with $G^{\text{descr}}_{R=r} \subset G^{\text{detect}}_{R=r} \subset G^{\text{ident}}_{R=r}$. Here we study $G^{\text{phys}}_{R=r}[M]$ and $G^{\text{union}}[M]$. We do not know, if $G^{\text{phys}}_{R=r}[M]$ is fully identifiable in general, or if the set of rules we provide is complete. It demonstrates however, that these graphs contain empirically testable information (see also example F.1 and discussion thereafter). We refer to these rules (§5.2) as "JCI-like" as they resemble [12; 9]. (Proofs are in §F.)

5.1 Inferring the Union-Graph

Recall from remark 4.4, that edges from R into directed union-cycles containing a child of R cannot be deleted by our independences. We will hence mostly focus on edges elsewhere in the graph ("away from R"), using the "barred" notation ($\bar{G}^{\text{descr}}_{R=r}$ etc.). Generally, a causal model is only Markov to the acyclification (see e. g. [3]) of its visible ("standard") graph $\mathrm{Acycl}(G^{\text{visible}}[M])$ while, for strongly regime-acyclic models we here have:

Lemma 5.1. *Let M be a strongly R-regime-acyclic, strongly R-faithful, causally sufficient model, then*

$$\bar{G}^{\text{visible}}[M] = \bar{G}^{\text{union}}[M] = \cup_r \bar{G}^{\text{detect}}_{R=r}[M]$$

is identifiable away from R by (R-context-specific) independences.

For edges out of R no context-specific tests are available, so (see Rmk. 4.4): $G^{\text{visible}}[M] = G^{\text{union}}[M] \subset G^{\text{union}}_{\text{detect}}[M] := \cup_r G^{\text{detect}}_{R=r}[M]$, where the difference $G^{\text{union}}_{\text{detect}}[M] - G^{\text{union}}[M]$ consists of edges from R to nodes in union-cycles only.

5.2 Interring the Physical Graph by JCI-like Rules

Similarly, there are properties of $G^{\text{phys}}_{R=r}$ that can be identified from data. We already know $\bar{G}^{\text{detect}}_{R=r}[M] \subset \bar{G}^{\text{phys}}_{R=r}[M] \subset \bar{G}^{\text{union}}[M]$ by lemma 4.2, where the left-hand-side is, by construction §D.4, identifiable (under our assumptions) from data via prop. 4.3 and lemma 4.5, and the right-hand-side is identifiable by lemma 5.1 above. So it will suffice, for understanding $\bar{G}^{\text{phys}}_{R=r}[M]$, to study edges in $\bar{G}^{\text{union}}_{\text{detect}}[M] - \bar{G}^{\text{ident}}_{R=r}[M]$ and decide if those should be in $\bar{G}^{\text{phys}}_{R=r}[M]$ or not. As already noted in lemma 3.13, physical changes occur only in regime-children:

Lemma 5.2. *If $R \notin \mathrm{Anc}^{\text{union}}(Y)$, then $\mathrm{Pa}^{\text{phys}}_{R=r}(Y) = \mathrm{Pa}^{\text{union}}(Y)$, i. e. the change is non-physical (by observational non-accessibility).*

Corollary 5.3. *If $R \notin \mathrm{Anc}^{\text{union}}_{\text{detect}}(Y)$, then $\mathrm{Pa}^{\text{phys}}_{R=r}(Y) = \mathrm{Pa}^{\text{union}}_{\text{detect}}(Y)$.*

If (conditioning on) R does not change the distribution of ancestors, no state-induced effects occur:

Lemma 5.4. *Assuming strong regime-acyclicity. If $X \in \mathrm{Pa}^{\text{union}}(Y) - \mathrm{Pa}^{\text{ident}}_{R=r}(Y)$ and $R \in \mathrm{Pa}^{\text{union}}(Y)$, and $\mathrm{Anc}^{\text{union}}(R) \cap \mathrm{Anc}^{\text{union}}(\mathrm{Pa}^{\text{union}}(Y) - \{R\}) = \emptyset$, then $X \notin \mathrm{Pa}^{\text{phys}}(Y)$ (i. e. the change is "physical" not just by state).*

Corollary 5.5. *Assuming strong regime-acyclicity. If $R \neq X \in \mathrm{Pa}^{\text{union}}_{\text{detect}}(Y) - \mathrm{Pa}^{\text{ident}}_{R=r}(Y)$ and $R \in \mathrm{Pa}^{\text{union}}_{\text{detect}}(Y)$, and $\mathrm{Anc}^{\text{union}}_{\text{detect}}(R) \cap \mathrm{Anc}^{\text{union}}_{\text{detect}}(\mathrm{Pa}^{\text{union}}_{\text{detect}}(Y) - \{R\}) = \emptyset$, then*

(a) there is a link into the strongly connected component of Y that vanishes in G^{phys}, but not in $G^{\text{union}}_{\text{detect}}$, i. e. there is a physical change.

(b) if Y is not part of a directed union-cycle, then $X \notin \mathrm{Pa}^{\text{phys}}(Y)$, i. e. there is a physical change of this particular link.

5.3 Validity of Transfer

JCI-arguments (§5) can exclude the possibility of physical changes, but they can only provide direct evidence in rare cases (lemma 5.4). But variable can depend quantitatively on R:

Example 5.6. If $Y = g(X) + \gamma R + \eta_Y$, with $\gamma \neq 0$, and if $g|_{\mathrm{supp}\, P(X|R=r_0)}$ is constant, $X \in \mathrm{Pa}^{\text{phys}}_{R=r_0}(Y)$, even though $X \notin \mathrm{Pa}^{\text{descr}}_{R=r_0}(Y)$ and $R \in \mathrm{Pa}^{\text{union}}(Y)$.

We sketch a statistical test (see also §F.3), that approaches this problem in analogy to the philosophy of constraint-based causal discovery (CD): For CD, the idea is, that in an Occam's razor sense, a link should be considered relevant to the causal model, if there is evidence for the link to be present, i. e. if independence can be rejected (see discussion of point (a) after example 1.1). For the multi-context case, from the perspective, that causal mechanisms are supposed to be robust, a reasonable null-hypothesis is, to assume, that g (in example 5.6) remains unchanged in the context $R = r_0$. So a link should be removed relative to the union-model if there is evidence for its vanishing (see discussion of point (b) after example 1.1).

In the example above, g is identifiable (in G^{union}), so we can learn g from data. Now, if we can show, that the independence-test we used for CD (of $G^{\text{descr}}_{R=r_0}$, see Rmk. D.10), would have (likely) rejected the independence $X \perp\!\!\!\perp Y | R = r_0$ given the observational distributions (e. g. bootstrapping from the observational distributions) if g had remained valid in $R = r_0$, then we have evidence for g vanishing in $R = r_0$. This formally is captured by the difference of G^{union} and $G^{\text{phys}}_{R=r_0}$ in the sense of Rmk. 3.5.

6 Conclusion

The assumption of positivity, $P > 0$, is quite common and very useful. However, it is not popular for its realism – finite data never gives empirical evidence outside a bounded support, even more so in light of Rmk. 3.5 – but because it dramatically simplifies the problem, by neglecting "purely formal" details that supposedly would not actually affect the conclusions we draw. Generally, this is certainly often true, but as we point out, there are a range of difficulties, where our *qualitative* understanding relies on the the understanding of available observational support. We formally capture such qualitative properties through our descriptive and physical graphs – this includes the example from the abstract, where once a physical and once a descriptive change occurs. Further, as we demonstrate, in multi-context systems, these qualitative properties become accessible, at least in part, from observations. Finally, we hope that the connection between the structure of context-specific independences and SCMs that our objects provide may help to better connect both worlds.

Future-Work: We focused on iid-data here, but time-series data seems like an interesting, even though potentially quite complicated, extension. For time-series, for example with persistent (slowly changing) regimes, the observable support of the stationary distribution should play an important and interesting role. What sounds very technical, captures some intuitive effects: As an example, consider a crossroads, where in one context (state of traffic-lights) the traffic flows in one direction, in the other context in the orthogonal direction. Now if states normally only accessible (by the stationary distribution) in one context (traffic in direction A) at a regime-boundary "lag" into the other context (traffic in direction B), then new phenomena arise.

A more immediate generalization would be in (transfer of) orientations. One can of course use standard orientation-rules per-context, or JCI-rules on the union, but really one would want to combine information from both where possible.

References

[1] A. Anonymous et al. Causal discovery in the presence of endogenous context variables. 2024. Concurrently submitted to NeurIPS (see supplementary materials).

[2] E. Bareinboim and J. Pearl. Transportability of causal effects: Completeness results. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 698–704, 2012.

[3] S. Bongers, P. Forré, J. Peters, and J. M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

[4] J. Corander, A. Hyttinen, J. Kontinen, J. Pensar, and J. Väänänen. A logical approach to context-specific independence. *Annals of Pure and Applied Logic*, 170(9):975–992, 2019.

[5] V. M. Gálfi, V. Lucarini, and J. Wouters. A large deviation theory-based analysis of heat waves and cold spells in a simplified model of the general circulation of the atmosphere. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(3):033404, 2019.

[6] B. Huang, K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf. Causal discovery from heterogeneous/nonstationary data. *The Journal of Machine Learning Research*, 21(1):3482–3534, 2020.

[7] A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Learning linear cyclic causal models with latent variables. *The Journal of Machine Learning Research*, 13(1):3387–3439, 2012.

[8] A. Hyttinen, J. Pensar, J. Kontinen, and J. Corander. Structure learning for bayesian networks over labeled dags. pages 133–144. PMLR, 2018.

[9] J. M. Mooij, S. Magliacane, and T. Claassen. Joint causal inference from multiple contexts. *The Journal of Machine Learning Research*, 21(1):3919–4026, 2020.

[10] H. Nyman, J. Pensar, T. Koski, and J. Corander. Stratified graphical models : Context-specific independence in graphical models. *BAYESIAN ANAL*, 9(4):883–908, 2014. doi: 10.1214/14-BA882. QC 20150225.

[11] J. Pearl. *Causality: Models, reasoning and inference*. Cambride University Press, 2000.

[12] J. Pearl and E. Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595, 2014.

[13] J. Pensar, H. Nyman, T. Koski, and J. Corander. Labeled directed acyclic graphs: a generalization of context-specific independence in directed graphical models. *Data mining and knowledge discovery*, 29:503–533, 2015.

[14] E. Perkovic, J. Textor, M. Kalisch, and M. H. Maathuis. Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs. 2018.

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

[16] J. Ramsey, J. Zhang, and P. L. Spirtes. Adjacency-faithfulness and conservative causal inference. *arXiv preprint arXiv:1206.6843*, 2012.

[17] J. M. Robins and T. S. Richardson. Alternative graphical causal models and the identification of direct effects. *Causality and psychopathology: Finding the determinants of disorders and their cures*, 84:103–158, 2010.

[18] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018. URL `http://jmlr.org/papers/v19/16-432.html`.

[19] D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):215–246, 2021.

[20] B. Saeed, S. Panigrahi, and C. Uhler. Causal structure discovery from distributions arising from mixtures of dags. In *International Conference on Machine Learning*, pages 8336–8345. PMLR, 2020.

[21] R. D. Shah and J. Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514, 2020.

[22] I. Shpitser and T. J. derWeele. A complete graphical criterion for the adjustment formula in mediation analysis. *The international journal of biostatistics*, 7(1), 2011.

[23] I. Shpitser and J. Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1219. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[24] P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT press, 2001.

[25] E. V. Strobl. Causal discovery with a mixture of dags. *Machine Learning*, 112(11):4201–4225, 2023.

[26] J. Tian and J. Pearl. A general identification condition for causal effects. In *Aaai/iaai*, pages 567–573, 2002.

[27] R. E. Tillman. Structure learning with independent non-identically distributed data. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1041–1048, 2009.

[28] H. Touchette. The large deviation approach to statistical mechanics. *Physics Reports*, 478(1-3): 1–69, 2009.

[29] S. Triantafillou and I. Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *arXiv preprint arXiv:1403.2150*, 2014.

[30] J. Zhang. *Causal inference and reasoning in causally insufficient systems*. PhD thesis, Citeseer, 2006.

A Details on Related Literature

The topic presented here has connections to many fields, so we give a more structured overview below. Also the connection to CSI and independence models [10; 13; 4] seems interesting (§D.4), but since we expect most potential readers to come from the causal community, a detailed treatment in the main-text seems ill-placed. Similarly, further details on the connection to counter-factuals are in §A.4.

A.1 Structured Overview

Combining Datasets from different contexts in causal inference has been studied e. g. by [2; 12; 6; 9]. The focus there is usually on gaining orientation-information or statistical power on finite data, i. e. gaining additional information about the union-model. The main technical ingredient is in adding the context-information (e. g. an index) to the pooled dataset as a "context-variable" and to then study the resulting system. We adopt this convention and call this (categorical) variable R ("regime"). [2; 12] in particular discuss transportability between contexts, but concerning identifiability (structure of hidden confounders), not available observational support. For example [20] also explicitly studies graphical models for mixtures, we will for example connect our results to their union-graph. Their focus is in defining graphs for the combined dataset, we focus on different graphs for a *single* context. The reason why there is not a unique (empirically accessible) graph is that we "enrich" this single context by our understanding of the other contexts. So our study also is inherently multi-context, but does not focus on the union-model (see 5.1 however).

Context Specific Independence (CSI) and Graphical Models have been combined from the perspective of encoding information about (the factorization of) a probability distribution, e. g. as stratified graphs [10], or labeled directed acyclic graphs (LDAGs) [13; 4]. The main distinction here is that we are interested primarily in *causal* properties and to this end study connections to SCMs, thereby e. g. to interventional properties. We also establish how one of the graphical models we define relates to CSI. The information our objects encode is subtly different from that encoded in LDAGs, or their induced "context-specific LDAGs" [13, Def. 8]. Both encode CSI properties however, and it should be possible via our results to leverage results about LDAGs for causal inference, and vice versa (for example the construction of counter-examples like D.13 seems much more accessible from an SCM perspective). See also §A.3.

Cyclic Models and Solution Functions have gained increased attention recently. There are other approaches to study cyclic union-models e. g. [7; 25] – cyclic union-models are a possible use-case for context-specific graphs, but not the core content of this paper. The type of cyclicity we allow in our models is extremely simple compared to the general treatment [3], even though they are not simple in the sense defined there. Simple SCMs [3] are cyclic models defined such that their solution-properties (and simple-ness) is stable under interventions and marginalization. It seems to be the case, that the ensuing problems (in particular unsolvable intervened models), occur, if we intervene to a system-state outside of the observational support, so in our "support-aware" philosophy, we should capture the problem "beforehand": We recognize the intervention as involving a transfer-problem and are thus warned, that it may not have a unique or clear solution without further information. We do not study this connection in detail here, but we use solution-functions in §4.1.

Interventions and Counterfactuals For interventions, e. g. by single-step adjustment [22; 14], a lack of support often becomes evident by a lack of training-data, and is comparatively easy to detect and simple to deal with (require expert-knowledge for extrapolation, there is not much to be done from data alone). For multi-step procedures (like the ID-Algo [26; 23]) and especially counterfactual quantities (like natural direct effects [22]) the situation becomes much more complicated. Here the question about which graphical properties even *could* be learned from data have been discussed, see e. g. [17, §5.1], even though the systematic connection to observational support does not seem to have been studied yet. See also §A.4.

Minimality and Faithfulness are also strongly intertwined with how to pick "the correct" graphical models. The most direct approach is by a minimality-condition on parent-sets [3, Def. 2.6] (even though there is a faithfulness assumption about non-determinism implicit for minimal parent-sets to be well-define / unique). For skeleton-discovery, we are primarily interested in adjacency-faithfulness [16], but e. g. [15, §6.5.3] also formalize a "causally minimal" condition which is faithfulness in the sense of independences only occur where they are guaranteed by the Markov-property, which turns out to be quite non-trivial here (§4.2). The context-specific absence of edges itself can be understood as a violation of faithfulness to the union-graph (as noted e. g. by [9, §4.3.7].

Missing Data in causal modeling in the literature usually concerns either latent variables [24; 30], or more abstractly missing data for certain interventions [29; 27] typically for the combination of datasets (see above) or robustness of causal models [18; 19]. The lack of overlap of observations and non-constant mechanism domain seems so far unexplored – certainly people are and have been aware of this issue, but the formal and systematic approach given here seem to be new (see also §3.3).

A.2 Relation to Method-Paper

The problem of an ambiguity in the definition of per-context graphs and its connection to observational access was encountered in [1] during the development of a constraint-based causal discovery method for this setting. There, the focus is on giving meaningful assumptions, under which this problem does not occur (i. e. assumptions under which $G^{\text{detect}}_{R=r} = G^{\text{phys}}_{R=r} = G^{\text{CF}}_{R=r}$), and when efficient (using few tests) causal discovery is possible. For the scenario with $G^{\text{detect}}_{R=r} = G^{\text{phys}}_{R=r} = G^{\text{CF}}_{R=r}$ a Markov-property is shown with (modified) standard tools (path-blocking), plus some tricks involving counter-factuals (see also the footnote in §4.1). The main distinction here is, that we focus on the usefulness, and identification from data, of the *difference between physical and descriptive changes*. This also means, that a Markov-property that holds only under the assumption of $G^{\text{detect}}_{R=r} = G^{\text{phys}}_{R=r} = G^{\text{CF}}_{R=r}$ is insufficient. The more general case shown here, requires a completely different approach (§4.1, §D). The subsequent study of union and *physical* graph, is relative to a suitable proxy $G^{\text{detect}}_{R=r}$ of $G^{\text{descr}}_{R=r}$ (see §D.4), for which, in light of prop. 4.1 as presented here (see also [1, Rmk. 4.2 on Thm. 1]), efficient causal discovery algorithms as in [1] are suitable. Generally, [1] evolves around developing assumptions for a method (and a method), that is both efficient and interpretable in terms of SCMs despite the difficulties that arise from these observations. The present paper is about studying the emerging structure: How do the different per-context graphs relate to each other and to the union-graph, which intuition do they capture, and how can they – in particular also the physical graph and their differences – be identified?

A.3 Connection to Independence Structures

We briefly recall the concept of labeled acyclic directed graphs LDAGs [13]. The underlying system is considered to consist of categorical variables only. Traditionally, the graphical representation of the independence-structure represents dependencies with links, independencies with missing links, in a sparse sense, i. e. if $X \perp\!\!\!\perp Y|Z$ the link from X to Y is also removed. The LDAG then labels these edges with a "stratum" [10] by the following idea (for simplicity we pretend we knew orientations): If $X \to Y$ then test for each combination of values of (other) parents $Z = \mathrm{Pa}(Y) - \{X\}$ of Y if $X \perp\!\!\!\perp Y|Z = z$, in this case add $Z = z$ as a label to the edge. In practice, some PC-like search-procedure can be used [8].

This, in our language, essentially treats every variable as a regime-indicator, thus also contains the information of any specific choice of regime-indicator (called "context-specific LDAG" in [13, Def. 8]). The full LDAG thus contains more information than only that of a context-specific LDAG corresponding to one choice of regime-indicator. The price for this additional generality is the restriction of the setup to categorical variables only, and for discovery from finite data, in cases where one is interested in a specific context-specific LDAG, the detour through the full LDAG is likely not sample-efficient. We think, it is also not to be underestimated, that LDAGs are hard to read, compared to the simpler (because less information-dense) context-specific ones.

Generally, the information encoded in a context-specific LDAG is very similar to our graphical models, there are some things to note however: The context in LDAGs is local – only strata of parents (adjacencies) are encoded – while our graphs also capture non-local effects (e. g. insert a mediator $R \to M \to T$ in example 1.1, then $T \to Y$ vanishes for specific R-contexts, even though R is not adjacent to either T or Y), which is also accessible from observations e. g. through intersection-graphs (Rmk. D.10). We do not know if the authors of [10; 13; 4; 8], were aware of this specific problem, e. g. the formulation used by Corander et al. [4, Conjecture 1 (p. 985)] about the completeness of their CSI-separation rules relative to a hypothesis (as complete for information contained in) I_{loc}, which contains this "local" CSI-information only, i. e. they were clearly very careful in avoiding potentially false claims or conjectures about such non-local problems. They also use positivity of the distribution, via the semi-graphoid axioms (see also last line on [4, p. 983]).

As Corander et al. [4, p. 983] (and others) point out, generally conclusions about information contained in a given CSI-structure (i. e. which other independencies can be derived) is a very hard problem (cf. [4, §2.3]). The results on this topic that [10; 13; 4; 8] and the "Bayesian network / independence-structure community" in general provide could be interesting to the causal community (e. g. for cross-validation of causal discovery results), and vice versa e. g. the construction of counter-examples like example D.13 that are "easy" in the SCM formalism might provide insights for the "Bayesian network community". Further the specific type of non-local CSI we encounter seems potentially interesting for understanding independence-structures as well. We hope our approach opens new connections between both perspectives.

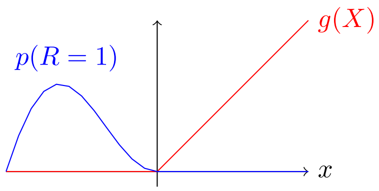
In [4], also a connection to support-properties is used to connect results to the abstract framework of "databases and team semantics". There the abstract model seems to describe the following observation: Given independence the joint distribution is a product, thus its support has certain symmetry-properties. What we study here is the overlap of observational supports and the support of mechanisms, thus a completely different concept.

A.4 Connection to Counter-Factuals

If we are worried about selection-bias, the systematic machinery developed for such questions is the do-calculus. While the "mutilated" graph $G_{\bar{R}}$ defined graphically (does not see qualitative change of mechanisms such as for $Y = \mathbb{1}(R) \times X + \eta_Y$) we may ask: What is the "correct" graph for the intervened model?

This requires additional information about the exogenous noises we consider. The most consistent approach seems to be assuming that the exogenous noises are *not* affected by the intervention in the model. In this case this becomes the counter-factual model [11] describing the world that would have been observed (given the "circumstances" encoded in exogenous noises) if R had been intervened to be r. We will hence call this concept the "counter-factual" graph. This is mostly a matter of perspective and to avoid overloading the term intervened graph typically used in the sense of mutilated graph (see above) with the do-calculus. For the example above, the counter-factual graphs seems to be the "descriptive" graph, but this is a coincidence, and is generally only true if R is exogenous. Indeed the counter-factual graph can even have *more* edges than the union graph:

Example A.1. The Counterfactual Graph can have more Edges than the Union:



Consider the following model with descriptive graph

$$X \to R \to Y,$$

where $f_Y(X, R) = \mathbb{1}(R) \times g(X) + R + \eta_Y$. Values of $X > 0$ and $R = 1$ together are *never* observed, so Y seems to be independent of X.

Here, the shown graph is the "correct" one by the usual means, but note, that intervening on R can make the link $X \to Y$ visible! In fact, if we have interventional ("experimental") data, than this is potentially testable in a multi-context setup, and should be considered a meaningful object. See [17, §5.1] however, where related problems (in a single-context setting) are discussed, and a number of subtleties are pointed out. We will subsequently focus on purely observational data and leave the problem of experimental data to future work.

Finally, we note, that this counter-factual model – under suitable assumptions – can be used as a mathematical trick to proof a (weaker version of) the Markov-property through "standard" path-blocking arguments [1] (because $G^{\text{CF}}_{R=r}[M] = G^{\text{visible}}[M_{\text{do}(R=r)}]$ is a causal graph associated to a causal model in the standard sense).

B Properties of Graphs

B.1 Proofs of Statements in the Main Text

In §3.2 we gave some properties of the studied graphical objects, here we give the corresponding proofs. We start – in slightly altered order compared to the main text – with

Lemma B.1 (Lemma 3.12)**.** *Relations of edge-sets:*

$$G^{\text{descr}}_{R=r}[M] \quad \subset \quad G^{\text{phys}}_{R=r}[M] \quad \subset \quad G^{\text{union}}[M]$$

writing "$G' \subset G$" if both G and G' are defined on the same nodes, and the subset-relation holds for the edge-sets. Generally (i. e. it can happen that) $G^{\text{CF}}_{R=r}[M] \not\subset G^{\text{union}}[M]$ (see example A.1) and $G^{\text{descr}}[M] \not\subset G^{\text{CF}}_{R=r}[M]$.

Proof. This follows directly from the definitions, by $\text{supp}(P(\text{Pa}(X)|R=r)) \subset P(P(\text{Pa}(X)))$. □

Lemma B.2 (Lemma 3.13)**.** *Physical changes are in regime-children:*
If $X \in \text{Pa}^{\text{phys}}_{R=r}(Y)$, and $Y \neq R$ with $R \notin \text{Pa}^{\text{union}}(Y)$, then $\text{Pa}^{\text{phys}}_{R=r}(Y) = \text{Pa}^{\text{union}}(Y)$.

Proof. By definition, $G^{\text{union}}[M] = G^{\text{visible}}[M] = G[\mathcal{F}, P_M]$ and $G^{\text{phys}}_{R=r}[M] = G[\mathcal{F}_{\text{do}(R=r)}, P_M]$. By definition, $\mathcal{F}$ and $\mathcal{F}_{\text{do}(R=r)}$ differ only in f_R and (by setting the parameter $R = r$) for f_i with $R \in \text{Pa}^{\text{union}}(X_i)$. For Y, by hypothesis neither of these two applies, so the same f_Y is in

$\mathcal{F}$ and $\mathcal{F}_{\mathrm{do}(R=r)}$. Since both graph-definitions further use the same support (that of P_M), their parent-definitions for Y agree: $\mathrm{Pa}^{\mathrm{phys}}_{R=r}(Y) = \mathrm{Pa}^{\mathrm{union}}(Y)$. □

Lemma B.3 (Lemma 3.11)**.** *Union Properties, for $G^{\mathrm{union}}[M] := G^{\mathrm{visible}}[M]$:*

(i) $G^{\mathrm{union}}[M]$ is the "union graph" in the sense of [20]

(ii) $G^{\mathrm{union}}[M] = \cup_r G^{\mathrm{phys}}_{R=r}[M]$

(iii) $G^{\mathrm{union}}[M] = \cup_r G^{\mathrm{descr}}_{R=r}[M]$, if M is strongly R-faithful (Def. 4.6)

Proof. (i) $G^{\mathrm{visible}}[M]$ corresponds to the causal graph in the standard sense given a suitable minimality-condition on parent-sets (see §3.2), so it is the graph of the union-model in the sense of [20].

(ii) "$\supset$": By lemma B.1 $G^{\mathrm{phys}}_{R=r}[M] \subset G^{\mathrm{union}}[M]$, so $\cup_r G^{\mathrm{phys}}_{R=r}[M] \subset G^{\mathrm{union}}[M]$.

"$\subset$": Let $X \in \mathrm{Pa}^{\mathrm{union}}(Y)$ be arbitrary. By lemma B.2 we only have to consider links to regime-children Y. By definition $X \in \mathrm{Pa}^{\mathrm{union}}(Y)$ means, there are values $\mathrm{pa}, \mathrm{pa}' \in \mathrm{supp}(P(\mathrm{Pa}^{\mathrm{union}}(Y)))$ which differ only in their X-coordinate (i. e. $\mathrm{pa} = (x, \mathrm{pa}^-)$, $\mathrm{pa}' = (x', \mathrm{pa}^-)$ with pa^- the same value for $\mathrm{Pa}^{\mathrm{union}}(Y) - \{X\}$) such that $f_Y(\mathrm{pa}) \neq f_Y(\mathrm{pa}')$. Since $R \in \mathrm{Pa}^{\mathrm{union}}(Y)$, the tuple pa^- also contains a value r_1 for R. For this r_1 we have $X \in \mathrm{Pa}^{\mathrm{phys}}_{R=r_1}(Y)$, because $G^{\mathrm{phys}}_{R=r_1} = G[\mathcal{F}_{\mathrm{do}(R=r_1)}, P(V)]$ uses the same support (that of $P(V)$) as $G^{\mathrm{union}} = G^{\mathrm{visible}} = G[\mathcal{F}, P(V)]$ and $f'_Y \in \mathcal{F}_{\mathrm{do}(R=r_1)}$ (forcing $R = r_1$ in f_Y) does agrees with the original $f_Y \in \mathcal{F}$ for pa and pa$'$ (as they contain $R = r_1$), so $f'_Y(\mathrm{pa}) = f_Y(\mathrm{pa}) = f_Y(\mathrm{pa}') = f'_Y(\mathrm{pa}')$.

(iii) "$\supset$": By lemma B.1, as in (ii).

"$\subset$": Let $X \in \mathrm{Pa}^{\mathrm{union}}(Y)$ be arbitrary. Define $N_r := F_R^{-1}(\{r\})$ (where F_R is the solution-function for R in terms of noises, see §C), and note that, by F_R being a well-define mapping, $P(\vec{\eta} \in \cup_r N_r) = 1$, using $V_r := F_{\mathrm{Pa}^{\mathrm{union}}(Y)}(N_r)$ and $F_{\mathrm{Pa}^{\mathrm{union}}(Y)}(\cup_r N_r) = \cup_r V_r$ thus so $P(\mathrm{pa} \in \cup_r V_r) = 1$.

By contradiction: Assume it were $X \notin \mathrm{Pa}^{\mathrm{descr}}_{R=r}(Y)$ for all r. Then, by definition, $f_Y|_{V_r}$ is constant in X with probability 1. We can thus define $g^r_Y(\mathrm{Pa}^{\mathrm{union}}(Y) - \{X\})$ such that $P(f_Y = g^r_Y | R = r) = 1$. Finally construct $f'_Y(\mathrm{Pa}^{\mathrm{union}}(Y) - \{X\}, R) := g^R_Y(\mathrm{Pa}^{\mathrm{union}}(Y) - \{X\})$ (i. e. depending on the value r of R choose the corresponding g^r). Then for $\mathcal{F}'$ defined as $\mathcal{F}$ with f_Y replaced by f'_Y, the same observations are obtained with probability 1, but parent-sets differ for Y. □

During the discussion of the Markov-property (§4.1) the graph $G^{\mathrm{ident}}_{R=r}$ is introduced, and the following property is claimed:

Lemma 4.2. *There are no physical ancestor–ancestor problems:*
$G^{\mathrm{descr}}_{R=r} \subset G^{\mathrm{ident}}_{R=r} \subset G^{\mathrm{union}}$ and if M is strongly regime-acyclic, then $G^{\mathrm{ident}}_{R=r} \subset G^{\mathrm{phys}}_{R=r}$.

Proof. Using $G^{\mathrm{descr}}_{R=r} \subset G^{\mathrm{union}}$ (lemma 3.12), by definition 4.1, $G^{\mathrm{ident}}_{R=r} \subset G^{\mathrm{union}}$. The first inclusions is also by definition.

"$G^{\mathrm{ident}}_{R=r} \subset G^{\mathrm{phys}}_{R=r}$": Let $e = (X, Y)$ and edge in $G^{\mathrm{ident}}_{R=r}$.
Case 1 ($X, Y \in \mathrm{Anc}^{\mathrm{union}}(R)$): By $G^{\mathrm{ident}}_{R=r} \subset G^{\mathrm{union}}$ (see above), $e \in G^{\mathrm{union}}$. By lemma 3.13, G^{union} and $G^{\mathrm{phys}}_{R=r}$ differ only by edges pointing into a (union-)child of R. By strong regime-acyclicity, children of R are not union-ancestors of R, so $e \in G^{\mathrm{phys}}_{R=r}$.
Case 2 (otherwise): By definition $e \in G^{\mathrm{descr}}_{R=r}$ in this case. So by lemma 3.12, $e \in G^{\mathrm{phys}}_{R=r}$. □

B.2 Formalization of Non-Constant on Support

In Def. 3.4, we require the restriction of f_Y to the support of a distribution $Q(\mathrm{Pa}(Y))$ to be non-constant in X. Usually this can be thought of as: $\exists\, \mathrm{pa}^-$, values of $\mathrm{Pa}(Y) - \{X\}$, and x, x' values of X such that $(\mathrm{pa}^-, x), (\mathrm{pa}^-, x') \in \mathrm{supp}(Q(\mathrm{Pa}(Y)))$ and $P(f_Y(\mathrm{pa}^-, x, \eta_Y) \neq f_Y(\mathrm{pa}^-, x', \eta_Y)) > 0$. Formally this requires regularity-assumptions (e. g. there are continuous densities, and the f_i are continuous) to exclude degenerate cases like:

Example B.4. Let $Q(X)$ uniform over $(\mathbb{R} - \mathbb{Q}) \cap [0, 1]$, and $f_Y(X, \eta_Y) = \mathbb{1}(X \in \mathbb{Q}) \times X + \eta_Y$. Then $\mathrm{supp}(Q(X)) = [0, 1]$ (it is defined as the closure, which includes the rationals), and f_Y is non-constant on $[0, 1]$, but really f_Y would never "see" the dependence on X.

The more relevant extension to our setting seems to be the finite-sample case §B.3. Nevertheless, the above problem could be fixed, e. g. by defining "non-constant on the support" as: $\exists U, U' \subset \mathcal{X}_X$ and $V \subset \mathcal{X}_{\mathrm{Pa}(Y)-\{X\}}$ such that $U \times V$ and $U' \times V$ are measurable (with respect to Q), and $E[f_Y \,|\, \mathrm{pa} \in U \times V] \neq E[f_Y \,|\, \mathrm{pa} \in U' \times V]$, so one can think of Def. 3.4 using this notion instead. Because measure-theoretic intricacies of the problem do not seem to aid the understanding of the main contents of this paper, we do not detail these problems in the main text.

B.3 Finite-Sample Generalizations

In practice, when only a finite number of samples is available, the distinctions (descriptive vs. physical changes) discussed in this paper also occur for reasons different from non-overlapping supports (of observations and mechanisms): Statistical power of independence tests often relies for example on sufficient width (compared to first derivative of the mechanism and noise on the target) of the observational distribution of the source. More generally, the specific choice of independence-test matters. In this section, we outline how our results generalize to the finite-sample case, how analogues of the previously introduced graphical objects lead to a very similar abstract structure, and why finite-sample properties are even more difficult: There is a "gap" (similar to §D.4) between never detectable (with probability less than a small p_0 detectable) and confidently detectable (with probability larger $1 - \epsilon$ detectable) that does not occur in the asymptotic case.

One may replace the definition 3.4 of $G[\mathcal{F}, Q]$ by the following harder to formalize, but for some problems more practical idea: For an estimator $\hat{d}$ of a dependence-measure d, let $G[\mathcal{F}, Q, \hat{d}, N, p_0, \epsilon]$ be the graph defined via by parent-sets with $X \in Pa(Y)$ if, fixing a sample-count N and error-rate p_0, the estimator $\hat{d}$ has enough (up to ϵ) statistical power to find dependence in the sense of $\exists d_0$: $Q(\hat{d} \geq d_0) > 1 - \epsilon$ – with $P_{\text{null}}(\hat{d} \geq d_0) < p_0$ in the product / independent null-distribution – where $Q(\hat{d})$ is the distribution of $\hat{d}$ evaluated on $(v_X, f_Y(v))$ on N samples v drawn from $Q(V)$. See §F.3. This does not seem to change the abstract structure (kinds of graphs and their relationships), except that an additional "gap" similar to §D.4 opens, because there are edges with effect-sizes that are detectable with probability between p_0 and $1 - \epsilon$.

This captures not only the reality of what we see (the observational support), but also the reality of how we see it (the dependence-test). In practice the result of a causal discovery algorithm does depend on the independence test used, so this describes what is identifiable from data. Its interpretation in terms of causal inference (e. g. effect estimation) is harder, but this is not a failure of the approach, but rather a "real" problem: Given e. g. an SCM with linear effects and Gaussian noise-terms, such that all (non-trivial) effects are large enough for a suitable to this data test (e. g. partial correlation) to have power $1 - \epsilon$, then the discovered graph is valid for effect estimation (up to error-rates bounded by p_0 and ϵ corrected for multiple-testing, we have $G[\mathcal{F}, Q, \hat{d}, N, p_0, \epsilon] = G^{\mathrm{xyz}}[M]$, where "xyz" stands for a graph corresponding to a specific choice of Q, which will also has implications for N). If the data is not suitable to the used test in this sense, we still discover $G[\mathcal{F}, Q, \hat{d}, N, p_0, \epsilon]$, but it is no longer trivially suitable for effect estimation (but e. g. a correlation-based test might still capture causal effect mean-values, even though no longer higher moments). We leave this general problem to future research, but it seem interesting, that statistically precise statements about validity of certain types of effect-estimations appear to be formally accessible. For counter-factual properties, one additionally to $\hat{d}$ needs an estimator for conditional densities.

The choice of independence test seems to usually be seen as governed by properties of available data (which is even in theory only possible to a certain degree [21]), our point here is, that there is an associated graphical object, whose practical usefulness depends on the application additional to the data.

C Solvability and Solution-Functions

Our graphical objects no longer have a simple connection to an set of mechanisms alone, rather they depend on observational support. This means many of the usual proof-techniques (most notably path-blocking) have no evident analogue when discovering these structures from data. A systematic treatment of "Markov"-properties needs a different approach. We show, that the problem can be studied via properties of solution-functions, hence we briefly study solvability of models.

Using only "context insensitive" independence-tests on the "pooled" data, fails to be Markov to the visible graph (some links cannot be detected as absent – actually exactly those links in the acyclification [3].

Some acyclicity-property *is* needed also with CSI. An easy to visualize property is the following "strong" regime-acyclicity (but we often only require the slightly weaker "solvable for R and weakly regime-acyclic", see lemma C.3):

Definition C.1. We call a SCM M weakly (R-)regime-acyclic, if $\forall r$, the regime-graph $G^{\text{descr}}_{R=r}[M]$ is acyclic.

We call a model M strongly (R-)regime-acyclic, if it is weakly (R-)regime-acyclic and no cycle in $G^{\text{union}}[M]$ involves any union-ancestor of R (including R itself).

Easily usable models are typically "solvable" as systems of equations from the noise-terms (this is a notion often employed e. g. to study counterfactuals [11] and has been used to study cyclic models e. g. in [3], see §A):

Definition C.2. A set of mechanisms $\mathcal{F}$ is (uniquely) solvable for X_i, on $\Omega \subset \mathcal{N}$ if there is a (unique) mapping $F_i : \Omega \to \mathcal{X}_i$ such that $X_i = F_i(\eta_1, \dots, \eta_N)$.

$\mathcal{F}$ is (uniquely) solvable on $\Omega \subset \mathcal{N}$, if for all i it is (uniquely) solvable for X_i.

A model M is (uniquely) solvable (for X_i), if its mechanisms $\mathcal{F}$ are (uniquely) solvable (for X_i) on $\operatorname{supp}(P_\eta)$.

We would expect such models to have "good" solution properties. There is a small caveat however: Our graph-definitions (and hence acyclicity-definitions) require a "weak" solvability, namely the observable distribution $P_{\mathcal{F},P_\eta}(V)$ has to exist (with unique support). In practice, when given observations – presumably from an SCM – than this SCM is evidently "weakly solvable" in this sense. Here, "weakly solvable" in turn implies (unique) solvability in the intuitive sense.

Lemma C.3. *Let M be weakly regime-acyclic and the observable distribution $P_{\mathcal{F},P_\eta}(V)$ exists. Then:*

$$M \text{ is strongly regime-acyclic} \quad \Rightarrow \quad M \text{ is uniquily solvable for } R \quad \Leftrightarrow \quad M \text{ is uniquily solvable.}$$

Proof. It is well-know, that acyclic SCMs are solvable. The idea is simply as follows: Let $l(i)$ be the length of the longest incoming path to X_i, i. e. the count of ancestors in a path $\gamma = [A_1 \to A_2 \to \ldots \to X_i]$ with all arrows pointing towards X_i. Then inductively (over l) show M is solvable for all X_i with $l(i) = l$. The inductive start $l = 0$ is trivial, as nodes with $l(X_i)$ are roots (i. e. do not have parents), so $l(i) = 0 \Rightarrow f_i = f_i(\eta_i)$, thus the solution $F_i = f_i$ works. For the inductive step, note, that $l(i) = l + 1 \Rightarrow l(\mathrm{Pa}_i) \leq l$, thus have solution functions F_{Pa_i}, the solution $F_i = f_i(F_{\mathrm{Pa}_i}, \eta_i)$ works for X_i.

Let M be strongly regime-acyclic. There are no cycles involving ancestors (in $G^{\text{union}}[M]$ of R (including R). Thus the above inductive argument works restricted to ancestors of R (including R), because parents of ancestors of R are also ancestors of R and within the support $\Omega = \operatorname{supp}(P_\eta)$ we only need union-parents. Therefore the model is solvable for ancestors (in $G^{\text{union}}[M]$) of R (including R).

Next, knowing F_R, we can "split" the space of noise-values into the disjoint union $N = \coprod_r F_R^{-1}(\{r\})$ and note, that for $\vec{\eta} \in F_R^{-1}(\{r\})$ we know $R = F_R(\vec{\eta}) = r$. Knowing $R = r$, each node depends (for these $\vec{\eta}$) at most on its parents in the respective $G^{\text{descr}}_{R=r}[M]$ (by definition of $G^{\text{descr}}_{R=r}[M]$). Hence we can repeat the argument above on the *acyclic* $G^{\text{descr}}_{R=r}[M]$ to find $X_i = F_i^{R=r}(\vec{\eta})$ for $\vec{\eta} \in F_R^{-1}(\{r\})$ (this $F_i^{R=r}$ is of course the same one as in definition C.4 below, as is immediate for the definition of F_i in the next paragraph).

Define $F_i := F_i^{R=F_R(\vec{\eta})}(\vec{\eta})$. By disjointness of the $F_R^{-1}(\{r\})$ this is well-defined, because every $\vec{\eta}$ is mapped to some r by F_R it is defined everywhere.

Finally the backwards direction M is solvable for $R \Rightarrow M$ is solvable is trivial. $\square$

For solvable models (with almost everywhere continuous densities), conditioning can be understood as restriction of the sample-space:

Definition C.4. If M is solvable, define,

$$F_i^{Z=z} := F_i|_{F_Z^{-1}(\{z\})} : F_Z^{-1}(\{z\}) \to \mathcal{X}_i$$

(we allow Z to be multivariate).

Corollary C.5. *Given a solvable, weakly regime-acyclic model, then, for an arbitrary variable X:*

(a) F_X depends only on noise-terms of ancestors of X in $G^{\text{union}}[M]$, i. e. is constant in all other noise-terms an can thus be written as a function of ancestors' noise-terms only.

(b) $F_X^{R=r}$ depends only on noise-terms of ancestors of X in $G^{\text{descr}}_{R=r}[M]$.

Proof. This is apparent from the proof of lemma C.3:

F_i was constructed inductively from parents and their noise, and from parents of parents and their noise etc. (in $G^{\text{union}}[M]$) thus from noises of ancestors in $G^{\text{union}}[M]$ (with roots depending only on their own noise).

$F_i^{R=r}$ was constructed in the same way from noises of ancestors in $G^{\text{descr}}_{R=r}[M]$. □

Note that corollary C.5 encodes information about support and parental relations on a given support. We use this knowledge to replace path-blocking arguments for obtaining a "Markov"-property.

D Markov-Property

Here, the detailed proof of the Markov-property (Prop. 4.3) is presented. See §4 in the main-text for a high-level overview.

We start from restrictions induced by the graphs on the form of solution-functions. Recall from §C, that because the system of structural equations can be solved "downstream" starting from root-nodes, successively working down their descendants, they depend only on noise-terms of ancestors within the respective graph:

Cor. C.5. *Given a solvable, weakly regime-acyclic model, then, for any set of variables X:*

(a) F_X depends only on noise-terms of ancestors of X in G^{union}.

(b) $F_X^{R=r} := F_X|_{F_R^{-1}(\{r\})}$ depends only on noise-terms of ancestors of X in $G^{\text{descr}}_{R=r}$.

D.1 Graphical Properties Reflected in the Joint Distribution

Next, recall that (generally) such restrictions on functional dependence translate to product-structures on distributions as follows:

Lemma D.1. *Given $A \perp\!\!\!\perp B$ and a mapping $f(A)$ of A only, then*

$$P(A, B|f(A)) = P(A|f(A)) \times P(B)$$

Proof. $P(A, B|f(A)) = P(B, A|f(A)) = P(B|A, f(A)) \times P(A|f(A)) = P(B|A) \times P(A|f(A)) = P(B) \times P(A|f(A))$, where the last equality is by $A \perp\!\!\!\perp B \Leftrightarrow P(B|A) = P(B)$. □

We can use this, to see which part of the "noise-space" is affected by conditioning. Note that the real power of this approach is hidden in the knowledge about ancestral relations via Cor. C.5 *combining* information about the two different graphs G^{union} and $G^{\text{descr}}_{R=r}$. We write "$P(\{\eta_i\})$" for $P(\eta_1, \ldots, \eta_N)$ for the N noise-terms of the N observables X_i. We then use set-notation to make restrictions more explicit (e. g. $\{\eta_i | i \in A\}$ instead of η_A).

Lemma D.2. *Given a solvable, weakly regime-acyclic, causally sufficient model, and a set Z of variables, then,* *(a) Using* $A := \mathrm{Anc}^{\text{union}}(Z)$*:*

$$P(\{\eta_i\}|Z) = P(\{\eta_i|i \in A\}|Z) \times \prod_{j \notin A} P(\eta_j)$$

In particular:

$$k \notin A \quad \Rightarrow \quad P(\eta_k|Z) = P(\eta_k)$$

(b) If $R \notin Z$ *and fixing a value* $R = r$*, using* $A_r := \mathrm{Anc}^{\text{union}}(R) \cup \mathrm{Anc}^{\text{descr}}_{R=r}(Z)$*:*

$$P(\{\eta_i\}|Z, R = r) = P(\{\eta_i|i \in A_r\}|Z, R = r) \times \prod_{j \notin A_r} P(\eta_j)$$

In particular:

$$k \notin A_r \quad \Rightarrow \quad P(\eta_k|Z, R = r) = P(\eta_k)$$

Proof. (a) By corollary C.5a, F_Z depends only on noise-terms of ancestors A of Z in G^{union}. In particular we can write $Z = F_Z(\{\eta_i|i \in A\})$. Using this and lemma D.1, which applies by causal sufficiency:

$$\begin{aligned}
&P(\{\eta_i\}|Z = z) \\
&= P\big(\{\eta_i|i \in A\}, \{\eta_j|j \notin A\} \quad | \quad F_Z(\{\eta_i|i \in A\}) = z\big) \\
&= P\big(\{\eta_i|i \in A\} \quad | \quad F_Z(\{\eta_i|i \in A\}) = z\big) \times P\big(\{\eta_j|j \notin A\}\big)
\end{aligned}$$

The first term is indeed just $P(\{\eta_i|i \in A\}|Z = z)$, while the second term is a product by causal sufficiency. The second claim (of part a) follows by marginalizing this.

(b) By corollary C.5a, F_R depends only on noise-terms of ancestors of R in G^{union}. In particular we can write $R = F_R(\{\eta_i|i \in A_r\})$ (with trivial dependence on elements in A_r not in $\mathrm{Anc}^{\text{union}}(R)$). By corollary C.5b, $F_Z^{R=r} = F_Z|_{F_R^{-1}(\{r\})}$ depends only on noise-terms of ancestors of Z in $G^{\text{descr}}_{R=r}$. In particular we can write $Z = F_Z(\{\eta_i|i \in A_r\})$ (with trivial dependence on elements in A_r not in $\mathrm{Anc}^{\text{descr}}_{R=r}(Z)$). Using this and lemma D.1, which applies by causal sufficiency:

$$\begin{aligned}
&P(\{\eta_i\}|Z = z, R = r) \\
&= P\big(\{\eta_i|i \in A_r\}, \{\eta_j|j \notin A_r\} \quad | \quad F_R(\{\eta_i|i \in A_r\}) = r, \quad F_Z(\{\eta_i|i \in A_r\}) = z\big) \\
&= P\big(\{\eta_i|i \in A_r\}, \{\eta_j|j \notin A_r\} \quad | \quad F_R(\{\eta_i|i \in A_r\}) = r, \quad F_Z^{R=r}(\{\eta_i|i \in A_r\}) = z\big) \\
&= P\big(\{\eta_i|i \in A_r\} \quad | \quad F_R(\{\eta_i|i \in A_r\}) = r, \quad F_Z^{R=r}(\{\eta_i|i \in A_r\}) = z\big) \times P\big(\{\eta_j|j \notin A_r\}\big)
\end{aligned}$$

Again, the first term is just $P(\{\eta_i|i \in A_r\}|R = r, Z = z)$, while the second term is a product by causal sufficiency. The second claim (of part b) follows by marginalizing this. □

I. e. on the "noise-space", selection-bias from conditioning is confined to "sources" η_i from A (or A_r respectively). The idea is now, to separate two variables, not by explicitly blocking all paths, but by building a "barrier" B to divide the system (by conditioning) into two regions of noise-terms affecting one variable vs. those affecting the other, and using the observation above (lemma D.2), to choose B such that selection-bias also does not mix those two regions.

Many ideas of the "standard" setup carry over, for example the "local Markov-Property" formalizes the observation, that, given its parents, a variable X_k, depends *only* on its "own" noise-term η_k. Hence the parents separate the "region" containing only η_k from all other noises (thus from upstream variables) and if X_k is not included in a directed cycle conditioning on the parents will not induce selection-bias ($\eta_k \perp\!\!\!\perp \eta_i | \mathrm{Pa}_k$). Here this can be formulated as a "barrier" against all other noise-terms (lemma D.5).

D.2 Definitions and their Properties

Immediately from the solution-properties Cor. C.5, we can relate variables to the sources of their randomness:

Definition D.3. Noise-sources of observations:

(a) The source of a set of variables X is $\mathrm{Source}(X) = \mathrm{Anc}^{\text{union}}(X)$.

(b) The r-source is $\mathrm{Source}_r(X) = \mathrm{Anc}^{\text{descr}}_{R=r}(X)$.

If we do not block paths, we need some other notion of separation, following the idea of studying the changes to the noise-space:

Definition D.4. Separation from noise-sources:

(a) A barrier B separating a set of variables Y from the noise-sources of another set of variables C is a set of variables disjoint from Y (i. e. $B \cap Y = \emptyset$; but *not* necessarily from C) such that $Y \perp\!\!\!\perp \eta_C | B$.

(b) An r-barrier B separating Y from the noise-sources of C is a set of variables disjoint from Y with $R \in B$ such that $Y \perp\!\!\!\perp \eta_C | B', R = r$ (where $B' = B - \{R\}$).

Such "barriers" exist: The "local" Markov property, essentially says, that parent-sets (from a suitable graph), block out all other (exogenous) noise-terms, it can be formulated in this language as:

Lemma D.5. *Local Markov Property for Barriers (assuming causal sufficiency):*

(a) For any variable Y which is not part of a directed cycle in G^{union}, the set $B = \mathrm{Pa}^{\text{union}}(Y)$ is a barrier separating Y from the noise-sources of any set C not containing Y.

(b) For any variable Y with $R \neq Y$, which is not part of a directed cycle in $G^{\text{descr}}_{R=r}$, and with $Y \notin \mathrm{Anc}^{\text{union}}(R)$, the set $B = \mathrm{Pa}^{\text{descr}}_{R=r}(Y) \cup \{R\}$ is an r-barrier separating Y from the noise-sources of any set C not containing Y.

Proof. (a) Let $B = \mathrm{Pa}^{\text{union}}(Y)$, then $Y = f_Y(B = b, \eta_y)$, we write (for fixed b) $f_Y(b, -)$ for the mapping $n_Y \mapsto f_Y(b, n_Y)$ in particular for measurable U_Y and almost all b

$$P(y \in U_Y | B = b) = P(n_Y \in f_Y(b, -)^{-1}(U_Y) | B = b),$$

or written as a pushforward $P(Y|B = b) = f_Y(b, -)_* P(\eta_Y | B = b)$, which is determined by $P(\eta_Y|B)$. Since, by hypothesis, Y is not part of a directed cycle $Y \notin \mathrm{Anc}^{\text{union}}(B)$, thus by lemma D.2a (second part), $P(\eta_Y|B) = P(\eta_Y)$. By causal sufficiency thus $Y \perp\!\!\!\perp \eta_C|B$ if $Y \notin C$.

(b) Let $B = \mathrm{Pa}^{\text{descr}}_{R=r}(Y) \cup \{R\}$, $B' = B - \{R\}$, then if $R = r$ we have almost surely $Y = f_Y(B' = b', \eta_y)$: By definition of $G^{\text{descr}}_{R=r}$, if $R = r$ then f_Y almost surely depends only on B (potentially trivially on R) and η_Y. Thus again, for measurable U_Y almost always (with $b = (b', r)$)

$$P(y \in U_Y | B' = b', R = r) = P(n_Y \in f_Y(b, -)^{-1}(U_Y) | B' = b', R = r),$$

or written as a pushforward $P(Y|B' = b', R = r) = f_Y(b, -)_* P(\eta_Y | B = b)$, which is determined by $P(\eta_Y|B)$. Since, by hypothesis, Y is not part of a directed cycle $Y \notin \mathrm{Anc}^{\text{descr}}_{R=r}(B)$. Further, by hypothesis, $Y \notin \mathrm{Anc}^{\text{union}}(R)$, thus by lemma D.2b (second part), $P(\eta_Y|B = b) = P(\eta_Y)$. By causal sufficiency thus $Y \perp\!\!\!\perp \eta_C | B', R = r$ if $Y \notin C$. $\square$

Most importantly, "any set C not containing Y" in the previous lemma includes $\mathrm{Source}(X)$ if $Y \notin \mathrm{Anc}_{\text{union}}(X)$ (similarly for (b)), so we will be able to relate noise-space properties back to properties of observables.

Definition D.6. Separation of observables:

(a) A barrier B separating two sets of variables X from Y is a barrier separating Y from the noise-sources of $\mathrm{Source}(X)$, with $B \cap X = \emptyset$. (Thus $B \cap (X \cup Y) = \emptyset$, by def. D.4.)

(b) A r-barrier B separating two sets of variables X from Y is a r-barrier separating Y from the noise-sources of $\mathrm{Source}_r(X)$, with $B \cap X = \emptyset$. (Thus $B \cap (X \cup Y) = \emptyset$, by def. D.4.)

Note, that the noise-barriers provided by the local Markov condition automatically "qualify" to separate $X \neq R$ and Y if X is not a (direct) parent (in the respective graph) of Y. Further, these (def. D.6) indeed relate to independences on the observables:

Lemma D.7. *Separation implies independence:*

(a) *If B is a barrier separating X from Y, then $X \perp\!\!\!\perp Y|B$.*

(b) *If B is a r-barrier separating X from Y, then $X \perp\!\!\!\perp Y|B', R = r$, with $B' = B - \{R\}$.*

Proof. (a) By definition, a barrier B between X and Y is a barrier separating Y from noise of $\mathrm{Source}(X) = \mathrm{Anc}_{\text{union}}(X)$. I.e. $Y \perp\!\!\!\perp \eta_{\mathrm{Anc}^{\text{union}}(X)}|B$, but by corollary C.5a, F_X depends only on noise-terms of ancestors of X in G^{union}, so that also $Y \perp\!\!\!\perp F_X(\eta_{\mathrm{Anc}^{\text{union}}(X)})|B$, with $X = F_X(\eta_{\mathrm{Anc}^{\text{union}}(X)})$ this proves claim (a).

(b) By definition, a r-barrier B between X and Y is an r-barrier separating Y from the noise of $\mathrm{Source}_r(X) = \mathrm{Anc}^{\text{descr}}_{R=r}(X)$. I.e. $Y \perp\!\!\!\perp \eta_{\mathrm{Anc}^{\text{descr}}_{R=r}(X)}|B', R = r$ (where $B' = B - \{R\}$), but by corollary C.5b, $F^{R=r}_X$ depends only on noise-terms of ancestors of X in $G^{\text{descr}}_{R=r}$. By conditioning on $R = r$ we restrict ourselves to noise-terms in $F_R^{-1}(\{r\})$, thereby considering the restriction $F^{R=r}_X = F_X|_{F_R^{-1}(\{r\})}$ suffices. Thus $Y \perp\!\!\!\perp F^{R=r}_X(\eta_{\mathrm{Anc}^{\text{descr}}_{R=r}(X)})|B', R = r$. Again the claim follows by $X = F^{R=r}_X(\eta_{\mathrm{Anc}^{\text{descr}}_{R=r}(X)})$ (whenever defined, i.e. whenever $R = r$). □

Now, that we have a framework for replacing path-blocking arguments in a way suitable to the problem at hand, we can return to the Markov-properties of our systems.

D.3 The Markov-Property

As illustrated in the main text §4.1 (see also example D.13), we will have to exclude relations between ancestors (beyond the union-graph) from our formal claims, as they are not generally accessible (see also §D.4 however):

Definition 4.1. Define the (identifiable) ancestor–ancestor correction $G^{\text{ident}}_{R=r}$ as follows: Start with $G^{\text{ident}}_{R=r} = G^{\text{descr}}_{R=r}$, then add all edges in G^{union}, between any two ancestors in G^{union} of R to $G^{\text{ident}}_{R=r}$.

Remark D.8. G^{union} and $G^{\text{phys}}_{R=r}$ differ only by edges pointing into a (union-)child of R (lemma 3.13), so "G_{union}" in the definition above may be replaced by "$G^{\text{phys}}_{R=r}$" as these always agree on edges between ancestors. In particular the "ancestor–ancestor" problem will never be an issue if we are interested in $G^{\text{phys}}_{R=r}$ (see §5).

Knowing, what separating barriers may look like (by the "local" Markov property lemma D.5), and how to use them to obtain independence-relations on observables (def. D.6, lemma D.7), we finally obtain:

Proposition 4.3. *Assume the model is strongly regime-acyclic and causally sufficient. If X and Y are non-adjacent in $G^{\text{ident}}_{R=r}$ and both $X, Y \neq R$, then either*

(a) *for $Z = \mathrm{Pa}^{\text{union}}(X)$ or $Z = \mathrm{Pa}^{\text{union}}(Y)$ it holds $X \perp\!\!\!\perp Y|Z$, or*

(b) *for $Z = \mathrm{Pa}^{\text{descr}}_{R=r}(X) - \{R\}$ or $Z = \mathrm{Pa}^{\text{descr}}_{R=r}(Y) - \{R\}$ it holds $X \perp\!\!\!\perp Y|Z, R = r$.*

Further, if either $X \notin \mathrm{Anc}^{\text{union}}(R)$ or $Y \notin \mathrm{Anc}^{\text{union}}(R)$, then (b) applies, otherwise (a) applies.

Proof. Case 1 (both X and Y are (union-)ancestors of R): By strong regime-acyclically w.l.o.g. $Y \notin \mathrm{Anc}^{\text{union}}(X)$. In this case, by construction of $G^{\text{ident}}_{R=r}$, X and Y are (non-)adjacent in $G^{\text{ident}}_{R=r}$ if and only if they are (non-)adjacent in G^{union}, thus $X \notin \mathrm{Pa}^{\text{union}}(Y)$. By the local Markov-property lemma D.5a – which applies, because Y is not part of any union-cycle by strong regime-acyclicity – $Z = \mathrm{Pa}_{\text{union}}(Y)$ is a barrier separating Y from the noise of $\mathrm{Anc}^{\text{union}}(X)$. As noted above $X \notin \mathrm{Pa}^{\text{union}}(Y) = Z$, so this is a barrier separating X from Y. Therefore, by lemma D.7a, $X \perp\!\!\!\perp Y|Z$ as claimed.

Case 2 (w.l.o.g. $Y \notin \mathrm{Anc}^{\text{union}}(R)$): Note, that we can further assume w.l.o.g. $Y \notin \mathrm{Anc}^{\text{descr}}_{R=r}(X)$, because, if we had $Y \in \mathrm{Anc}^{\text{descr}}_{R=r}(X)$:

Case 2a ($X \in \mathrm{Anc}^{\text{union}}(R)$): Then if it were $Y \in \mathrm{Anc}^{\text{descr}}_{R=r}(X) \subset \mathrm{Anc}^{\text{union}}(X)$ (by lemma 3.12), this would imply $Y \in \mathrm{Anc}^{\text{union}}(R)$ in contradiction to the hypothesis of the case 2.

Case 2b ($X \notin \mathrm{Anc}^{\text{union}}(R)$), then by weak regime-acyclicity, $X \notin \mathrm{Anc}^{\text{descr}}_{R=r}(Y)$ and we can swap the roles of X and Y to satisfy the w. l. o. g. assumption of the case and $Y \notin \mathrm{Anc}^{\text{descr}}_{R=r}(X)$.

Thus by lemma D.5b – which applies, because by weak regime-acyclicity Y is not part of any cycle in $G^{\text{descr}}_{R=r}$ and $Y \notin \mathrm{Anc}^{\text{union}}(R)$ by hypothesis of the case – using $Z = \mathrm{Pa}^{\text{descr}}_{R=r}(Y)$, we find $Z \cup \{R\}$ is a r-barrier separating Y from the noise of X. Again $X \notin \mathrm{Pa}^{\text{union}}(Y)$ and $X \neq R$, so $X \notin Z \cup \{R\}$, and this is a barrier separating X from Y. By lemma D.7a, $X \perp\!\!\!\perp Y|Z, R = r$ as claimed. $\square$

Remark 4.4. If one of the variables is R then (for univariate R) no regime-specific tests are available and we have to fall back to the "standard" result (see e. g. [3]): Assume the model is causally sufficient. If R and Y are non-adjacent in $\mathrm{Acycl}(G^{\text{union}})$, then there is $Z = \mathrm{Pa}^{\text{union}}(R)$ or $Z = \mathrm{Pa}^{\text{union}}(Y)$ with $R \perp\!\!\!\perp Y|Z$. If Y is an ancestor of R this does not change the result if the model is strongly regime acyclic. However, if Y is part of a directed cycle involving at least one child of R, then the edge $R \to Y$ in $\mathrm{Acycl}(G^{\text{union}})$ cannot be deleted from our independence-constraints, even if it is not in G^{union}. By the above, together with prop. 4.3, this is the only such issue, that can occur.

The restriction on where to search for Z is relevant for causal discovery algorithms in practice, and the following reformulation is helpful to that end:

Corollary D.9. *Given a strongly regime-acyclic, causally sufficient model, X, Y not adjacent in $G^{\text{ident}}_{R=r}$ and both $X, Y \neq R$, then either*

(a) it exists $Z \subset \mathrm{Adj}^{\text{ident}}_{R=r}(X)$ or $Z \subset \mathrm{Adj}^{\text{ident}}_{R=r}(Y)$ with $X \perp\!\!\!\perp Y|Z$, or

(b) it exists $Z \subset \mathrm{Adj}^{\text{ident}}_{R=r}(X) - \{R\}$ or $Z \subset \mathrm{Adj}^{\text{ident}}_{R=r}(Y) - \{R\}$ with $X \perp\!\!\!\perp Y|Z, R = r$.

Proof. We have to show, that the conditioning sets in proposition 4.3 are in the adjacencies of $G^{\text{ident}}_{R=r}$. If (b) applies, then either $Z \subset \mathrm{Pa}^{\text{descr}}_{R=r}(X) \subset \mathrm{Pa}^{\text{ident}}_{R=r}(X) \subset \mathrm{Adj}^{\text{ident}}_{R=r}(X)$ or $Z \subset \mathrm{Pa}^{\text{descr}}_{R=r}(Y) \subset \mathrm{Pa}^{\text{ident}}_{R=r}(Y) \subset \mathrm{Adj}^{\text{ident}}_{R=r}(Y)$ and there is nothing to show. If either $X \notin \mathrm{Anc}^{\text{union}}(R)$ or $Y \notin \mathrm{Anc}^{\text{union}}(R)$, then (b) applies. So the only remaining case is where both X and Y are in $\mathrm{Anc}^{\text{union}}(R)$. In this case, since parents of ancestors of R are again ancestors of R, and edges between nodes in $\mathrm{Anc}^{\text{union}}(R)$ are in G^{union} if and only if they are in $G^{\text{ident}}_{R=r}$ we have (from part (a)) $Z \subset \mathrm{Pa}^{\text{union}}(X) = \mathrm{Pa}^{\text{ident}}_{R=r}(X) \subset \mathrm{Adj}^{\text{ident}}_{R=r}(X)$ or $Z \subset \mathrm{Pa}^{\text{union}}(Y) = \mathrm{Pa}^{\text{ident}}_{R=r}(Y) \subset \mathrm{Adj}^{\text{ident}}_{R=r}(Y)$. $\square$

Remark D.10. There is still a subtle difficulty left: Generally, there is no reason why a model – even if it is faithful to $G^{\text{descr}}_{R=r}$ – would be faithful to $G^{\text{ident}}_{R=r}$. We cannot guarantee links as in example D.13 will be deleted, but they *might* be nevertheless (see also §D.4). So generally by causal discovery using the proposition, one finds a graph $G^{\text{detect}}_{R=r}$, with $G^{\text{descr}}_{R=r} \subset G^{\text{detect}}_{R=r} \subset G^{\text{ident}}_{R=r}$, but for rule (a) one has to test all conditioning-sets contained in the parents in $G^{\text{ident}}_{R=r}$.

An "easy" fix would be to first discover the (acyclification of) the union graph with standard methods, and restrict the search for separating-sets by $G^{\text{ident}}_{R=r} \subset G^{\text{union}} \subset \mathrm{Acycl}(G^{\text{union}})$ to $\mathrm{Acycl}(G^{\text{union}})$. This will do more tests than actually required however.

In practice it might be preferable to either:

(i) Learn $\mathrm{Acycl}(G^{\text{union}})$, then $G^{\text{mask}}_{R=r}$ by masking on $R = r$ (only using rule (b), avoiding the problem discussed above) and then consider "intersection graphs" $(G^{\text{detect}})'_{R=r} := \mathrm{Acycl}(G^{\text{union}}) \cap G^{\text{mask}}_{R=r}$, which in the end also deletes all edges that can be deleted either by (a) or by (b).

(ii) Find suitable assumptions, that allow for more efficient (requiring fewer test, on the pooled data when consistent) algorithms [1].

While the first option sounds simpler and theoretically elegant, the issue of state-access induced vanishing of links between ancestors of R precluding required tests (by searching the wrong adjacencies) in the indicated way seems a bit esoteric for most potential applications, with stability on finite data being a major concern for causal discovery, the second option certainly mandates closer investigation.

D.4 Detectable Graph

As briefly discussed in §4.2, there is a gap between links that always can be removed $G^{\text{ident}}_{R=r}$ (prop. 4.3) and those that never will be removed $G^{\text{descr}}_{R=r}$ (by faithfulness, ass. 4.5). In part, this gap is genuine – counterexamples exist (example D.13) – but in cases where X and Y are ancestors of R, but not both direct parents of R, there might be more generally applicable result than prop. 4.3: From a path-separation perspective, a path containing R as collider being opened by conditioning on R could still be blocked "elsewhere" along the path.

We do not know, if single graph encoding all independence constraints while also being consistent for conclusions drawn via paths exists. However, there is a practical approach via directly encoding independence structure (slightly different from LDAGs, see §A.3) with the connection to SCMs given by prop. 4.3 and assumption 4.5 encoded in lemma D.12:

Definition D.11. Define the "detectable" (independence-)graph $G^{\text{detect}}_{R=r}$ as the "causally minimal" representation [15, §6.5.3]: There is an edge between X and Y if there is no Z with $X, Y \notin Z$ for which at least one of the independence $X \perp\!\!\!\perp Y|Z$ or (if $X, Y \neq R$) the CSI $X \perp\!\!\!\perp Y|Z, R = r$ holds. Orient edges not involving R as in G^{union} (this is well-defined, by lemma D.12 and lemma 4.2 showing $\bar{G}^{\text{detect}}_{R=r} \subset \bar{G}^{\text{union}}$) and edges out of R not in G^{union}, see rmk. 4.4, are oriented out of R, all other edges involving R are also oriented as in G^{union}.

Lemma D.12. *: Connection of $G^{\text{detect}}_{R=r}$ to SCM:*

$$\bar{G}^{\text{descr}}_{R=r} \subset \bar{G}^{\text{detect}}_{R=r} \subset \bar{G}^{\text{ident}}_{R=r}$$

For edges involving R, $G^{\text{detect}}_{R=r}$ contains at least the edges in G^{union}, but may additionally contain edges in $\mathrm{Acycl}(G^{\text{union}})$ out of R.

Proof. The first inclusion is by ass. 4.5, the second one by prop. 4.3. The last statement follows from rmk. 4.4. By strong regime-acyclicity, there are no additionally edges in $\mathrm{Acycl}(G^{\text{union}})$ into R. $\square$

This provides a tight enough connection between CSI-structure and SCMs for the arguments in §5. In practice the results in §5 work for

$$\bar{G}^{\text{descr}}_{R=r} \subset \bar{G}^{\text{detect}}_{R=r} \subset \bar{G}^{\text{phys}}_{R=r}$$

which has the advantage of physical changes being restricted to regime-children (lemma 3.13), which reduces the search-space for CSI-testing and allows for more efficient methods [1].

D.5 Counter-Example to General Case

The following example illustrates the problem of links between ancestors, vanishing by observational access, becoming invisible due to selection bias. See start of §4.

Example D.13. "Selection-bias between ancestors can lead to violations of the Markov-property":

Let $X, Y \in U := \{a_0, a_1, b_0, b_1\}$ categorical variables. Let $X = \eta_X$ (with $P(\eta_X) > 0$), fix $A := \{a_0, a_1\} \subset U$, and $B := \{b_0, b_1\} \subset U$ and the "letter" l and "index" i indicators on U as follows

$$l : U \to \{a, b\}, \begin{cases} a_0, a_1 \mapsto a \\ b_0, b_1 \mapsto b \end{cases}$$

$$i : U \to \{0, 1\}, \begin{cases} a_0, b_0 \mapsto 0 \\ a_1, b_1 \mapsto 1 \end{cases}$$

Then, define for $\eta_Y \in \{0, 1\}$ (with $P(\eta_Y) > 0$):

$$Y = f_Y(X, \eta_Y) := \begin{cases} a_{i(X) \,\mathrm{xor}\, \eta_Y} \text{ if } X \in A \\ b_{\eta_Y} \text{ if } X \in B \end{cases}$$

(On binary variables, the natural choice of binary operators are those of boolean algebra, i. e. of the field $\mathbb{Z}/2\mathbb{Z}$, so that "$\mathrm{xor} = +$" and "$\mathrm{and} = *$". If the reader feels confused by the xor notation, they may think "$+$" (formally mod 2) instead.)

Note, that $l(X) = l(Y)$, and Y clearly depends on X in general. However, for $X \in B$, Y does *not* further depend on the value within B taken by X, i. e. $f_Y|_B$ is independent of X.

Finally, the "regime-indicator" $R \in U$ for $\eta_R \in \{0, 1\}$ (with $P(\eta_R) > 0$ and $P(\eta_R = 1) = p \neq 1/2$):

$$R = l(X)_{\eta_R \operatorname{xor} i(X) \operatorname{xor} i(Y)}$$

This construction has the following interesting properties: $l(R) = l(X)$, hence $l(R) = b \Leftrightarrow l(X) = b$, therefore $\operatorname{supp} P(X|R = b_0) = B$. But $f_Y|_B$ is independent of X (see above), so $X \notin \mathrm{Pa}_{G_{R=b_0}}(Y)$.

However, due to selection bias, this non-adjacency is never detectable: Given $R = b_0$, we know $l(X) = l(R) = b$. Thus also $l(Y) = b$. Further, knowing $0 = i(R) = \eta_R \operatorname{xor} i(X) \operatorname{xor} i(Y)$, we can use information about X to infer the following. If $i(X) = 0$, then the equation above becomes $0 = i(R) = \eta_R \operatorname{xor} i(Y) \operatorname{xor} 0$ with $P(\eta_R = 1) = p$, thus $P(i(Y) = 1|R = b_0, X = b_0) = 1 - p$. On the other hand if $i(X) = 1$, then the equation above becomes $0 = i(R) = \eta_R \operatorname{xor} i(Y) \operatorname{xor} 1$ with $P(\eta_R = 1) = p$, thus $P(i(Y) = 1|R = b_0, X = b_1) = p$.

If it were $X \perp\!\!\!\perp Y|R = b_0$, then $P(i(Y)|R = b_0, X) = P(i(Y)|R = b_0)$ would hold, thus also $p = P(i(Y) = 1|R = b_0, X = b_1) = P(i(Y) = 1|R = b_0) = P(i(Y) = 1|R = b_0, X = b_0) = 1 - p$. But we assumed $p \neq 1/2$. So $X \not\perp\!\!\!\perp Y|R = b_0$ must hold, and we will *always* fail to delete this link from conditional independences alone.

E Faithfulness

There are multiple ways in which faithfulness can fail to hold: Finetuning (cancelations) between paths might be the most discussed one, but also deterministic relations between variables lead to non-unique parent-sets and thus non-well-defined graphs. But also regime-specific changes of mechanism (as for $Y = \mathbb{1}(R) \times X + \eta_Y$) can be understood as a faithfulness violation (the intervened model $\mathcal{F}_{\mathrm{do}(R=r)}$ is not faithful to $G[\mathcal{F}]$), as has also been observed e. g. by [9].

One may thus take a more general perspective: We can think of faithfulness as an assumption "bridging" the gap between observations and a graphical object associated to the model. The "width" of this gap depends on what aspects of the above mentioned problems are encoded in this graphical object! E. g. for a regime-specific change of mechanism (as above), instead of saying "$\mathcal{F}_{\mathrm{do}(R=r)}$ is not faithful to $G[\mathcal{F}]$" and giving up, we clearly want to *learn and understand* a "regime-specific graph", which captures the difference and for which the context-specific independence is "expected" rather than a violation of assumptions.

The additional inclusion of the support into the definition of the graphical object is, from this perspective, just the logical next step. For example looking at the discussion around the definition 3.6 of the "visible" graph, the reader will notice, that we moved the support-related aspects of faithfulness into the graph, while all other aspects (including minimality of the parent-sets) are left in the "gap" that is bridged by assuming "M is faithful to $G^{\text{visible}}[M]$".

Clearly the abstract argument is in no way specific to support aspects of faithfulness, similarly one could e. g. weaken determinism-assumptions encapsulated in the faithfulness assumption by changing the graphical objects etc., however, a thorough and systematic treatment of faithfulness from this perspective turned out to be quite complex, so we will leave this issue to future research for now.

Another faithfulness-related problem is discussed in §D.4.

E.1 Justification of Assumptions in the Main Text

We briefly repeat the argument given in [1], to justify assumption 4.5.

Generally, a probability distribution P is faithful to a DAG G if independence $X \perp\!\!\!\perp_P Y|Z$ with respect to P implies d-separation $X \perp\!\!\!\perp_G Y|Z$ with respect to G. As discussed in [1], this means if $G' \subset G$ is (strictly) sparser, then faithfulness to G' is (strictly) weaker than faithfulness to G. Now, $G^{\text{descr}}_{R=r} \subset G^{\text{union}} = G^{\text{visible}}$, so "$P_M(\ldots)$ is faithful to $G^{\text{descr}}_{R=r}$" is weaker than the standard assumption "$P_M(\ldots)$ is faithful to G^{visible}", and similarly (excluding links involving R), $\bar{G}^{\text{descr}}_{R=r}$ is sparser than what one would expect for a "graph of the conditional model" (there is no selection-bias induced edges in $\bar{G}^{\text{descr}}_{R=r}$) so "$P_M(\ldots|R = r)$ is faithful to $\bar{G}^{\text{descr}}_{R=r}$" is also weaker than what one would expect to assume. One can thus give an adjacency-faithfulness result that essentially corresponds to standard-assumptions as explained above:

Lemma E.1. *Given r, assume both P_M is faithful to $G^{\text{descr}}_{R=r}$ and $P_M(\dots|R=r)$ is faithful to $\bar{G}^{\text{descr}}_{R=r}$ (we will refer to this condition as r-faithfulness, or R-faithfulness if it holds for all r). Then:*

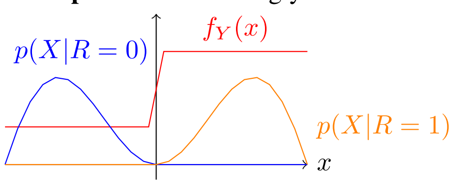
$$\exists Z \text{ s.t. } \left\{ \begin{array}{c} X \perp\!\!\!\perp Y|Z \quad \text{or} \\ X, Y \neq R \text{ and } X \perp\!\!\!\perp Y|Z, R=r \end{array} \right\} \quad \Rightarrow \quad X \text{ and } Y \text{ are not adjacent in } G^{\text{descr}}_{R=r}$$

Proof. The statement is symmetric under exchange of X and Y, so it is enough to show $X \notin \text{Pa}^{\text{descr}}_{R=r}(Y)$. We do so by contradiction: Assume $X \in \text{Pa}^{\text{descr}}_{R=r}(Y)$ and let Z be arbitrary. Z can never block the direct path $X \to Y$, so they are never d-separated $X \not\perp\!\!\!\perp_{G^{\text{descr}}_{R=r}} Y|Z$. By (the contra-position of) the faithfulness assumptions, thus $X \not\perp\!\!\!\perp_P Y|Z$ and if $X, Y \neq R$ also $X \not\perp\!\!\!\perp_P Y|Z, R=r$ (the second statement is by definition the same as $X \not\perp\!\!\!\perp_{P(\dots|R=r)} Y|Z$). □

E.2 An Example that is not Strongly Faithful

Below are an example and discussion to shed some light on why the union-property (lemma 3.11) required an additional faithfulness assumption.

Example E.2. Not strongly R-faithful:



For the functional relationships on the left, Y is a function of X and $X \in \text{Pa}^{\text{union}}(Y)$, but $X \notin \text{Pa}^{\text{descr}}_{R=r}(Y)$ for both $r=0$ and $r=1$.

This is a non-determinism issue (we could write f_Y as a function of R only in the observational support of the union), and is supposed to be excluded by faithfulness (of the union-model). There should be $Z = \text{Pa}^{\text{union}}(Y)$ with $X \not\perp\!\!\!\perp_{G^{\text{union}}} Y|Z, R$ (because the direct path cannot be blocked), but $X \perp\!\!\!\perp Y|Z, R$ (because of the deterministic relation R explains away X). For cyclic models there is a subtle problem however: If Y is part of a directed cycle where X is a parent of another node Z in that cycle, then possibly $X \not\perp\!\!\!\perp Y|Z, R$, i. e. faithfulness may not be violated (formally), because there is a link in the acyclification [3], that "saves" us.

The problem formally also reveals itself as follows: Faithfulness of the union-model implies, that for every Z (again because the direct path cannot be d- or σ-blocked) $X \not\perp\!\!\!\perp Y|Z, R$, which is equivalent (as can be seen e. g. by the characterization of independence as factorization of the joint) to $\exists r$ with $X \not\perp\!\!\!\perp Y|Z, R=r$, which suggests, that there is a context with this link. But there could e. g. be $Z \neq Z'$ with $X \not\perp\!\!\!\perp Y|Z, R=0$ and $X \not\perp\!\!\!\perp Y|Z', R=1$, which in the cyclic case can (non-trivially) happen by union-parents potentially not being valid separating-sets.

This cannot easily be solved by a minimality-condition [3, Def. 2.6] on parents either: In the example above both possible parent-sets of Y, which are $\{X\}$ or $\{R\}$ are of cardinality 1 so no unique minimal parent-set exists, and e. g. the choice via [3, Def. 2.6] is not well-defined (which is not a problem, because normally a suitable faithfulness assumption excludes deterministic relation- ships; this is really a determinism issue, not a minimality issue).

F Details on Connections to JCI- and Transfer-Arguments

This section contains proofs of the statements in §5 and examples.

F.1 Inferring the Union-Graph

Recall from remark 4.4, that edges from R into directed union-cycles containing a child of R cannot be deleted by our independences. We will hence mostly focus on edges elsewhere in the graph, using the "barred" notation ($\bar{G}^{\text{descr}}_{R=r}$ etc.). Generally, a causal model is only Markov to the acyclification of its visible ("standard") graph $\text{Acycl}(G^{\text{visible}}[M])$ while, for strongly regime-acyclic models we here have:

Lemma 5.1. *Let M be a strongly R-regime-acyclic, strongly R-faithful, causally sufficient model, then*

$$\bar{G}^{\text{visible}}[M] = \bar{G}^{\text{union}}[M] = \cup_r \bar{G}^{\text{detect}}_{R=r}[M]$$

is identifiable away from R by (R-context-specific) independences.

Proof. By lemma 3.11, $G^{\text{union}} = \cup_r G^{\text{descr}}_{R=r}$, thus (a) $\bar{G}^{\text{union}} = \cup_r \bar{G}^{\text{descr}}_{R=r}$. While $G^{\text{detect}}_{R=r} \neq G^{\text{descr}}_{R=r}$ in general, by prop. 4.3 and ass. 4.5 (see §D.4), $G^{\text{descr}}_{R=r} \subset G^{\text{detect}}_{R=r} \subset G^{\text{ident}}_{R=r}$ thus (b) $\bar{G}^{\text{descr}}_{R=r} \subset \bar{G}^{\text{detect}}_{R=r} \subset \bar{G}^{\text{ident}}_{R=r}$.

Combining (a) with (b), thus

$$\cup_r \bar{G}^{\text{detect}}_{R=r} \overset{(b)}{\supset} \cup_r \bar{G}^{\text{descr}}_{R=r} \overset{(a)}{=} \bar{G}^{\text{union}}.$$

On the other hand, by lemma 4.2, $G^{\text{ident}}_{R=r} \subset G^{\text{union}}$ and thus (c) $\bar{G}^{\text{ident}}_{R=r} \subset \bar{G}^{\text{union}}$, so that

$$\cup_r \bar{G}^{\text{detect}}_{R=r} \overset{(d)}{\subset} \cup_r \bar{G}^{\text{ident}}_{R=r} \overset{(b)}{\subset} \bar{G}^{\text{union}}.$$

□

F.2 Interring the Transfer-Graph

Lemma 5.2. *If $R \notin \mathrm{Anc}^{\text{union}}(Y)$, then $\mathrm{Pa}^{\text{phys}}_{R=r}(Y) = \mathrm{Pa}^{\text{union}}(Y)$, i. e. the change is non-physical (by observational non-accessibility).*

Proof. This follows directly from lemma 3.13. □

Cor. 5.3. *If $R \notin \mathrm{Anc}^{\text{union}}_{\text{detect}}(Y)$, then $\mathrm{Pa}^{\text{phys}}_{R=r}(Y) = \mathrm{Pa}^{\text{union}}_{\text{detect}}(Y)$.*

Proof. This follows directly from lemma 5.2 and rmk. 4.4 (see also lemma D.12). □

If R (or conditioning on R) does not change the distribution of ancestors, no state-induced effects occur:

Lemma 5.4. *Assuming strong regime-acyclicity. If $X \in \mathrm{Pa}^{\text{union}}(Y) - \mathrm{Pa}^{\text{ident}}_{R=r}(Y)$ and $R \in \mathrm{Pa}^{\text{union}}(Y)$, and $\mathrm{Anc}^{\text{union}}(R) \cap \mathrm{Anc}^{\text{union}}(\mathrm{Pa}^{\text{union}}(Y) - \{R\}) = \emptyset$, then $X \notin \mathrm{Pa}^{\text{phys}}(Y)$ (i. e. the change is "physical" not just by state).*

Proof. By lemma D.2, the noise-terms of nodes in $\mathrm{Anc}^{\text{union}}(Y)$ are unchanged by conditioning on R i. e. $P(\eta_{\mathrm{Anc}^{\text{union}}(Y)}|R) = P(\eta_{\mathrm{Anc}^{\text{union}}(Y)})$ and by corollary C.5a applied to $R \neq W \in \mathrm{Pa}^{\text{union}}(Y)$ shows $W = F_W(\eta_{\mathrm{Anc}^{\text{union}}(W)})$, with $\mathrm{Anc}^{\text{union}}(W) \subset \mathrm{Anc}^{\text{union}}(Y)$ thus $P(X_{\mathrm{Pa}^{\text{union}}(Y)-\{R\}}|R) = P(X_{\mathrm{Pa}^{\text{union}}(Y)-\{R\}})$. Therefore the support on parents did not change and the change must be physical. □

Cor. 5.5. *Assuming strong regime-acyclicity. If $R \neq X \in \mathrm{Pa}^{\text{union}}_{\text{detect}}(Y) - \mathrm{Pa}^{\text{ident}}_{R=r}(Y)$ and $R \in \mathrm{Pa}^{\text{union}}_{\text{detect}}(Y)$, and $\mathrm{Anc}^{\text{union}}_{\text{detect}}(R) \cap \mathrm{Anc}^{\text{union}}_{\text{detect}}(\mathrm{Pa}^{\text{union}}_{\text{detect}}(Y) - \{R\}) = \emptyset$, then*

(a) *there is a link into the strongly connected component of Y that vanishes in G^{phys}, but not in $G^{\text{union}}_{\text{detect}}$, i. e. there is a physical change.*

(b) *if Y is not part of a directed union-cycle, then $X \notin \mathrm{Pa}^{\text{phys}}(Y)$, i. e. there is a physical change of this particular link.*

Proof. Excluding R, $X \in \mathrm{Pa}^{\text{union}}_{\text{detect}}(Y) \Rightarrow X \in \mathrm{Pa}^{\text{union}}(Y)$. Similarly both $\mathrm{Anc}^{\text{union}}_{\text{detect}}(R)$ and $\mathrm{Anc}^{\text{union}}_{\text{detect}}(\mathrm{Pa}^{\text{union}}_{\text{detect}}(Y) - \{R\})$ exclude R, so we can replace them by $\mathrm{Anc}^{\text{union}}$. Since $G^{\text{union}} \subset G^{\text{union}}_{\text{detect}}$, also $R \in \mathrm{Pa}^{\text{union}}_{\text{detect}}(Y) \Rightarrow R \in \mathrm{Pa}^{\text{union}}(Y)$.

Thus the lemma applies. the vanishing link starts at $X \neq R$ (thus is away from R) and ends at an element of the strongly-connected component of Y. If Y is not part of a directed cycle, the strongly-connected component of Y is simply $\{Y\}$, and there is only a unique choice. □

F.3 Validity of Transfer

One can also use a transfer-argument to construct a test which deletes edges from the union-graph only if there is evidence that the mechanism did in fact change. See also §B.3.

Fix dependency measure d and estimator $\hat{d}$. Assume, using $\hat{d}$ (and some null-distribution and p-value threshold), we found a link $X \to Y$ with identifiable (e. g. by adjusting for Z) controlled direct effect of X on Y and such that this link vanishes in one context r_0. We want to distinguish between:

- The nullhypothesis: The change in $P(X)$ suffices to explain the failure to reject independence on finite-data.
- The alternative: The mechanism (or the noise on Y) have changed.

On the $\hat{d}$-dependent context, learn an estimator $\hat{P}_X$ of $P(X, Z|R = r_0)$ and $\hat{P}_{Y|X}$ of $P(Y|X, Z)$ (i. e. of the kernel $x \mapsto f_Y(x, -)_* \eta_Y$ containing the observable information about f_Y and η_Y)[4] by some conditional-density learning method. For a total of K datasets of size N each, draw $((x_1, z_1), \ldots, (x_N, z_N))$ from $\hat{P}_X$, then draw y_i from $\hat{P}_{Y|X}(Y|X = x_i, Z = z_i)$. On these datasets, generate dependence-measures (or test for independence) using $\hat{d}$ leading to a distribution $\hat{P}_d$. If the result for $\hat{d}$ on the original data in the $\hat{d}$-independent regime is plausible under $\hat{P}_d$ (or the test results on the K many datasets are $1 - \alpha$ often "independent"), then the changed support of X is sufficient to explain the "independence" (or rather the failure of $\hat{d}$ to detect any dependence) in this regime – assuming $\hat{P}_{Y|X}$ approximates the true $P(Y|X, Z)$ sufficiently well (see below). Otherwise we can reject the null-hypothesis that the change in support of X alone could explain the absence of this link.

The reliance on sufficiently fast convergence of $\hat{P}_{Y|X}$ is conceptually similar to the convergence of regressors in conditional independence testing with regressing out. I. e. when using a parametric model, for evaluating p-values, one has to take into account the additional number of degrees of freedom, for non-parametric models, e. g. bootstrapping approaches could be used. We acknowledge that this is in practice a very difficult problem. We leave it to future research, our present intent is to illustrate, that this seems – in principle – to also be a testable hypothesis.

F.4 Limiting (Extreme) Cases

The following "extreme" case is formally trivial, but provides some insights:

Example F.1. Given $P(R = r_0) = 1$ (which we typically exclude by the way we define regime-indicators, but which we can think of as a limiting case in practice), we observe: $P(\ldots|R = r_0) = P(\ldots)$, so also the supports agree and $G^{\text{union}} = G^{\text{descr}}_{R=r_0} = G^{\text{phys}}_{R=r_0}$. I. e., in this case our results collapse to the standard results for G^{union}.

From the perspective that, for the single-context case, the question about what is happening outside the support should probably be considered purely philosophical, this is a good sign: If our objects capture empirically accessible information, then they should not make claims about the single-context case.

[4] Under the null-hypothesis, learning g from the pooled data is ok, so even though in the alternative hypothesis g changes, for rejecting the null, learning g from the pooled data is fine, even though learning from a single or all other contexts might improve power.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The example from the abstract is explicitly picked up in the discussion of the (similar, but simpler) example 1.1 and in the conclusion §6. Observations (a-c) from the introduction are discussed throughout 3. The advertised identifiability results are in §4 and §5, the potential applications to anomalies and extreme events are discussed in §3.3. The mathematical framework for relating the presented graphical object to CSI is detailed in §D, and outlined in §4.1. Our graphical objects are defined in §3, as we note in the introduction, these extend the observations of [1], with details given transparently in §A.2.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: While there is no separate limitations section, we discuss assumptions made and applicability in many places in the main text, detailed discussions are also given e. g. in e. g. in §D.4, §B.3 or §E. Also the counterexample D.13 (referenced and explained in §4.1 in the main text) illustrates some fundamental limitations of the approach in general.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Those proofs central to the paper are outlined in the main text, there is at least reference to the appendix. Full proofs are in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: –

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: –

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: –

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: –

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: –

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

Justification: The presented work is not immediately applicable in a "plug-and-play" fashion, as it contains only theoretical ideas. While any idea can be abused, the contributions here are mostly towards explainability and human understandable results, so it more likely contributes to a saver and more transparent future ML. We do not discuss societal impact explicitly, because such a discussion without a particular use-case does not seem to contribute much benefit here.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: cf. "Broader Impacts"

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: –

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: –

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: –

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: –

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.