Necessary and sufficient graphical conditions for optimal adjustment sets in causal graphical models with hidden variables

Jakob Runge
German Aerospace Center
Institute of Data Science
07745 Jena, Germany
and
Technische Universität Berlin
10623 Berlin, Germany
jakob.runge@dlr.de

Abstract

The problem of selecting optimal backdoor adjustment sets to estimate causal effects in graphical models with hidden and conditioned variables is addressed. Previous work has defined optimality as achieving the smallest asymptotic estimation variance and derived an optimal set for the case without hidden variables. For the case with hidden variables there can be settings where no optimal set exists and currently only a sufficient graphical optimality criterion of limited applicability has been derived. In the present work optimality is characterized as maximizing a certain adjustment information which allows to derive a necessary and sufficient graphical criterion for the existence of an optimal adjustment set and a definition and algorithm to construct it. Further, the optimal set is valid if and only if a valid adjustment set exists and has higher (or equal) adjustment information than the Adjust-set proposed in Perković et al. [Journal of Machine Learning Research, 18: 1–62, 2018] for any graph. The results translate to minimal asymptotic estimation variance for a class of estimators whose asymptotic variance follows a certain information-theoretic relation. Numerical experiments indicate that the asymptotic results also hold for relatively small sample sizes and that the optimal adjustment set or minimized variants thereof often yield better variance also beyond that estimator class. Surprisingly, among the randomly created setups more than 90% fulfill the optimality conditions indicating that also in many real-world scenarios graphical optimality may hold.

1 Introduction

A standard problem setting in causal inference is to estimate the causal effect between two variables given a causal graphical model that specifies qualitative causal relations among observed variables [Pearl, 2009], including a possible presence of hidden confounding variables. The graphical model then allows to employ graphical criteria to identify valid adjustment sets, the most well-known being the backdoor criterion [Pearl, 1993] and the generalized adjustment criterion [Shpitser et al., 2010, Perković et al., 2015, 2018], providing a complete identification of all valid adjustment sets. Estimators of causal effects based on such a valid adjustment set as a covariate are then unbiased, but for different adjustment sets the estimation variance may strongly vary. An optimal adjustment set may be characterized as one that has minimal asymptotic estimation variance. In current work, following Kuroki and Cai [2004] and Kuroki and Miyakawa [2003], Henckel et al. [2019] (abbreviated

We consider causal effects in causal graphical models over a set of variables $\mathbf{Z}$ for $\text{forb}(\text{henceforth just denoted as valid})$ if and only if a valid adjustment set exists and has higher (or equal) adjustment information than manually checked directly from the graph, but, more conveniently, Perković et al. [2018] define an minimal is called generalizes Pearl’s back-door criterion [Pearl, 2009]. To this end define a given causal graph using the generalized adjustment criterion [Perković et al., 2015, 2018] which factorizes as $p_{\text{valid}}$ set notation and denote unions of variables as preliminaries, including the definition of open and blocked paths, see Supplementary Section A. In nodes on causal paths from $\mathbf{X}$ des for " Fig. 1A for an example. Kinships are defined as usual: parents $\mathbf{pa}$ nodes can have possibly more than one edge which can be directed $P = \mathcal{P}(\mathbf{V})$ that is consistent with an acyclic directed mixed graph (ADMG) $\mathcal{G} = (\mathbf{V}, \mathcal{E})$. Two nodes can have possibly more than one edge which can be directed ($\leftarrow$) or bi-directed ($\leftrightarrow$). See Fig. 1A for an example. Kinships are defined as usual: parents $\mathbf{pa}$ for "$\rightarrow\mathbf{X}$", spouses $\mathbf{sp}(\mathbf{X})$ for $\mathbf{X} \rightarrow \mathbf{X}$", children $\mathbf{ch}(\mathbf{X})$ for "$\rightarrow \mathbf{X}$". These sets all exclude $\mathbf{X}$. Correspondingly descendants $\text{des}(\mathbf{X})$ and ancestors $\text{an}(\mathbf{X})$ are defined, which, on the other hand, both include $\mathbf{X}$. The mediator nodes on causal paths from $\mathbf{X}$ to $\mathbf{Y}$ are denoted $\mathbf{M} = \mathbf{M}(\mathbf{X}, \mathbf{Y})$ and exclude $\mathbf{X}$ and $\mathbf{Y}$. For detailed preliminaries, including the definition of open and blocked paths, see Supplementary Section A. In this work we only consider a univariate intervention variable $\mathbf{X}$ and effect variable $\mathbf{Y}$. We simplify set notation and denote unions of variables as $\{ \mathbf{W} \} \cup \mathbf{M} \cup \mathbf{A} = \mathbf{WMA}$. A (possibly empty) set of adjustment variables $\mathbf{Z}$ for the total causal effect of $\mathbf{X}$ on $\mathbf{Y}$ in an ADMG is called valid relative to ($\mathbf{X}, \mathbf{Y}$) if the interventional distribution for setting $\text{do}(\mathbf{X} = \mathbf{x})$ [Pearl, 2009] factorizes as $p(\mathbf{Y}|\text{do}(\mathbf{X} = \mathbf{x})) = \int p(\mathbf{Y}|\mathbf{x}, \mathbf{z})p(\mathbf{z})d\mathbf{z}$ for non-empty $\mathbf{Z}$ and as $p(\mathbf{Y}|\text{do}(\mathbf{X} = \mathbf{x})) = p(\mathbf{Y}|\mathbf{x})$ for empty $\mathbf{Z}$. Valid adjustment sets, the set of which is here denoted $\mathbf{Z}$, can be read off from a given causal graph using the generalized adjustment criterion [Perković et al., 2015, 2018] which generalizes Pearl’s back-door criterion [Pearl, 2009]. To this end define

$$\text{forb}(\mathbf{X}, \mathbf{Y}) = \mathbf{X} \cup \text{des}(\mathbf{Y}\mathbf{M})$$  \hspace{1cm} (1)$$

(henceforth just denoted as forb). A set $\mathbf{Z}$ is valid if both of the following conditions hold: (i) $\mathbf{Z} \cap \text{forb} = \emptyset$, and (ii) all non-causal paths from $\mathbf{X}$ to $\mathbf{Y}$ are blocked by $\mathbf{Z}$. An adjustment set is called minimal if no strict subset of $\mathbf{Z}$ is still valid. The validity conditions can in principle be manually checked directly from the graph, but, more conveniently, Perković et al. [2018] define an

My main theoretical contribution is a solution to this problem. Optimality for conditional causal effects in the hidden variables case is fully characterized by an information-theoretic approach involving a certain difference of conditional mutual informations among the observed variables termed the adjustment information. Maximizing the adjustment information formalizes the common intuition to choose adjustment sets that maximally constrain the effect variable and minimally constrain the cause variable. This allows to derive a necessary and sufficient graphical criterion for the existence of an optimal adjustment set. The derived optimal adjustment set also has the property of minimum cardinality, i.e., no node can be removed without sacrificing optimality. Further, the optimal set is valid if and only if a valid adjustment set exists and has higher (or equal) adjustment information than the Adjust-set proposed in Perković et al. [2018] for any graph, whether graphical optimality holds or not. The results translate to minimal asymptotic estimation variance for a class of estimators whose asymptotic variance follows a certain information-theoretic relation that, at present, I could only verify theoretically for the linear case. As practical contributions the paper provides extensive numerical experiments that corroborate the theoretical results and show that the optimal adjustment set or minimized variants thereof often yield better variance also beyond the theoretically analyzed estimator class. Code is available in the python package [https://github.com/jakobrunge/tigramite](https://github.com/jakobrunge/tigramite). More detailed preliminaries, proofs, algorithms, and further numerical experiments are given in the Supplementary Material.

1.1 Preliminaries and problem setting

We consider causal effects in causal graphical models over a set of variables $\mathbf{V}$ with a joint distribution $p = \mathcal{P}(\mathbf{V})$ that is consistent with an acyclic directed mixed graph (ADMG) $\mathcal{G} = (\mathbf{V}, \mathcal{E})$. Two nodes can have possibly more than one edge which can be directed ($\leftarrow$) or bi-directed ($\leftrightarrow$).
Throughout the present paper we will assume the following. We assume a causal graphical model over a set of variables $V$ with a joint distribution $P = \mathcal{P}(V)$ that is consistent with an ADMG $\mathcal{G} = (V, E)$. We assume a non-zero causal effect from $X$ on $Y$, potentially through a set of mediators $M$, and given selected conditioned variables $S$, where $S \cap \text{forb} = \emptyset$. We assume that at least one valid adjustment set (given $S$) exists and, hence, the causal effect is identifiable (except when stated otherwise). Finally, we assume the usual Causal Markov Condition (implicit in semi-Markovian models) and Faithfulness.
2 Optimal adjustment sets

2.1 Information-theoretic characterization

Figure 1B illustrates two causal effect estimates for a linear Gaussian model consistent with the graph in Fig. 1A. With \( Z = Z_1Z_2 \) (blue) the error is much larger than with \( O = Z_2Z_3 \) (orange) for two reasons: \( Z \) constrains the residual variance \( \text{Var}(Y|Z) \) of the effect variable \( Y \) less than \( O \) and, on the other hand, \( Z \) constrains the residual variance \( \text{Var}(X|Z) \) of the cause variable \( X \) more than \( O \). Smaller estimator variance also holds for \( O \) compared to any other valid set in \( Z \) here.

We information-theoretically formalize the resulting intuition to choose an adjustment set \( Z \) that maximally constrains the effect variable \( Y \) and minimally constrains the cause variable \( X \). In terms of CMIs and given selected fixed conditions \( S \) the quantity to maximize can be stated as follows.

**Definition 1** (Adjustment information). Consider a causal effect of \( X \) on \( Y \) for an adjustment set \( Z \) given a condition set \( S \). The (conditional) adjustment (set) information, abbreviated \( J_Z \), is defined as

\[
J_{X|Y|S;Z} = I_{Z;Y|X,S} - I_{X;Z|S} = J_{X|Y|S} - H_{X|S} - (H_{Y|X,Z} - H_{X|Z})
\]

\( J_Z \) is not necessarily positive if the dependence between \( X \) and \( Z \) (given \( S \)) is larger than that between \( Z \) and \( Y \) given \( XS \). Equation (6) follows from the CMI definition. Fig. 1C illustrates the two CMIs in Eq. (5) in a Venn diagram.

Before discussing the range of estimators for which maximizing the adjustment information \( J_Z \) leads to a minimal asymptotic estimation variance in Sect. 2.2 we characterize graphical optimality in an information-theoretic framework. Our goal is to provide graphical criteria for optimal adjustment sets, i.e., criteria that depend only on the structure of the graph \( G \) and not on the distribution.

**Definition 2** (Information-theoretical graphical optimality). Given Assumptions 1 we say that (information-theoretical) graphical optimality holds if there is a \( Z \in Z \) such that either there is no other \( Z' \neq Z \in Z \) or for all other \( Z' \neq Z \in Z \) and all distributions \( P \) consistent with \( G \) we have \( J_{Z} \geq J_{Z'} \).

My main result builds on the following lemma which relates graphical optimality to information-theoretic inequalities in a necessary and sufficient comparison condition for an optimal set to exist.

**Lemma 1** (Necessary and sufficient comparison criterion for existence of an optimal set). Given Assumptions 1, if and only if there is a \( Z \in Z \) such that either there is no other \( Z' \neq Z \in Z \) or for all other \( Z' \neq Z \in Z \) and all distributions \( P \) consistent with \( G \) it holds that

\[
I_{Z;Y|Z,X,S} \geq I_{Z;Y|Z,X,S} \quad \text{(i)}
\]

\[
I_{X;Z|Z,X,Z,S} \geq I_{X;Z|Z(X,S)} \quad \text{(v)}
\]

then graphical optimality holds and \( Z \) is optimal implying \( J_Z \geq J_{Z'} \).

In SSR20 and HPM19 the corresponding conditional independence statements to the terms (iii) and (iv) in the inequalities (7) are used as a sufficient pairwise comparison criterion. However, Lemma 1 shows that for graphical optimality it is not necessary that terms (iii) and (iv) vanish, they just need to fulfill the inequalities (7) for a necessary and sufficient criterion.

In principle, Lemma 1 can be used to cross-compare all pairs of sets, but firstly, it is difficult to explicitly evaluate (7) for all distributions \( P \) consistent with \( G \) and, secondly, iterating through all valid adjustment sets is computationally prohibitive even for small graph sizes. As an example, consider a confounding path consisting of 5 nodes. Then this path can be blocked by \( 2^5 - 1 \) different subsets. In the main result of this work (Thm. 3) a necessary and sufficient criterion based purely on graphical properties is given.

2.2 Applicable estimator class

The above characterization only relates optimality of adjustment sets to the adjustment information \( J_Z \) defined in Eq. (5), but not to any particular estimator. Now the question is for which class of
causal effect estimators $\hat{\Delta}_{y|x'|s,z}$, the intuition of maximizing the adjustment information $J_Z$ leads to a minimal asymptotic estimation variance. In its most general form this class is characterized as fulfilling

$$Z_{\text{optimal}} \in \arg\max_{\mathbf{Z} \in \mathcal{Z}} J_Z \iff \text{Var}(\hat{\Delta}_{y|x'|s,z_{\text{optimal}}}) = \min_{\mathbf{Z} \in \mathcal{Z}} \text{Var}(\hat{\Delta}_{y|x'|s,z}), \quad (8)$$

where we assume that $\hat{\Delta}_{y|x'|s,z}$ is consistent due to a valid adjustment set and correct functional model specification. One can also further restrict the class to estimators whose (square-root of the) asymptotic variance can be expressed as

$$\sqrt{\text{Var}(\hat{\Delta}_{y|x'|s,z})} = f(H_{Y|XZS} - H_{X|ZS}), \quad (9)$$

for a real-valued, strictly monotonously increasing function of the adjustment entropy. Minimizing the adjustment entropy is by Eq. (6) equivalent to maximizing the adjustment information. The following assumption and lemma then relates $J_Z \geq J_{Z'}$ to the corresponding asymptotic variances of a given estimator.

**Assumptions 2** (Estimator class assumption). The model class of the estimator for the causal effect $\hat{\Delta}_{y|x'|s,z}$ is correctly specified and its asymptotic variance can be expressed as in relation (9).

**Lemma 2** (Asymptotic variance and adjustment information). Given Assumptions 2 and an estimator fulfilling Assumptions 2 if and only if for two different adjustment sets $\mathbf{Z}, \mathbf{Z}' \in \mathcal{Z}$ we have $J_Z \geq J_{Z'}$, then the adjustment set $\mathbf{Z}$ has a smaller or equal asymptotic variance compared to $\mathbf{Z}'$.

**Proof.** By Equations (6) and (9) $J_Z \geq J_{Z'}$ (for fixed $X, Y, S$) is directly related to a smaller or equal asymptotic variance for $\mathbf{Z}$ compared to $\mathbf{Z}'$, and vice versa. \qed

The paper’s theoretical results currently hold for estimators fulfilling relation (9), but at least the main result on graphical optimality in Thm. 3 can also be relaxed to estimators fulfilling the less restrictive relation (8). In this work, we leave the question of which general classes of estimators fulfill either relation (8) or the more restricted relation (9) to further research and only show that it holds for the OLS estimator $\hat{\beta}_{Y|XZS}$ for Gaussian distributions.

For Gaussians the entropies in (9) are given by $H(Y|XZS) = 1/2 + 1/2 \ln(2\pi\sigma^2_{y|x|z})$ and $H(X|ZS) = 1/2 + 1/2 \ln(2\pi\sigma^2_{x|z})$ where $\sigma(\cdot|\cdot)$ denotes the square-root of the conditional variance. Then

$$\sqrt{\text{Var}(\hat{\Delta}_{y|x'|s,z})} = \frac{1}{\sqrt{n}} e^{H_{Y|XZS} - H_{X|ZS}} = \frac{1}{\sqrt{n}} \frac{\sigma_{Y|XZS}}{\sigma_{X|ZS}}. \quad (10)$$

This relation is also the basis of the results for the causally sufficient case in [Henckel et al. 2019] where it is shown that it holds more generally for causal linear models that do not require the noise terms to be Gaussian.

### 2.3 Definition of O-set

The optimal adjustment set for the causally sufficient case is simply $\mathbf{P} = pa(YM) \setminus \text{forb}$ and was derived in HPM19 and Romitzyk and Smucler [2019]. In Section 3.2 the derivation is discussed from an information-theoretic perspective. In the case with hidden variables we need to account for bidirected edges “$\leftrightarrow$” which considerably complicate the situation. Then the parents of $YM$ are not sufficient to block all non-causal paths. Further, just like conditioning on parents of $YM$ leads to optimality in the sufficient case since parents constrain information in $YM$, in the hidden variables case also conditioning on spouses of $YM$ constrains information about $YM$.

**Example A.** A simple graph (ADMG) to illustrate this is $X \rightarrow Y \leftrightarrow Z_1$ (shown with an additional $S$ in Fig. 2A below, or Fig. 4 in SSR20). Here $Z' = \emptyset = \text{vancs}$ is a valid set, but it is not optimal. Consider $O = Z_1$, then term (iii) $= 0$ in the inequalities (7) since $Z' \setminus O = \emptyset$. Even though not needed to block non-causal paths (there is none), $Z_1$ still constrains information in $Y$ while being independent of $X$ (hence, term (iv) $= 0$) which leads to $J_O > J_0$ according to the inequalities (7).

Not only direct spouses can constrain information in $Y$ as Fig. 2B below illustrates. Since for $W \in YM$ the motif “$W \leftrightarrow C_1 \rightarrow C_2 \leftrightarrow C_3$” (“$\leftrightarrow$” denotes either edge mark) is open, it holds that $I(C_1; Y) + I(C_2; W|C_1) \geq I(C_1; Y)$ and we can even further increase the first term in the adjustment
We now move to the question of optimality. It is known that there are graphs where no graphical
adjustment set exists, then
\[ O \subseteq \text{vancs} \]
which can be implemented with (bi-directional) breadth-first search as proposed in van der Zander
et al. [2019].
Numerical experiments in Section 3 will show that further interesting adjustment sets are the
valid collider paths. This leads to the notion of a valid collider path (related to the

Definition 3 (Valid collider paths). Given a graph \( G \), a collider path of \( W \) for \( k \geq 1 \) is defined by
a sequence of edges \( W \leftarrow C_1 \leftarrow \cdots \leftarrow C_k \). We denote the set of path nodes (excluding \( W \)) along
a path indexed by \( i \) as \( \pi^i_W \). Using the set of valid ancestors \( \text{vancs} = \text{an}(XY|S) \setminus \text{forb} \)
for the causal effect of \( X \) on \( Y \) given \( S \) we call a collider path node set \( \pi^i_W \) for \( W \in YM \) valid wrt.
\( (X,Y,S) \) if for each path node \( C \in \pi^i_W \) both of the following conditions are fulfilled:

\[ (1) \ C \notin \text{forb}, \text{ and } (2a) \ C \in \text{vancs} \text{ or } (2b) \ C \perp \perp X \mid \text{vancs}. \tag{11} \]

Condition (1) is required for any valid adjustment set. If jointly (2a) and (2b) are not fulfilled, i.e.
\( C \notin \text{vancs} \) and \( C \not\perp \perp X \mid \text{vancs} \), then the collider path stops before \( C \). Our candidate optimal
adjustment set is now constructed based on the parents of \( Y \), valid collider path nodes of \( Y \),
and their parents to ‘close’ these collider paths.

Definition 4 (O-set). Given Assumptions [1] and the definition of valid colliders in Def. 3, define the
set \( O(X,Y,S) = P \cup C \cup P_C \) where
\[ P = \text{pa}(YM) \setminus \text{forb}, \quad C = \{ \pi^i_W : \pi^i_W \text{ is valid wrt. } (X,Y,S) \}, \quad P_C = \text{pa}(C). \]

In the following we will abbreviate \( O = O(X,Y,S) \). Algorithm C.1 states efficient pseudo-code to
construct the O-set and detect whether a valid adjustment set exists. Since none of the conditions
of Def. 3 for adding collider nodes depends on previously added nodes, the algorithm is order-
independent. The statement occurring in lines 11 and 21 (“No valid adjustment set exists.”) is proven
in Thm 1. If the graph is a DAG, then lines 4-22 can be omitted. The algorithm is of low complexity
independent. The statement occurring in lines 11 and 21 (“No valid adjustment set exists.”) is proven
in Thm 1. If the graph is a DAG, then lines 4-22 can be omitted. The algorithm is of low complexity
and the most time-consuming part is checking for a path in line 12, Def. 3(2b)

Definition 3 for adding collider nodes depends on previously added nodes, the algorithm is order-
independent. The statement occurring in lines 11 and 21 (“No valid adjustment set exists.”) is proven
in Thm 1. If the graph is a DAG, then lines 4-22 can be omitted. The algorithm is of low complexity
and the most time-consuming part is checking for a path in line 12, Def. 3(2b)

\[ (Valid collider paths) \]

For each path node \( X \), \( i \)

\[ \forall X \in \pi^i_W \text{ for each path node } C, \text{ if } C \in \pi^i_W \forall X \in \pi^i_W \text{ both of the following conditions are fulfilled:} \]

\[ (1) C \notin \text{forb}, \text{ and } (2a) C \in \text{vancs} \text{ or } (2b) C \perp \perp X \mid \text{vancs}. \tag{11} \]

Condition (1) is required for any valid adjustment set. If jointly (2a) and (2b) are not fulfilled, i.e.
\( C \notin \text{vancs} \) and \( C \not\perp \perp X \mid \text{vancs} \), then the collider path stops before \( C \). Our candidate optimal
adjustment set is now constructed based on the parents of \( Y \), valid collider path nodes of \( Y \),
and their parents to ‘close’ these collider paths.

Definition 4 (O-set). Given Assumptions [1] and the definition of valid colliders in Def. 3, define the
set \( O(X,Y,S) = P \cup C \cup P_C \) where
\[ P = \text{pa}(YM) \setminus \text{forb}, \quad C = \{ \pi^i_W : \pi^i_W \text{ is valid wrt. } (X,Y,S) \}, \quad P_C = \text{pa}(C). \]

In the following we will abbreviate \( O = O(X,Y,S) \). Algorithm C.1 states efficient pseudo-code to
construct the O-set and detect whether a valid adjustment set exists. Since none of the conditions
of Def. 3 for adding collider nodes depends on previously added nodes, the algorithm is order-
independent. The statement occurring in lines 11 and 21 (“No valid adjustment set exists.”) is proven
in Thm 1. If the graph is a DAG, then lines 4-22 can be omitted. The algorithm is of low complexity
and the most time-consuming part is checking for a path in line 12, Def. 3(2b) \( C \perp \perp X \mid \text{vancs}, \)
which can be implemented with (bi-directional) breadth-first search as proposed in van der Zander
et al. [2019].

Numerical experiments in Section 3 will show that further interesting adjustment sets are the
minimized O-set \( O_{\min} \), where \( O \) is minimized such that no subset can be removed without making
\( O_{\min} \) invalid, and the collider-minimized O-set \( O_{\min} \) where only \( CP_C \setminus P \subseteq O \) is minimized
such that no collider-subset can be removed without making \( O_{\min} \) invalid. Both adjustment sets

Validity of O-set

Given Assumptions [1] but without any priori assuming that a valid adjustment set exists (apart from the requirement \( S \cap \text{forb} = \emptyset \)). If and only if a valid backdoor adjustment set exists, then \( O \) is a valid adjustment set.

2.4 Graphical optimality

We now move to the question of optimality. It is known that there are graphs where no graphical
criterion exists to determine optimality. Examples, discussed later, are the graphs in Figs. 2-6.

Before stating necessary and sufficient conditions for graphical optimality, I mention that next to
the O-set defined above and the Adjust set \( \text{vancs} \) [Perković et al. 2018], I am not aware of any
other systematically constructed set that will yield a valid adjustment set for the case with hidden
variables. van der Zander et al. [2019] provide algorithms to list all valid adjustment sets, but the
question is which of these a user should choose. As mentioned above, Lemma [1] can be used to
cross-compare all pairs of sets, but this is not really feasible. Hence, for automated causal effect
estimation, rather than the question of whether graphical optimality holds, it is crucial to have a set
with better properties than other systematically constructable sets. The following theorem states that
the adjustment informations follow $J_\mathcal{O} \geq J_{\text{vancs}}$ for any graph (whether graphical optimality holds or not).

**Theorem 2** (O-set vs. Adjust-set). Given Assumptions 7 with $\mathcal{O}$ defined in Def. 4 and the Adjust-set defined in Eq. (2), it holds that $J_\mathcal{O} \geq J_{\text{vancs}}$ for any graph $\mathcal{G}$. We have $J_\mathcal{O} = J_{\text{vancs}}$ only if (1) $\mathcal{O} = \text{vancs}$, or (2) $\mathcal{O} \subseteq \text{vancs}$ and $X \perp \text{vancs} \setminus \mathcal{O} | \mathcal{O}$.

In the following the $\mathcal{O}$-set is illustrated and conditions for graphical optimality are explored. SSR20 provide a sufficient condition for optimality, which states that either all nodes are observed (no bidirected edges exist) or for all observed nodes $V \subseteq \text{vancs}$. This is a very strict assumption and not fulfilled for any of the examples (except for Example G) discussed in the following.

**Example B.** Figure 2B depicts a larger example to illustrate the $\mathcal{O}$-set $\mathcal{O} = \text{PCP}_C$ with $\mathcal{P} = Z_1Z_2Z_3Z_4$ (blue boxes) and $\text{CP}_C \setminus \mathcal{P} = Z_5Z_6Z_7Z_8$ (green boxes). We also have a conditioned variable $\mathcal{S}$. Among $\mathcal{P}$, only $Z_1Z_2$ are needed to block non-causal paths to $X$, $Z_3Z_4$ are only there to constrain information in $Y$. Here the same holds for the whole set $\text{CP}_C \setminus \mathcal{P}$ which was constructed from the paths $Y \leftrightarrow Z_5 \leftrightarrow Z_6 \leftrightarrow Z_7$ and $Y \leftrightarrow Z_5 \leftrightarrow Z_8$ which does not include $Z_9$ since it is a descendant of $YM$. Including an independent variable like $Z_{12}$ in $\mathcal{O}$ would not decrease the adjustment information $J_\mathcal{O}$, but then $\mathcal{O}$ would not be of minimum cardinality anymore (proven in Cor. B.3). Here, again, the condition of SSR20 does not hold (e.g., $Z_5$ is not an ancestor of $XYS$). $\mathcal{O}$ is optimal here which can be seen as follows: For term (iii) in the inequalities (7) to even be non-zero, we would need a valid $\mathcal{Z}$ such that $\mathcal{Z} \setminus \mathcal{O}$ has a path to $Y$ given $OSX$. But these are all blocked. Note that while $Z_{10}$ or $Z_9 \in \mathcal{Z}$ would open a path to $Y$, both of these are descendants of $M$ or $Y$ and, hence, cannot be in a valid $\mathcal{Z}$. For term (iv) to even be non-zero $\mathcal{O} \setminus \mathcal{Z}$ would need to have a path to $X$ given $Z\mathcal{S}$. But since any valid $\mathcal{Z}$ has to contain $Z_1$ and $Z_2$ (or $Z_{11}$), the only nodes in $\mathcal{O}$ with a path to $X$ are parents of $YM$ and paths from these parents to $X$ all need to be blocked for a valid $\mathcal{Z}$. Hence, $\mathcal{O}$ is optimal here.

**Example C.** In Fig. 2C a case is shown where $\mathcal{O} = Z_1Z_5$. $Z_2$ is not part of $\mathcal{O}$ because none of the conditions in Def. 4 are fulfilled: $Z_2 \notin \text{vancs} = Z_1Z_4Z_5$ and $Z_2 \not\perp X | \text{vancs}$. Hence, we call $Z_2$ an N-node. But $Z_2$ cannot be part of any valid $\mathcal{Z}$ because it has a collider path to $X$ through $Z_1$ which is always open because it is part of $\text{vancs}$. Hence, term (iii) is always zero. Term (iv) is zero because $\mathcal{O} \setminus \mathcal{Z}$ is empty for any valid $\mathcal{Z}$ here. Here even $J_\mathcal{O} > J_Z$ since $\mathcal{O}$ is minimal and term (ii) $I_{X,Z;O|O} > 0$ for any $\mathcal{Z} \neq \mathcal{O}$ (generally proven in Corollary B.1).

**Example D.** The example in Fig. 2D depicts a case with $\mathcal{O} = Z_2$ where $Z_1$ is an N-node. Next to $Z = \emptyset$ another valid set is $\mathcal{Z} = Z_1$. Then term (iii) is non-zero and in the same way term (iv) is non-zero. The sufficient pairwise comparison criterion in SSR20 and HPM19 is, hence, not applicable. However, it holds that always term (iii) $\leq$ (i) because the dependence between $Z_1$ and $Y$ given $X$ is
always smaller than the dependence between \(Z_2\) and \(Y\) given \(X\) and correspondingly term (iv) \(\leq\) (ii). Hence, \(O\) is optimal here. If a link \(Y \rightarrow Z_1\) exists, then the only other valid set is \(Z = \emptyset\) and both terms are strictly zero.

**Example E.** The example in Fig. 2E (Fig. 3 in SSR20 and also discussed in HPM19) is not graphically optimal. Here \(O = Z_1Z_2\). Other valid adjustment sets are \(Z_1\) or the empty set. From using \(Z_1 \perp \perp Y|X\) and \(X \perp \perp Z_2|Z_1\) in the inequalities \(7\) one can derive in information-theoretic terms that both \(Z_1Z_2\) and \(\emptyset\) are better than \(\text{van}cs = Z_1\), but since \(J_{Z_1Z_2} = J_\emptyset + I_{Z_2Y|XZ_1} - I_{X;Z_1}\), a superior adjustment set depends on how strong the link \(Z_1 \rightarrow X\) vs. \(Z_2 \rightarrow Y\) is. The graph stays non-optimal also with a link \(Z_1 \leftrightarrow Z_2\).

**Example F.** The example in Fig. 2F is also not graphically optimal. Here \(O = \emptyset\) and \(Z_2\) is an N-node with a non-collider path to \(X\). Other valid adjustment sets are \(Z_1\) and \(Z_1Z_2\). Higher adjustment information here depends on the distribution. Also the same graph with the link \(Z_1 \leftrightarrow X\) is non-optimal. If, however, there is another link \(Z_1 \rightarrow Y\), then \(O = \emptyset\) is optimal (then \(Z_1\) is a mediator).

**Example G.** The example in Fig. 2G is only a slight modification of Example E with an added selected condition \(S\). Then \(Z_1, Z_2 \in \text{van}cs\). We still get \(O = Z_1Z_2\) and this is now optimal since \(Z_2\) is always open and any valid set has to contain \(Z_1\).

The main result of this work is a set of necessary and sufficient conditions for the existence of graphical optimality and the proof of optimality of the \(O\)-set which is based on the intuition gained in the preceding examples.

**Theorem 3** (Necessary and sufficient graphical conditions for optimality and optimality of \(O\)-set). Given Assumptions \(7\) and with \(O = \text{PCP}_C\) defined in Def. \(4\). Denote the set of N-nodes by \(N = sp(YMC) \setminus \text{forbOS}\). Finally, given an \(N \in N\) and a collider path \(N \leftrightarrow \ldots \leftrightarrow C \leftrightarrow \ldots \leftrightarrow W\) (including \(N \rightarrow W\)) for \(C \in C\) and \(W \in YM\) (indexed by \(i\)) with the collider path nodes denoted by \(\pi^N_i\) (excluding \(N\) and \(W\)), denote by \(O_{\pi^N_i} = O(X, Y, S' = S \cap \pi^N_i)\) \(O\)-set for the causal effect of \(X\) on \(Y\) given \(S' = S \cup \{N\} \cup \pi^N_i\). If and only if exactly one valid adjustment set exists, or both of the following conditions are fulfilled, then graphical optimality holds and \(O\) is optimal:

(I) For all \(N \in N\) and all its collider paths \(i\) to \(W \in YM\) that are inside \(C\) it holds that \(O_{\pi^N_i}\) does not block all non-causal paths from \(X\) to \(Y\), i.e., \(O_{\pi^N_i}\) is non-valid,

and

(II) for all \(E \in O \setminus P\) with an open path to \(X\) given \(SO \setminus \{E\}\) there is a link \(E \leftrightarrow W\) or an extended collider path \(E \leftrightarrow C \leftrightarrow \ldots \leftrightarrow W\) inside \(C\) for \(W \in YM\) where all colliders \(C \in \text{van}cs\).

Condition (I) and (II) essentially rule out the two canonical cases in Examples F and E, respectively, on which non-optimality in any graph is based. Applied to the examples, we obtain that in Example F Cond. (I) holds since no N-node exists and Cond. (II) holds since \(X \perp \perp Z_1|S\). In Example B also no N-node exists and Cond. (II) holds as \(X \perp \perp E|SO \setminus \{E\}\) for every \(E \in O \setminus P\). In example C \(Z_2\) is an N-node, but there is a collider path to \(X\) through \(Z_1\) which is in \(\text{van}cs\) such that Cond. I is fulfilled. Further, while \(X \not \perp Z_5|SO \setminus \{Z_5\}\), there is a link \(Z_5 \rightarrow Y\) such that Cond. II holds. In example D \(Z_1\) is an N-node, but it has a bidirected link with \(X\) and Cond. (II) holds since \(X \perp \perp Z_2|SO \setminus \{Z_2\}\). In Example E optimality does not hold, but Cond. (I) actually holds since there is no N-node. Cond. (II) is not fulfilled for \(E = Z_1\), which has a path to \(X\) given \(O\) and on the extended collider path \(Z_1 \leftrightarrow Z_2 \leftrightarrow Y\) \(Z_2 \notin \text{van}cs\). For \(Z' = \emptyset\) and a distribution \(P'\) where the link \(Z_2 \leftrightarrow Y\) almost vanishes we then have \(J_O < J_{Z'}\). Example F has an N-node \(Z_2\) and \(O_{\pi^N_i} = O(X, Y, S' = Z_2) = Z_1Z_2\) is valid implying that Cond. (I) does not hold, while Cond. (II) is actually fulfilled with \(O = \emptyset\). For \(Z' = O_{\pi^N_i} = Z_1Z_2\) and a distribution \(P'\) where the link \(X \rightarrow Z_1\) almost vanishes we then have \(J_O < J_{Z'}\). Example G is optimal since there are no N-nodes and \(Z_2 \in \text{van}cs\).

Similar to SSR20, HPM19, and Witte et al. [2020], I also provide results regarding minimality and minimum cardinality for the hidden variables case in the Supplement.

### 3 Numerical experiments

We now investigate graphical optimality empirically to answer three questions: Firstly, whether for a linear estimator under Assumptions \(6\), the asymptotically optimal variance also translates into better
Secondly, how the O-set performs in non-optimal settings (according to Thm. [3]). Thirdly, how the O-set and variants thereof perform for estimators not captured by the class for which the theoretical results were derived (Assumptions [2]). To this end, we compare the performance of O, Adjust, O_{\text{Cmin}}, O_{\text{min}}, Adjust_{X_{\text{min}}}, and Adjust_{\text{min}} (see definitions in Section 2.3) together with linear least squares estimation (LinReg) on linear models. In the Supplement we also investigate nonlinear models using nearest neighbor regression (kNN), a multilayer perceptron (MLP), random forest regression, and double machine learning for partially linear regression models (DML) [Chernozhukov et al., 2018]. The experiments are based on a generalized additive model and described in detail in Section D. Among these 12,000 randomly created configurations 93% fulfill the optimality conditions in Thm. [3].

The results in Fig. [3] confirm our first hypothesis that for linear experiments with an estimator fulfilling Assumptions [2] and in settings where graphical optimality is fulfilled (Thm. [3]) the O-set either has similar RMSE or significantly outperforms all other tested variants. In particular, O_{\text{min}} and Adjust_{\text{min}} are bad choices for this setting. Adjust is intermediate and O_{\text{Cmin}} and Adjust_{X_{\text{min}}} come closest to O, but may still yield significantly higher variance.

Secondly, in non-optimal settings (only 7% of configurations) the O-set still outperforms Adjust (as expected by Thm. [3]). Compared to O_{\text{Cmin}} and Adjust_{X_{\text{min}}} the O-set leads to worse results for about half of the studied configurations, while O_{\text{min}} and Adjust_{\text{min}} are still bad choices. Cardinality is slightly higher for O compared to all other sets. In Fig. [S7] we further differentiate the results by the cardinality of the O-set and find that for small cardinalities (up to 4) the O-set has the lowest variance in a majority of cases, but for higher cardinalities either O_{\text{Cmin}} or O_{\text{min}} have the lowest variance (slightly beating Adjust_{X_{\text{min}}}). Hence, either O or O_{\text{Cmin}} performs best in non-optimal configurations. For very small sample sizes n = 30 (see Fig. [S2]) that become comparable to the adjustment set cardinality, there tends to be a trade-off and smaller cardinality helps. Then O_{\text{Cmin}} tends to be better than O for high cardinalities also in optimal settings, but here this effect is only present for n = 30 and for n = 50 already negligible compared to the gain in J_{\text{c}}. In Appendix D.2 are RMSE ratios for all combinations of adjustment approaches considered here and it is shown that, in general, results are very similar for other sample sizes.

Thirdly, we investigate non-parametric estimators on linear as well as nonlinear models (implementations described in Section D) results in the figures of Section D.3. The different classes of estimators exhibit quite different behavior. For kNN (Figs. [S8],[S9]) the O-set has the lowest variance in around 50% of the configurations followed by O_{\text{Cmin}} and O_{\text{min}}. More specifically (Figs. [S15],[S16]), for small O-set cardinalities up to 2 the O-set and for higher either O_{\text{min}} or O_{\text{Cmin}} (the latter only in non-optimal configurations) perform best. For nonlinear experiments the results are less clear for O-set cardinalities greater than 2, but O_{\text{min}} is still a good choice. Regarding RMSE ratios, we see that, for the cases where O is not the best, the O-set can have considerably higher variance, while O_{\text{min}} seems to be most robust and may be a better choice if O is too large. MLP (Figs. [S10],[S11]) behaves much differently. Here in optimal cases neither method outperforms any other for small
O-set cardinalities, but for higher cardinalities (Figs. S15–S16) the O-set is best in more than 50% of configurations (slightly less for nonlinear experiments) and the others share the rest (except Adjust\(\text{X}_{\text{min}}\)). For non-optimal cases O, \(O_{\text{Cmin}}\), and Adjust\(\text{X}_{\text{min}}\) share the ranks. Regarding RMSE, for linear experiments the O-results are almost as optimal as for the LinReg estimator in the optimal setting. However, for non-optimal cases \(O_{\text{Cmin}}\) can have considerably smaller variance and seems to be a robust option then, similarly to Adjust\(\text{X}_{\text{min}}\). Also for nonlinear experiments \(O_{\text{Cmin}}\) is more robust. The RF estimator (Figs. S12–S13) is again different. Here no method clearly is top-ranked, \(O_{\text{min}}\) and Adjust\(\text{X}_{\text{min}}\) are slightly better for linear experiments and O for nonlinear experiments. \(O_{\text{Cmin}}\) and \(O_{\text{min}}\) are more robust regarding RMSE ratios (similar to Adjust\(\text{X}_{\text{min}}\)). Finally, the DML estimator (Fig. S14) was here applied only to linear experiments since its model assumption does not allow for fully nonlinear settings. For optimal settings here O is top-ranked in a majority of cases, but closely followed by \(O_{\text{Cmin}}\) and Adjust\(\text{X}_{\text{min}}\). In non-optimal cases for higher O-set cardinalities these two seem like a better choice. Quantitatively, \(O_{\text{Cmin}}\) and Adjust\(\text{X}_{\text{min}}\) are the most robust choices. Overall, the O-set and its variants seem to outperform or match the Adjust-variants and whether higher cardinality of the O-set reduces performance depends strongly on the estimator and data.

4 Discussion and Conclusions

The proposed adjustment information formalizes the common intuition to choose adjustment sets that maximally constrain the effect variable and minimally constrain the cause variable. The main theoretical contributions are a necessary and sufficient graphical criterion for the existence of an optimal adjustment set in the hidden variables case and a definition and algorithm to construct it. To emphasize, graphical optimality implies that the O-set is optimal for any distribution consistent with the graph. Note that in cases where graphical optimality does not hold, there will still be distributions for which the O-set has maximal adjustment information.

Further, the optimal set is valid if and only if a valid adjustment set exists and has smaller (equal) asymptotic variance compared to the Adjust-set proposed in Perković et al. [2018] for any graph, whether graphical optimality holds or not. This makes the O-set a natural choice in automated causal inference analyses. Practical contributions comprise Python code to construct adjustment sets and check optimality, as well as extensive numerical experiments that demonstrate that the theoretical results also hold for relatively small sample sizes.

The theoretical optimality results are limited to estimators for which the asymptotic variance becomes minimal for adjustment sets with maximal adjustment information (relation (8)). This is fulfilled for least-squares estimators, where even the direct relation (9) holds, but it is unclear whether this also holds for more general classes. The numerical results show that the O-set or minimized variants thereof often yield smaller variance also in non-optimal settings and beyond that estimator class. I speculate that further theoretical properties of maximizing adjustment information can be shown because relation (9) for \(f(\cdot) = \frac{1}{\sqrt{n}} e^{H_{Y/X} - H_{X/Z}}\) seems related to the lower bound of the estimation variance counterpart to Fano’s inequality (Theorem 8.6.6 in Cover and Thomas [2006]). For estimators sensitive to high-dimensionality one may consider data-driven criteria or penalties to step-wisely minimize the O-set. However, estimating, for example, the adjustment information from a potentially small sample size carries considerable errors itself. Another current limitation is that relation (9) only holds for univariate singleton cause variables X. The information-theoretical results, however, also hold for multivariate X and preliminary results indicate that, while relation (9) does not hold for multivariate X, the less restrictive relation (8) still seems to hold.

The proposed information-theoretic approach can guide further research, for example, to theoretically study relations (8,9) for other estimators and to address other types of graphs as emerge from the output of causal discovery algorithms and the setting where the graph is unknown (Maathuis et al. 2009, 2010). At present, the approach only applies to ADMGs and Maximal Ancestral Graphs (MAG) [Richardson and Spirtes 2002] without selection variables. Last, it remains an open problem to identify optimal adjustment estimands for the hidden variables case based on other criteria such as the front-door formula and Pearl’s general do-calculus [Pearl 2009].

The results may carry considerable practical impact since, surprisingly, among the randomly created configurations more than 90% fulfill the optimality conditions indicating that also in many real-world scenarios graphical optimality may hold. Code is available in the python package https://github.com/jakobrunge/tigramite.
Acknowledgments and Disclosure of Funding

I thank Andreas Gerhardus for very helpful comments. This work was funded by the ERC Starting Grant CausalEarth (grant no. 948112).

References


Supplementary Material: Necessary and sufficient graphical conditions for optimal adjustment sets in causal graphical models with hidden variables

Jakob Runge
German Aerospace Center
Institute of Data Science
07745 Jena, Germany
and
Technische Universität Berlin
10623 Berlin, Germany
jakob.runge@dlr.de

A Problem setting and preliminaries

A.1 Graph terminology

We consider causal effects in causal graphical models over a set of variables $\mathbf{V}$ with a joint distribution $\mathcal{P} = \mathcal{P}(\mathbf{V})$ that is consistent with an acyclic directed mixed graph (ADMG) $\mathcal{G} = (\mathbf{V}, \mathcal{E})$. Two nodes can have possibly more than one edge which can be directed ($\rightarrow$) or bi-directed ($\leftrightarrow$). We use “∗” to denote either edge mark. There can be no loops or directed cycles. See Fig. 1A for an example.

The results also hold for Maximal Ancestral Graphs (MAG) [Richardson and Spirtes, 2002] without selection variables. A path between two nodes $X$ and $Y$ is a sequence of edges such that every edge occurs only once. A path between $X$ and $Y$ is called directed or causal from $X$ to $Y$ if all edges are directed towards $Y$, else it is called non-causal. A node $C$ on a path is called a collider if “$\leftrightarrow C \leftrightarrow$”. Kinships are defined as usual: parents $\text{pa}(X, \mathcal{G})$ for “$\bullet \rightarrow X$”, spouses $\text{sp}(X, \mathcal{G})$ for “$X \leftrightarrow \bullet$”, children $\text{ch}(X, \mathcal{G})$ for “$X \rightarrow \bullet$”, and correspondingly descendants des and ancestors an. We omit the $\mathcal{G}$ in the following since all relations are relative to the graph $\mathcal{G}$ in this paper. Our approach does not involve modified graph constructions as in van der Zander et al. [2019] and other works. A node is an ancestor and descendant of itself, but not a parent/child/spouse of itself. The mediator nodes on causal paths from $X$ to $Y$ in $\mathcal{G}$ are denoted $\mathcal{M} = \mathcal{M}(X, Y)$ and exclude $X$ and $Y$ (different from definitions in other works). For sets of variables the kinship relations correspond to the union of the individual variables. For parent/child/spouse-relationships these exclude the set of variables itself. A path $\pi$ between $X$ and $Y$ in $\mathcal{G}$ is blocked (or closed) by a node set $\mathcal{Z}$ if (i) $\pi$ contains a non-collider in $\mathcal{Z}$ or (ii) $\pi$ contains a collider that is not in $\text{anc}(\mathcal{Z})$. Otherwise the path $\pi$ is open (or active/connected) given $\mathcal{Z}$. Nodes $X$ and $Y$ are said to be m-separated given $\mathcal{Z}$ if every path between them is blocked by $\mathcal{Z}$, denoted as $X \perp \perp Y | \mathcal{Z}$. In the following we will simplify set notation and denote unions of variables as $\{W\} \cup \mathcal{M} \cup \mathcal{A} = W \mathcal{M} \mathcal{A}$.

B Further theoretical results and proofs

B.1 Properties of adjustment information

$J_\mathcal{Z}$ is not necessarily positive if the dependence between $X$ and $\mathcal{Z}$ (given $\mathcal{S}$) is larger than that between $\mathcal{Z}$ and $Y$ given $\mathcal{X}$. By the properties of CMI, it is bounded by

$$- \min(H_X|S, H_Z|S) \leq J_{X|S, Z} \leq \min(H_Y|X, S, H_Z|X, S).$$

(S1)

B.2 Causally sufficient case

Figure S1: DAG version of graph in Fig. 1A with O-set shown as blue boxes.

The optimal adjustment set for the causally sufficient case was derived in HPM19 and Rotnitzky and Smucler [2019]. Here the derivation is discussed from an information-theoretic perspective.

Definition B.1 (O-set in the causally sufficient case). Given Assumptions 7 restricted to DAGs with no hidden variables, define the set

\[ O = P = pa(YM) \setminus \text{forb} . \]

In the causally sufficient case a valid adjustment set always exists and the O-set is always valid since O contains no descendants of YM and all non-causal paths from X to Y are blocked since P blocks all paths from X through parents of YM.

Figure S1 shows an example DAG with a mediator M and conditioned variable S. The O-set \( O = Z_2Z_3 \) is depicted by blue boxes. Compare O with \( \text{vancs} = Z_1Z_2Z_3S \) (Adjust-set in Perkovic et al. [2018]) in the inequalities (7). Since \( Z_1 \perp \perp Y \mid OXS \), term (iii) is zero and since \( O \setminus \text{vancs} = \emptyset \), also term (iv) is zero. Further, terms (i) and (ii) are both strictly greater than zero (under Faithfulness). Then \( J_O > J_{\text{vancs}} \) and under Assumptions 2 by Lemma 2 the O-set has a smaller asymptotic variance than \( \text{vancs} \). Since the parents of YM block all paths from any other valid adjustment sets to Y and because any valid adjustment set Z has to block paths from X to \( pa(YM) \setminus Z \), \( J_O \geq J_Z \) holds in general for any valid set Z as proven from an information-theoretic perspective in Proposition B.1.

Proposition B.1 (Optimality of O-set in causally sufficient case). Given Assumptions 7 restricted to DAGs with no hidden variables and with \( O = P \) defined in Def. 7, graphical optimality holds for any graph and \( O \) is optimal.

Similar to HPM19 and Witte et al. [2020], there also exist results regarding minimality and minimum cardinality which are covered for the hidden variables case in Corollary B.1.

B.3 Hidden variables case

Here we provide some further theoretical results for the general hidden variables case in addition to the lemmas and theorems in the main text.

Corollary B.1 (Minimality and minimum cardinality). Given Assumptions 7 assume that graphical optimality holds, and, hence, \( O \) is optimal. Further it holds that:

1. If \( O \) is not minimal, then \( J_O > J_Z \) for all minimal valid \( Z \neq O \),

2. If \( O \) is minimal valid, then \( O \) is the unique set that maximizes the adjustment information \( J_Z \) among all minimal valid \( Z \neq O \),

3. \( O \) is of minimum cardinality, that is, there is no subset of \( O \) that is still valid and optimal.

Another relevant Proposition states that \( O_{C_{\text{min}}} \) is a subset of \( \text{vancs} \), similar to corresponding Lemmas in van der Zander et al. [2019].

Proposition B.2 (Collider-minimized O-set is a subset of Adjust.). Given Assumptions 7 with \( O = PCP_C \) defined in Def. 7 and the \( O_{C_{\text{min}}} \)-set constructed with Alg. C.2 it holds that \( O_{C_{\text{min}}} \subseteq \text{vancs} \).
B.4 Proof of Lemma [1]

**Lemma** (Necessary and sufficient comparison criterion for existence of an optimal set). Given Assumptions [1] if and only if there is a \( Z \in \mathcal{Z} \) such that either there is no other \( Z' \neq Z \in \mathcal{Z} \) or for all other \( Z' \neq Z \in \mathcal{Z} \) and all distributions \( \mathcal{P} \) consistent with \( \mathcal{G} \) it holds that

\[
\begin{align*}
I_{Z;Y|Z'X} & \geq I_{Z;Y|ZXS}, \\
I_{X;Z'|ZS} & \geq I_{X;Z'Z'S},
\end{align*}
\]

and

\[
\begin{align*}
I_{Z;Y|Z'X} & \geq I_{Z;Y|ZXS}, \\
I_{X;Z'|ZS} & \geq I_{X;Z'Z'S},
\end{align*}
\]

(S2)

then graphical optimality holds and \( Z \) is optimal implying \( J_Z \geq J_{Z'} \).

**Proof.** If there is no other \( Z' \), the statement trivially holds. Assuming there is another \( Z' \), we prove the two implications as follows by an information-theoretic decomposition.

Define disjunct (possibly empty) sets \( R, B, A \) with \( Z = AB \) and \( Z' = BR \) with \( B = Z \cap Z' \). Note that if both \( R = \emptyset \) and \( A = \emptyset \), then \( Z = Z' \). Consider two different ways of applying the chain rule of CMI,

\[
I_{ABR;Y|XS} = I_{X;ABR|S}
\]

\[
= I_{AB;Y|XS} + I_{R;Y|ABXS} - I_{X;AB|S} - I_{X;RB|S}, \tag{S3}
\]

\[
= I_{BR;Y|XS} + I_{A;Y|BRXS} - I_{X;BR|S} - I_{X;A|BR}, \tag{S4}
\]

from which with \( J_Z = I_{AB;Y|XS} - I_{X;AB|S} \) and \( J_{Z'} = I_{RB;Y|XS} - I_{X;RB|S} \) it follows that

\[
J_Z = J_{Z'} + I_{A;Y|BRXS} + I_{X;RB|S} - I_{X;A|BR}|Y. \tag{S5}
\]

The inequalities [S2] then read

\[
I_{A;Y|BRXS} \geq I_{R;Y|ABXS}, \quad \text{and} \quad I_{X;RB|S} \geq I_{X;A|BR}, \tag{S6}
\]

“if”: If term (i) is greater or equal to term (iii) and term (ii) greater or equal to term (iv), then trivially \( J_Z \geq J_{Z'} \) for all distributions \( \mathcal{P} \).

“only if”: We prove the contraposition that if for all valid \( Z \) there exists a valid \( Z' \neq Z \) and a distributions \( \mathcal{P} \) consistent with \( \mathcal{G} \) such that

\[
I_{A;Y|BRXS} < I_{R;Y|ABXS}, \quad \text{or} \quad I_{X;RB|S} < I_{X;A|BR}. \tag{S7}
\]

then there always exists a modification \( \mathcal{P}' \) of the distribution \( \mathcal{P} \) such that \( J_Z < J_{Z'} \). This is because, in both cases, we can always construct a distribution for which terms (ii) and (i), respectively, become arbitrarily close to zero. Consider the two cases as follows:

1) there exists a distribution \( \mathcal{P} \) with \( I_{A;Y|BRXS} < I_{R;Y|ABXS} \): Since CMI are always non-negative, it holds that \( R \neq \emptyset \) and there must exist at least one open path between \( R \) and \( Y \) where every collider is in \( ABXS \) and no non-collider is in \( ABXS \). No such open path can pass through \( X \) because if \( X \) is a non-collider (as for paths continuing on causal paths from \( X \) to \( Y \)), then the path is blocked, and if \( X \) is a collider, then there would be a non-causal path from \( X \) to \( Y \) given \( ZS \) which would make \( Z \) invalid while \( Z \) is assumed valid. Correspondingly, no open path from \( A \) (if \( A \neq \emptyset \)) to \( Y \) given \( BRXS \), if a path exists at all, can pass through \( X \) if \( Z' \) is assumed valid. Now we can construct a distribution \( \mathcal{P}' \) with associated structural causal model (SCM) consistent with \( \mathcal{G} \) where \( I_{A;Y|BRXS} < I_{R;Y|ABXS} \) holds as in \( \mathcal{P} \) and still all links “\( U \rightarrow X \)” for \( X \in X \) and \( U \notin XMY \).
We have to show that $X \perp \perp U$ with $U := f_U(\ldots, X, \ldots)$, (2) $X \leftarrow U$ with $X := f_X(\ldots, U, \ldots)$, and (3) $X \leftrightarrow U$ with $X := f_X(\ldots, L_U, \ldots)$ where $L_U$ denotes one or more latent variables. In each case, to go from $\mathcal{P}$ to $\mathcal{P}'$, we can modify $f \rightarrow f'$ where in $f'$ the dependence on the respective argument is replaced by $X \rightarrow cX, U \rightarrow cU$, or $L_U \rightarrow cL_U$ for $c \in \mathbb{R}$, and where we consider the limit $c \rightarrow 0$. This modification does not affect $I_{A,Y|BRS} < I_{R,Y|ABXS}$ because the paths contributing to the two CMIs cannot pass through $X$. On the other hand, then term (ii) $I_{X,R|ABS} \rightarrow 0$ because all paths passing through $X$ contain almost zero links and there cannot be a path from $R$ to $X$ through $YM'$ for a valid $Z$. Hence, since in Eq. (S5) term (i) is smaller than term (iii) by assumption, and term (ii) is almost zero, it holds that $J_z < J_{Z_z}$.

2) there exists a distribution $\mathcal{P}$ with $I_{X,R|ABS} < I_{X,A|BRS}$: As before, since CMIs are always non-negative, it holds that $A \neq \emptyset$ and there must exist at least one open path between $A$ and $X$ where every collider is in $BRS$ and no non-collider is in $BRS$. No such open path can pass through $YM$ because if any node in $YM$ is a collider, then the path is blocked, and no path can contain any node in $YM$ as a non-collider since then either the graph is cyclic or $Z'$ contains descendants of $YM$ leading to $Z' \cap \text{forb} \neq \emptyset$ while $Z'$ is assumed valid. Correspondingly, no open path from $R$ (if $R \neq \emptyset$) to $X$ given $ABS$, if a path exists at all, can pass through $YM$ if $Z$ is assumed valid. Then, analogous to before, we can construct a $\mathcal{P}'$ with associated SCM consistent with $\mathcal{G}$ where $I_{X,R|ABS} < I_{X,A|BRS}$ holds and where all links $U \rightarrow W$ for $W \in YM$ and $U \notin X \cup MY'$ almost vanish. Then term (i) $I_{A,Y|BRS} \rightarrow 0$ because all paths contain almost zero links and there cannot be a path from $A$ to $Y$ where $X$ contains a collider for a valid $Z'$ since this would constitute a non-causal path. Hence, since in Eq. (S5) term (ii) is smaller than term (iv) by assumption, and term (i) is almost zero, it holds that $J_z < J_{Z_z}$.

### B.5 Proof of Proposition 4.1

**Proposition (Optimality of O-set in causally sufficient case).** Given Assumptions [1] restricted to DAGs with no hidden variables and with $O = \mathcal{P}$ defined in Def. 4.1 graphical optimality holds for any graph and $O$ is optimal.

**Proof.** The proof is based on Lemma 4.1 and relation (S5). We will prove that for any DAG $\mathcal{G}$ term (i)$\geq$(iii) and term (ii)$\geq$(iv) from which optimality follows by Lemma 4.1

We have to show that $I_{A,Y|BRS} \geq I_{R,Y|ABXS}$ and $I_{X,R|ABS} \geq I_{X,A|BRS}$ where $O = AB$ and $Z' = RB$ with $B = O \cap Z'$.

Any path from $X$ or $V \setminus YMOSX$ to $YM$ given OS (denoted by [2]), excluding the causal path from $X$ to $Y$, features at least one of the following motifs: “$X, V \leftarrow [P] \rightarrow W$” (excluding “$X \rightarrow [P] \rightarrow W$”), or “$V \leftarrow W$” where, hence, $V \in \text{forb}$.

Now all paths from a valid adjustment set $Z'$ with $Z' \in Z$ to $Y$ are blocked given OS: Motif “$X, V \leftarrow [P] \rightarrow W$” contains a non-collider in OS and is, hence, blocked. In motif “$V \leftarrow W$” $V \in \text{forb}$. Since $X \notin \text{des}(Y)$ (acyclicity) and $Z' \cap \text{des}(Y) = \emptyset$ (validity of $Z'$), the paths from $Z'$ to $V$ either end with a head at $V$ or there must be a collider $K$ that is a descendant of $V$ and hence, $K \in \text{forb}$. Then $K \notin \text{pa}(OS)$ and $K \notin Z'$ and the path is therefore blocked. Hence, with $R \subseteq Z'$, term (iii) is zero by Markovity.

Term (iv) $I_{X,A|Z|S} = 0$ for any valid $Z'$ because $A \subseteq \text{pa}(YM)$ and then otherwise there would be a non-causal path from $X$ through $A$ to $YM$.

### B.6 Further Lemmas

**Lemma B.1** (relevant path motifs wrt. the O-set). Given Assumptions [1] but without a priori assuming that a valid adjustment set exists (apart from the requirement $S \cap \text{forb} = \emptyset$). With $O = \text{PCP}_C$ defined in Def. 4.2, any path from $X$ or $V \setminus YMOSX$ to $YM$ given OS (denoted by [2]), excluding the causal path from $X$ to $Y$, features at least one of the following motifs with certain constraints as indicated. We denote $V \in V \setminus YMOSX$ and further differentiate nodes in $YM$ as $W \in YM$ and in $O = \text{PCP}_C$ as $C \in C$ or $P \in \mathcal{P} \cap \mathcal{PC}$. Last, we denote those collider path nodes not included in the O-set in Alg. 4.1 due to not sufficing Def. 3.1) as $F$ with $F \in \text{forb}$ and those not sufficing Def. 3.2a,b) as $N$ with $N \notin \text{forb}$, $N \notin \text{vans}$, and $N \nsubseteq X \setminus \text{vans}$.}


(1a) “$\leftarrow X \rightarrow C \leftarrow$”
(1b) “$\leftarrow X \rightarrow P \rightarrow C \rightarrow$”

(2a) “$X, V \leftarrow P \rightarrow W$” excluding “$X \rightarrow P \rightarrow W$”
(2b) “$X, V \leftarrow P \rightarrow C \rightarrow$” excluding (1b)

(3a) “$V \leftarrow W$” where, hence, $V \in \text{forb}$
(3b) “$X, V \leftarrow C \rightarrow$”

(4a) “$\leftarrow F \rightarrow W$” with the constraint $F \notin \text{vanCS}$
(4b) “$\leftarrow F \rightarrow C \rightarrow$” with the constraints $F \notin \text{pa}(C)$ and $F \notin \text{vanCS}$

(5a) “$\leftarrow N \rightarrow W$” with the constraints $N \notin \text{pa}(W)$ and $W \notin \text{pa}(N)$
(5b) “$\leftarrow N \rightarrow C \rightarrow$” with the constraint $N \notin \text{pa}(C)$

Further it holds that $F, N, X \notin S$.

Proof. Any path from $X$ or $V \setminus \text{YMOS}X$ to $Y \text{M}$ has to contain a link “$A \leftarrow B$” where $A = X$ or $A \in V \setminus \text{YMOS}X$ and $B \in Y \text{M}$O where $\leftrightarrow \in \{\to, \leftarrow, \leftrightarrow\}$. If we differentiate the left node by $X$ or $V \in V \setminus \text{YMOS}X$ and the right node by $W \in Y \text{M}$ or $C \in C$ or $P \in P$ or $P \in \text{PC}$, we can in principle have $2 \cdot 4 \cdot 3 = 24$ link types which are motifs if we consider the adjacent links to $A$ and $B$. These are listed in the Lemma except for “$\leftarrow \rightarrow X \rightarrow W$” which is part of the causal path from $X$ to $Y$, “$X \rightarrow P \rightarrow W$” which cannot occur since then $P \in \text{M}$, “$V \rightarrow W$” which cannot occur since $P$ would contain $V$ or $V \in \text{des}(Y \text{M})$ leading to a cyclic graph, “$V \rightarrow C$” which cannot occur since $P \text{C}$ would contain $V$, and “$X \rightarrow W$” which cannot occur since this implies a cyclic graph.

Regarding the constraints listed in motifs (4a,b) for $F \in \text{forb}$ it holds that $F \notin \text{vanCS}$ because $\text{vanCS} = an(YS) \setminus \text{forb}$ by definition. Further, in (4b) $F \notin \text{pa}(C)$ holds because otherwise $C \in \text{forb}$. In motif (5a) $N \notin \text{pa}(W)$ holds because $N \notin \text{vanCS}$ and $W \notin \text{pa}(N)$ holds because $N \notin \text{forb}$. In motif (5b) $N \notin \text{pa}(C)$ holds because $C \in \text{vanCS}$ contradicts $N \notin \text{vanCS}$ and $N \notin X \mid \text{vanCS}$ with $N \rightarrow C$ contracts $C \perp \perp X \mid \text{vanCS}$. Last, it holds that $F, N, X \notin S$ because $S \cap \text{forb} = \emptyset$, $S \cap X = \emptyset$ by Assumptions and $N \notin \text{vanCS}$ while $S \subseteq \text{vanCS}$.

Lemma B.2 (Sufficient condition for non-identifiability). Given Assumptions but without a priori assuming that a valid adjustment set exists (apart from the requirement $S \cap \text{forb} = \emptyset$). With $O = \text{PCP} \text{C}$ defined in Def. if on any non-causal path from $X$ to $Y$ given $OS$ any of the motifs (1a) or (4a) or (4b) for $F = X$ occurs as listed in Lemma B.2 then the causal effect of $X$ on $Y$ (potentially through $M$) is not identifiable by backdoor adjustment.

Proof. If motif (4a) “$X \leftarrow W$” for $W \in YM$ occurs, the case is trivial [Pearl 2009, Thm. 4.3.1]. In motifs (1a) “$X \rightarrow C \leftarrow$” and (4b) “$X \leftarrow C \rightarrow$” we have that since Def. 3, (2b) $C \perp \perp X \mid \text{vanCS}$ is not fulfilled, Def. 3, (2a) $C \in \text{vanCS}$ must be the case. Then every $C_k$ on collider paths to $W$ also fulfills $C_k \in \text{vanCS}$ because for all of them $C_k \perp \perp X \mid \text{vanCS}$ does not hold since each collider is opened. Hence, there exists a collider path $X \leftarrow C \leftarrow \cdots \leftarrow W$ where every collider $C \in \text{vanCS} = an(YS) \setminus \text{forb}$. This path cannot be blocked by any adjustment set (given $S$): colliders with $C \in an(S)$ are always open. For colliders with $C \in an(X)$ or $C \in an(Y)$ there is a directed path to $X$ or $Y$ and either this path is open leading to a non-causal path, or an adjustment set contains a non-collider on that directed path which opens the collider $C$.

In Theorem we will prove that the condition in Lemma B.2 is also necessary for non-identifiability by backdoor adjustment. To this end, consider the following Lemmas.

Lemma B.3 (Collider parents fulfill Def. 3). Given Assumptions. With $O = \text{PCP} \text{C}$ defined in Def. for every $P \in \text{PC}$ conditions (1), and (2a) or (2b) in Def. hold.

Proof. Denote a pair $P \rightarrow C$ for $C \in C$ fulfilling conditions (1), and (2a) or (2b) in Def. Firstly, (1) $P \notin \text{forb}$ since if $P \in \text{des}(YM)$ also $C \in \text{des}(YM)$ and if $P \rightarrow X$, then by Lemma B.2 no valid adjustment set exists, contrary to Assumptions. Secondly, it cannot be that (2a) $P \notin \text{vanCS}$ and (2b) $P \leftarrow X \mid \text{vanCS}$ because then the path from $X$ to $P$ would extend to $C$ and would not be
Lemma B.4 (Blockness of parent-child-motifs). Given Assumptions 7 with \( \Omega = \text{PCP}_C \) defined in Def. 7 Any path from \( X \) or a valid adjustment set \( Z \) with \( Z \in Z \rightarrow Y \) containing the motifs (1b), (2a), (2b), (3a), (3b) is blocked given \( \Omega \).

Proof. Motifs (1b), (2a), (2b), and (3b) contain a non-collider in \( \Omega \) and are, hence, all blocked. In motif (3a) \( V \in \text{forb} \). Since \( X \notin \text{des}(Y) \) (acyclicity) and \( Z \cap \text{des}(Y) = \emptyset \) (validity of \( Z \)), the paths from \( Z \) to \( V \) either end with a head at \( V \) or there must be a collider \( K \) that is a descendant of \( V \) and hence, \( K \in \text{forb} \). Then \( K \notin \text{an}(\Omega) \) and \( K \notin Z \) and the path is therefore blocked. \( \square \)

Lemma B.5 (Blockness of F-motifs). Given Assumptions 7 with \( \Omega = \text{PCP}_C \) defined in Def. 7 Firstly, any path from \( X \) to \( Y \) containing the motifs (4a) or (4b) for \( F \in \text{des}(YM) \) is blocked given \( \Omega \). Secondly, any path from a valid adjustment set \( Z \) with \( Z \in Z \rightarrow Y \) containing the motifs (4a) or (4b) for \( F \in \text{des}(YM) \) is blocked given \( X \Omega \).

Proof. First statement: \( F \notin \text{vancs} \) by Lemma B.1 and, hence, in particular \( F \notin \text{an}(X) \). Then, if a path exists, either the paths from \( X \) to \( F \) end with a head at \( F \) or there must be at least one collider \( K \) with \( F \in \text{an}(K) \) on a path to \( X \). Now \( F, K \notin \text{an}(\Omega) \) because \( \Omega \cap \text{forb} = \emptyset \) and the path is blocked. Secondly, \( F \notin \text{an}(Z) \) since \( Z \) is valid. Then similarly, if a path exists, either the paths from \( Z \) to \( F \) end with a head at \( F \) or there must be at least one collider \( K \) on a path to \( Z \) with \( F \in \text{an}(K) \). Now \( F, K \notin \text{an}(X \Omega) \) because \( \Omega \cap \text{forb} = \emptyset \) and \( F \notin \text{vancs} \) by Lemma B.1 and the path is blocked. \( \square \)

Lemma B.6 (Blockness of N-motifs). Given Assumptions 7 with \( \Omega = \text{PCP}_C \) defined in Def. 7 Firstly, any path from \( X \) to \( Y \) containing the motifs (5a) or (5b) is blocked given \( \Omega \). Secondly, any path from a valid adjustment set \( Z \) to \( Y \) containing the motifs (5a) or (5b) is blocked given \( X \Omega \) if \( Z \notin \text{an}(N) \).

Proof. First statement: \( N \notin \text{vancs} \) by definition of \( N \) and, hence, in particular \( N \notin \text{an}(X) \). Then, if a path exists, either the paths from \( X \) to \( N \) end with a head at \( N \) or there must be at least one collider \( K \) with \( N \in \text{an}(K) \) on a path to \( X \). Now \( N, K \notin \text{an}(\Omega) \) can be seen by considering the different parts of \( \Omega \): \( N, K \notin \text{an}(PS) \) since \( N, K \notin \text{vancs} \) and \( N, K \notin \text{an}(C) \) for \( C \in \text{vancs} \cap \text{CP}_C \). Finally, \( N, K \notin \text{an}(C) \) for \( C \in \text{CP}_C \cap C \perp X \mid \text{vancs} \) because \( N, K \notin \text{an}(C) \mid \text{vancs} \). Hence, the path is blocked. Second statement: If \( Z \) does not contain any descendants of \( N \), then \( N \notin \text{an}(Z) \). Then any path from a \( Z \) is blocked by the same reasoning as in the first part with the addition that \( N \notin \text{an}(X) \) and hence the motif is blocked given \( X \Omega \). \( \square \)

The following Lemma is not needed in this paper, but may be of interest for further research.

Lemma B.7 (Existence of X-N-path). Given Assumptions 7 with \( \Omega = \text{PCP}_C \) defined in Def. 7 There must be at least one path from \( X \) to \( N \) (defined in the motifs (5a) or (5b)) that ends with a head at \( N \) and where every collider is in \( \text{vancs} \) and every non-collider is not in \( \text{vancs} \).

Proof. By definition of the N-node, \( N \notin X \mid \text{vancs} \). Now all paths that end with a tail at \( N \) are blocked given \( \text{vancs} \) because \( N \notin \text{an}(X) \) and the first collider \( K \) coming from \( N \) must be blocked because \( K \notin \text{vancs} \). Hence, there must be an open path that ends with a head at \( N \) and where every collider is in \( \text{vancs} \) and every non-collider is not in \( \text{vancs} \) as stated. \( \square \)

B.7 Proof of Theorem 1

Theorem (Validity of O-set). Given Assumptions 1 but without a priori assuming that a valid adjustment set exists (apart from the requirement \( S \cap \text{forb} = \emptyset \)). If and only if a valid backdoor adjustment set exists, then \( \Omega \) is a valid adjustment set.

Proof. “if”: Given that a valid backdoor adjustment set exists, we need to prove that (i) \( \Omega \cap \text{forb} = \emptyset \) with \( \text{forb} = X \cup \text{des}(YM) \) and (ii) all non-causal paths from \( X \) to \( Y \) are blocked by \( \Omega \) (given \( S \)). (i) is true by the construction of \( \Omega \) in Def. 1 and Alg. C.1 where nodes \( \in \text{des}(YM) \) are not added and nodes that are \( X \) indicate non-identifiability (see Lemma B.2). By Lemma B.3 also \( \text{PC} \cap \text{des}(YM) = \emptyset \) and \( X \notin \text{PC} \) because otherwise no valid adjustment set exists by Lemma B.2.

Lemma B.1 lists all possible motifs on non-causal paths. By Lemma B.2 the occurrence of the motifs (1a) or (4a) or (4b) for \( F = X \) renders the effect non-identifiable, contrary to the assumption. Hence
only the remaining motifs can occur. By Lemma B.4, the motifs (1b), (2a), (2b), (3a), (3b) are blocked given \( OS \). By Lemma B.5 (part one) the motifs (4a,b) for \( F \in \text{des}(YM) \) are blocked given \( OS \). By Lemma B.6 (part one) motifs (5a) and (5b) are blocked given \( OS \).

"only if" is trivially true since \( O \) is then assumed valid. □

### B.8 Proof of Theorem 2

**Theorem (O-set vs Adjust-set)**. Given Assumptions 1 with \( O \) defined in Def. 4 and the Adjust-set vanes defined in Eq. (2), it holds that \( J_O \geq J \text{vanes} \) for any graph \( G \). We have \( J_O = J \text{vanes} \) only if \( O = \text{vanes} \) or \( O \subseteq \text{vanes} \) and \( X \perp\!\!\!\perp \text{vanes} \mid O \cup OS \).

**Proof.** We directly use the decomposition in Eq. (5) with \( Z = O \) and \( Z' = \text{vanes} = BR \) with \( \text{vanes} = an(XS) \setminus \text{forb} \) and the definitions of \( R, B, A \) as in Eq. (5). For term (iii), \( I_{R,Y \mid OXS} \) to be non-zero, there must be an active path from \( R \subseteq \text{vanes} \) to \( Y \) given \( XOS \). By Lemma B.1, Lemma B.4, Lemma B.5 (second part), and Lemma B.6 (second part), the only possibly open motifs on paths from \( R \) to \( Y \) given \( OS \) are \( \{ e \leftarrow C \leftrightarrow \} \) or \( \{ e \leftarrow C \leftarrow \} \) where \( R \cap \text{des}(N) \) \( \neq \emptyset \). But since \( R \subseteq \text{vanes} \) and \( N \notin \text{vanes} \), \( R \) cannot contain descendants of \( N \). Hence, term (iii) is zero. For term (iv), \( I_{X,A \mid \text{br} X} = I_{X,A \mid \text{vanes}} \), note that \( A = O \setminus \text{vanes} \) and, hence, for all \( A \in A \) it holds that \( A \perp\!\!\!\perp X \mid \text{vanes} \) since all \( A \in A \) then fulfill Def. 3(2b) (for \( A \in CP \) see Lemma B.3). Hence, \( I_{X,A \mid \text{vanes}} = 0 \) by Markovity. This proves that \( J_O \geq J \text{vanes} \).

We are now left with terms (i) and (ii) in Eq. (5). By construction of the collider path nodes, \( A \subseteq CP \) is connected to \( Y \) (potentially through \( M \)) conditional on \( \text{vanes} X \) since \( \text{vanes} \) contains all remaining collider nodes in \( C \). Then by Faithfulness term (i) \( I_{A,Y \mid \text{br} X} = I_{A,Y \mid \text{vanes} X} \) can only be zero if \( A = \emptyset \). Then \( O \subseteq \text{vanes} \). Term (ii), \( I_{X,R \mid OS} = 0 \) if \( R = \text{vanes} \setminus O = \emptyset \) or \( X \perp\!\!\!\perp \text{vanes} \mid O \cup OS \). □

### B.9 Proof of Proposition B.2

**Proposition (Collider-minimized O-set is a subset of Adjust.).** Given Assumptions 1 with \( O = \text{PC}_C \) defined in Def. 4 and the \( O_{C\text{min}} \)-set constructed with Alg. C.2, it holds that \( O_{C\text{min}} \subseteq \text{vanes} \).

**Proof.** Define \( C_{\text{min}} = O_{C\text{min}} \setminus P \). We need to show that \( C \in C_{\text{min}} \Rightarrow C \in \text{vanes} \) for all \( C \in O \setminus P \). Assume \( C \notin \text{vanes} \). Since then all \( C \in O \setminus P \) fulfill Def. 3(2b) (for \( C \in PC \) see Lemma B.3), it holds that \( C \perp\!\!\!\perp \text{vanes} \) implying that no link \( X \leftrightarrow C \) exists. If a path exists at all, either (i) there must be at least one collider \( K \) with \( C \in an(K) \) and \( K \notin \text{vanes} \) on a path to \( X \) or (ii) \( C \in \text{des}(X) \).

We now show that for case (i) (\( C \) has no open path to \( X \) given \( SO \setminus \{C\} \)). \( K \notin \text{os} \) can be seen by considering the different parts of \( OS \): \( K \notin an(PS) \) since \( K \notin \text{vanes} \) and \( an(PS) \subseteq \text{vanes} \). Further, \( K \notin an(\text{vanes} \cap C) \). Finally, \( K \notin an(CP \setminus \text{vanes}) \) since \( C \subseteq CP \) (\( \text{vanes} \) fulfilling (by Def. 3(2b)) \( C \perp\!\!\!\perp \text{vanes} \) \( \land \) \( K \not\perp\!\!\!\perp X \)). Hence, \( X \perp\!\!\!\perp C \mid SO \setminus \{C\} \) implying that \( C \) would be removed in the first loop of Alg. C.2 and \( C \notin C_{\text{min}} \), contrary to assumption.

In case (ii) the directed path from \( X \) to \( C \) for \( C \in C \setminus PC \) is blocked because \( PC \subseteq O \) contains all parents of \( C \) and \( X \notin PC \) since we assume identifiability. This implies that \( C \) would be removed in the first loop of Alg. C.2 and \( C \notin C_{\text{min}} \), contrary to assumption. Finally, if there exists a directed path from \( X \) to \( C = PC \setminus PC \) for \( PC \notin \text{vanes} \) we know that all children \( C \in ch(PC) \) \( \land \) \( CP \) were removed in the first loop of Alg. C.2. Denote the remaining nodes after the first loop of Alg. C.2 by \( O_{C\text{min}} \). \( PC \notin \text{vanes} \) has no directed path to \( Y \) and is separated from \( Y \) given \( SO_{C\text{min}} \) because the motif \( PC \rightarrow C \leftrightarrow \) is blocked since \( C \notin an(O_{C\text{min}}) \). This implies that \( PC \) would be removed in the second loop of Alg. C.2 and \( PC \notin C_{\text{min}} \), contrary to assumption. □

### B.10 Proof of Theorem 3

**Theorem (Necessary and sufficient graphical conditions for optimality and optimality of O-set).** Given Assumptions 1 and with \( O = \text{PC}_C \) defined in Def. 4. Denote the set of \( N \)-nodes by \( N = \text{sp}(YM) \setminus \text{forb} OS \). Finally, given an \( N \in N \) and a collider path \( N \leftrightarrow \cdots \leftrightarrow C \leftrightarrow \cdots \leftrightarrow W \) (including \( N \leftrightarrow W \)) for \( C \in C \) and \( W \in YM \) indexed by \( i \) with the collider path nodes denoted by \( \pi_i^N \) (excluding \( N \) and \( W \)), denote by \( O_{\pi_i^N} = O(X, Y, \pi_i^N = SN \pi_i^N) \) the O-set for the causal effect of \( X \) on \( Y \) given \( \pi_i^N \).
If and only if exactly one valid adjustment set exists, or both of the following conditions are fulfilled, then graphical optimality holds and \( O \) is optimal:

(I) For all \( N \in N \) and all its collider paths \( i \) to \( W \in YM \) that are inside \( C \) it holds that \( O_{\pi_i} \) does not block all non-causal paths from \( X \) to \( Y \), i.e., \( O_{\pi_i} \) is non-valid,

and

(II) for all \( E \in O \setminus P \) with an open path to \( X \) given \( SO \setminus \{ E \} \) there is a link \( E \rightarrow W \) or an extended collider path \( E \rightarrow C \leftarrow \cdots \rightarrow W \) inside \( C \) for \( W \in YM \) where all colliders \( C \in \text{vances} \).

**Proof.** If exactly one valid adjustment set exists, then optimality holds by Def. 2 and then this set is \( O \) because \( O \) is always valid if a valid set exists (Lemma 1).

The proof is based on Lemma 1 and relation (S5). We will first prove the “if”-statement by showing that Cond. (I) leads to term (i) \( \geq \) (iii) and Cond. (II) leads to term (ii) \( \geq \) (iv) from which optimality follows by Lemma 1. Then we prove the “only if”-statement by showing that if either of the two conditions is not fulfilled, then (i) \( < \) (iii) or (ii) \( < \) (iv) for some distribution \( \mathcal{P} \) consistent with \( \mathcal{G} \) and graphical optimality does not hold.

“if”: We have to show that if both conditions hold, then \( I_{A;Y|B;XS} \geq I_{R;Y|A;BXS} \) and \( I_{X;R|A;BS} \geq I_{X;A|B;BS} \) where \( O = AB \) and \( Z' = RB \) with \( B = O \cap Z' \). Further, we use \( A_P = A \cap P \) and \( A_C = (A \cap CP_C) \setminus A_P \) where \( A = A_P \cup A_C \).

Condition (I) directly leads to \( I_{A;Y|B;XS} \geq I_{R;Y|A;BXS} \) as follows.

We subdivide condition (I) into two cases where the former implies the latter: (I.1) There are no \( N \)-nodes, i.e., \( N = \emptyset \), or (I.2) for all \( N \in N \) and all its collider paths \( i \) it holds that \( O_{\pi_i} \) does not block all non-causal paths from \( X \) to \( Y \).

If condition (I.1) holds, then there are no \( N \)-nodes. If there are no \( N \)-motifs on any path from \( R \) to \( Y \), then by Lemma B.1 and Lemma B.4 (second part) all paths given \( XOS \) are blocked and term (iii) is zero by Markovity.

If condition (I.2) holds, then there are \( N \)-nodes. By Lemma B.6 (second part) the only possibly open motifs on paths from \( R \) to \( Y \) given \( OXS \) are “\( \leftarrow \rightarrow N \leftarrow \rightarrow W \)” or “\( \leftarrow \rightarrow N \leftarrow \rightarrow C \leftarrow \rightarrow \)” where \( R \cap \text{des}(N) \neq \emptyset \). Term (iii), \( I_{R;Y|B;XS} = I_{R;Y|X;SO} \), is then always non-zero since, by definition of the \( N \)-nodes, there exists at least one collider path \( N \leftarrow \cdots \leftarrow C \leftarrow \cdots \rightarrow W \) (including \( N \rightarrow W \)) for \( C \in C \) and \( W \in YM \). To see under which conditions still term (i) \( \geq \) (iii) consider two ways of decomposing the following CMI:

\[
I_{AR;Y|B;XS} = I_{A;Y|B;XS} + I_{R;Y|B;XS} \quad \text{term (i)}
\]

\[
= I_{R;Y|B;XS} + I_{A;Y|B;XS} \quad \text{term (ii)}
\]

\[
= I_{R;Y|B;XS} + I_{A;Y|B;XS} \quad \text{term (iii)}
\]

From this decomposition we see that term (i) \( \geq \) (iii) if and only if term (i') \( \geq \) (iii'). Paths from \( R \) to \( Y \) via \( X \) given \( SXZ' \setminus R = BSX \) are blocked because if \( X \) is a collider, then there would be a non-causal path rendering \( Z' \) invalid. Therefore, for term (iii') to be non-zero \( Z'S \) must contain at least descendants of an \( N \)-node \( N \) and all its collider path nodes towards \( W \), denoted \( \pi_i^N \), for at least one path \( i \). Then \( R \cap \text{des}(N) \neq \emptyset \) and \( \pi_i^N \subseteq BS \) such that there exists an open path “\( N \leftarrow \rightarrow C \leftarrow \cdots \rightarrow C \leftarrow \rightarrow W \)” (or \( N \rightarrow W \)).

Condition (I.2) now guarantees that for all \( N \in N \) and all collider paths indexed by \( i \) the O-set \( O_{\pi_i} \), which includes \( N \pi_i \) as a subset, does not block all non-causal paths. By Theorem 1 if \( O_{\pi_i} \) is not valid, then no valid adjustment set \( Z' \) containing \( N \pi_i \) as a subset exists. And this in turn implies that no valid set with \( R \cap \text{des}(N) \neq \emptyset \) exists. To show this, assume the contraposition: If there was such a valid set \( Z' \) with \( R \cap \text{des}(N) \neq \emptyset \) and \( \pi_i^N \subseteq Z' \), then it would open the collider motif \( \leftarrow \rightarrow N \leftarrow \rightarrow W \) since \( R \) contains descendants of \( N \) and lead to an open path “\( N \leftarrow \rightarrow C \leftarrow \cdots \rightarrow C \leftarrow \rightarrow W \)” (or \( N \rightarrow W \)). If \( Z' \) is still valid, it must block all paths from \( X \) that end with an arrowhead at \( N \). But then also \( Z' \cup \{ N \} \) is invalid. Note that since \( N \notin \text{forb} \), \( S \cap \text{forb} = \emptyset \), and \( \pi_i^N \cap \text{forb} = \emptyset \) since
\[ \pi_N \subseteq C, \text{ the validity of } O_{Z_N} \text{ depends only on its ability to block non-causal paths. Hence, term (iii') is zero and by Eq. (S8) term (i) \geq (iii).} \]

Condition (II) directly leads to \( I_{X; R\mid A_{BS}} \geq I_{X; A\mid BR S} \) as follows.

Define \( E = \{ E \in O \mid P : X \not\perp \perp E \mid \mathcal{SO} \setminus \{ E \} \} \). By condition (II) there exists a link \( E \leftrightarrow W \) or an extended collider path \( E \leftrightarrow \cdots \leftrightarrow W \) inside \( C \) for \( W \in YM \) where all colliders \( C \in \text{vanss} \). There are two types: (1) \( E \rightarrow C \leftrightarrow \cdots \leftrightarrow W \) (then \( E \in P_C \)) and (2) \( E \leftrightarrow W \) or \( E \leftrightarrow C \leftrightarrow \cdots \leftrightarrow W \). We consider two cases:

Case (1): \( E \in \mathcal{E} \) for which there exists at least one path of type (1). Any valid \( Z' \) with \( E \notin \mathcal{Z}' \) has to block paths from \( X \) to \( E \) since otherwise there is a non-causal open path from \( X \) to \( Y \) through the motif chain \( \cdots \leftrightarrow E \leftrightarrow C \leftrightarrow \cdots \leftrightarrow W \) for \( W \in YM \): \( E \) is open since \( E \notin \mathcal{Z}' \) and the path from \( E \) to \( W \) is open since all colliders \( C \in \text{vanss} \): if \( C \in \text{an}(S) \), the collider is always opened and if \( C \in \text{an}(XY) \) then either the directed path to \( X \) or \( Y \) is open, or \( C \) is opened if \( Z' \) contains a node on that path.

Case (2): \( E \in \mathcal{E} \) for which all paths are of type (2). Firstly, all paths from \( X \) to \( E \) that end with a tail at \( E \) must be blocked by \( Z' \) since otherwise there is a non-causal path as for case (1). The same holds for paths that end with a head at \( E \) if \( E \in \text{vanss} \). Consider paths that end with a head at \( E \) and \( E \notin \text{vanss} \) which implies \( E \perp \perp X \mid \text{vanss} \) by Def. 3. Then it follows that \( E \perp \perp X \mid \text{SO} \setminus \{ E \} \) and, hence, \( E \notin \mathcal{E} \) which we can show by considering where \( E \) can occur with respect to the different motifs listed in Lemma B.1 as follows (see the definitions of \( W, V, F, N, C, P_C \) there): Motif (1a) \( \cdots \leftrightarrow E \rightarrow C \leftrightarrow \cdots \rightarrow W \) is not relevant since then non-identifiability holds and motif (2a) \( \cdots \leftrightarrow E \rightarrow [P] \rightarrow W \) is not relevant since \( E \notin \mathcal{P} \). Motifs (3a) \( V \leftarrow W \), (4a) \( \cdots \rightarrow F \rightarrow W \), and (5a) \( \cdots \rightarrow N \rightarrow W \) are not relevant since no \( E \in \mathcal{O} \) is involved. For the motifs (1b) \( \cdots \leftrightarrow E \rightarrow [P_C] \rightarrow W \), (2b) \( \cdots \leftrightarrow E \rightarrow [P_C] \rightarrow C \leftrightarrow \cdots \rightarrow W \), and (3b) \( \cdots \leftrightarrow E \rightarrow [C] \leftrightarrow \cdots \rightarrow W \) the path to \( X \) is blocked by \( \text{SO} \setminus \{ E \} \). For motif (4b) \( \cdots \leftrightarrow E \rightarrow C \leftrightarrow \cdots \rightarrow W \) or \( \cdots \leftrightarrow F \rightarrow E \), since \( \text{SO} \cap \text{forb} = \emptyset \) and \( X \cap \text{desc(forb)} = \emptyset \), there must exist a collider in \( \text{forb} \) or \( \cdots \rightarrow F \rightarrow E \) on a path to \( X \) which is then blocked. Hence, \( E \notin \mathcal{E} \). Finally, for (5b) \( \cdots \rightarrow N \rightarrow E \rightarrow C \leftrightarrow \cdots \rightarrow W \) or \( \cdots \rightarrow E \rightarrow N \rightarrow W \) with \( N \notin \text{vanss} \) and \( X \not\perp \perp N \mid \text{vanss} \) we either have \( \cdots \rightarrow N \rightarrow E \) or there exists a collider on any path to \( X \) with \( K \in \text{desc}(N) \) and, hence, \( N \notin \text{vanss} \). \( E \perp \perp X \mid \text{SO} \setminus \{ E \} \) would only be possible if \( N \) or \( K \in \text{an} (O \mid \text{vanss}) \). The subset \( O \mid \text{vanss} \) fulfills \( O \mid \text{vanss} \mid \perp \perp X \mid \text{vanss} \) by Def. 3. However, \( N \) or \( K \in \text{an} (O \mid \text{vanss}) \) implies that there is a path from \( O \mid \text{vanss} \) to \( N \). Then \( X \not\perp \perp N \mid \text{vanss} \) contradicts \( O \mid \text{vanss} \mid \perp \perp X \mid \text{vanss} \) implying that \( N, K \notin \text{an} (O \mid \text{vanss}) \) and, hence, \( E \perp \perp X \mid \text{SO} \setminus \{ E \} \) and \( E \notin \mathcal{E} \).

Both cases taken together, it holds that \( X \perp \perp E \mid \text{SZ} \setminus \{ E \} \) for any valid \( Z' \). Furthermore, \( X \perp \perp P \mid \text{SZ} \setminus \{ P \} \) with \( P \in \mathcal{P} \) for any valid \( Z' \) since \( P \) is directly connected to \( Y \) and, therefore, a valid \( Z' \) has to block a non-causal path between \( X \) and \( Y \) through \( P \).

Now decompose term (iv) as

\[
I_{X; A\mid Z' \mid S} = \underbrace{I_{X; A\mid (A_P A_E)\mid Z' \mid S}}_0 + \underbrace{I_{X; A\mid (A_P A_E)\mid Z' \mid S A_P A_E}}_{(S9)}
\]

with \( A_P = A \cap P \) and \( A_E = (A \cap E) \setminus A_P \). The preceding derivations imply \( X \perp \perp A_P A_E \mid Z' \mid S \) for any valid \( Z' \) and, hence, the first term vanishes.

Consider the set \( E' = \{ E' \in O \mid P : X \not\perp \perp E' \mid \mathcal{SO} \setminus \{ E' \} \} \). This implies that \( A_{E'} = A \setminus (A_P A_E) \) fulfills \( A_{E'} \mid \perp \perp X \mid \mathcal{SO} \setminus A_{E'} \) and since \( A_{E'} = S B A_P A_E \) we have

\[
I_{X; A_{E'} \mid S B A_P A_E} = 0. \tag{S10}
\]

This now leads to term (ii) \( \geq \) term (iv) by considering two ways of decomposing the following CMI:

\[
I_{X; R A_{E'} \mid S B A_P A_E} = \underbrace{I_{X; A_{E'} \mid S B A_P A_E}}_{=0 \text{ by Eq. (S10)}} + \underbrace{I_{X; R \mid S B A_P A_E A_{E'}}}_{\text{term (ii)}} \geq \underbrace{I_{X; A_{E'} \mid S B A_P A_E}}_{\text{term (iv) by Eq. (S9)}} + \underbrace{I_{X; A_{E'} \mid S B A_P A_E R}}_{\text{term (ii) by Eq. (S9)}}.
\]

“only if”: We need to prove that if either Condition (I) or Condition (II) or both are not fulfilled, then graphical optimality does not hold (implying that also \( O \) is not optimal).
We define $E$Alternatively, the negation of Condition (II) directly leads to (1,2b') “$A\rightarrow U$” for $SO$ (or (2) “$A\rightarrow C$”) $\in A \astern$ denote one or more latent variables and $c \in \mathbb{R}$.

Thirdly, we consider the modified motifs with unconditioned $C, P$ paths through $W$ inside $C$ for $W \in YM$ contain at least one collider $C \notin \text{vancs}$. Define the set of these non-ancestral colliders as $C_E = \{C \in C : E\rightarrow \cdots \rightarrow C\rightarrow \cdots \rightarrow Y\} \cup \{Y\}$.

Ad (1): As a subset of $O$ we have that $Z' \cap \text{forb} = \emptyset$. We investigate whether $Z'$ blocks all non-causal paths between $X$ and $Y$ by considering the motifs in Lemma B.3. In addition to all those motifs listed there are modified motifs where unconditioned $C$-nodes and $P$-nodes occur (denoted without a $\Box$) due to removing $E_C$ from $O$.

Firstly, the unmodified motifs are blocked as before (see Theorem I): Motif (1a) “$\astern X\rightarrow \Box C \leftarrow \astern$” is not relevant since then non-identifiability holds. By Lemma B.4 the motifs (1b), (2a), (2b), (3a), (3b) all contain a non-collider in $SO \setminus E_C$ and are blocked. By Lemma B.5 (part one) the motifs (4a,b) for $F \in \text{des}(YM)$ are blocked because $Z' \cap \text{forb} = \emptyset$. By Lemma B.5 (part one) motifs (5a) and (5b) are blocked given $SO \setminus E_C$ because the proof in Lemma B.6 requires that on paths to $X$ either $N$ is a collider or there exists a descendant collider $K$ and that $N, K \notin \text{an}(OS)$. The latter is fulfilled because $SO \setminus E_C$ is a subset of $SO$.

Secondly, all paths from $X$ through the removed node $E$ to $W \in YM$ are blocked by $SO \setminus E_C$: Paths through $P$ are blocked since $E \notin P$ and $\text{des}(C_E) \cap \text{vancs} = \emptyset$ and, hence, $P \subseteq SO \setminus E_C$. Paths through colliders are blocked by the negation of condition (II): there is no link $E\leftrightarrow W$ and all extended collider paths $E\rightarrow C\rightarrow \cdots \rightarrow W$ inside $C$ for $W \in YM$ contain at least one collider $C \notin \text{vancs}$. By construction, $E_C = \{E\} \cup (\text{des}(C_E) \cap O)$, implying that all these non-ancestral colliders are blocked.

Thirdly, we consider the modified motifs with unconditioned $C, P \in \{\text{des}(C_E) \cap O\}$. By definition of $C_E$ in (S13), $C, P \notin \text{vancs}$. (As a remark, $E$ can potentially be in $\text{vancs}$.)

Motif (1a) “$\astern X\rightarrow \Box C \leftarrow \astern$” cannot occur since then non-identifiability holds. Modified motifs (1.2b) “$X, V \astern P_C \rightarrow C \astern Y$” and (1.2b) “$X, V \astern P_C \rightarrow C \astern \cdots \astern C \astern Y$” are blocked since they contain a non-collider. Motifs (1.2b) “$X, V \astern P_C \rightarrow C \astern Y$” are blocked since $C \notin \text{des}(SO \setminus E_C)$. Motif (2a) “$X, V \astern P \rightarrow W$” is not possible since $P \in P \subseteq \text{vancs}$. Motifs (3a) “$V\rightarrow W$”, (4a) “$\astern F\rightarrow W$”, and (5a) “$\astern N\rightarrow W$” are not modified since no conditioned node occurs. Motif (3b) “$X, V \rightarrow C \astern Y$” is blocked because due to $C \notin \text{vancs}$ there must exist a descendant of $C$ that is a collider $K \notin \text{vancs}$ on the path to $X$. Since $E_C$ contains all descendants of $C$, also $K$ and all its descendants are not in $SO \setminus E_C$ and $K$ is blocked. Finally, motifs (4b) “$\astern F\rightarrow C \astern Y$” and (5b) “$\astern N\rightarrow C \astern Y$” are blocked since $SO \setminus E_C$ does not contain any descendant of $C$. This proves the validity of $Z'$. 

10
Ad (2): To show that \( I_{X:A|\mathcal{BRS}} = I_{X:E|\mathcal{SO}\setminus \mathcal{EC}} > I_{X:R|\mathcal{ABS}} = 0 \) we start from the assumption that \( E \not\subseteq X \mid \mathcal{SO} \setminus \{E\} \). This implies that there exists a path from \( X \) to \( E \) where no non-collider is in \( \mathcal{SO} \setminus \{E\} \) and for every collider \( K \) it holds that \( \text{des}(K) \cap \mathcal{SO} \setminus \{E\} \neq \emptyset \). With \( Z' = \mathcal{O} \setminus \mathcal{EC} \) all non-colliders are still open. Consider those colliders \( K \) with \( \text{des}(K) \cap (\mathcal{O} \setminus \{E\}) \subseteq \mathcal{EC} \setminus \{E\} \). Then these colliders are closed on the path from \( E \) to \( X \). However, for each such \( K \) there is a \( C \in \mathcal{EC} \setminus \{E\} \) with \( C \in \text{des}(K) \). Then the path from \( X \) through \( \cdots \rightarrow K \rightarrow \cdots \rightarrow C \) is open given \( \mathcal{SO} \setminus \mathcal{EC} \). Hence, at least for the last such collider on the path from \( E \) to \( X \) there is an open path from \( C \in \mathcal{EC} \setminus \{E\} \) to \( X \) given \( \mathcal{SO} \setminus \mathcal{EC} \). Then Faithfulness implies that \( I_{X:A|\mathcal{BRS}} = I_{X:E|\mathcal{SO}\setminus \mathcal{EC}} > I_{X:R|\mathcal{ABS}} = 0 \) and, hence, term (ii) < term (iv) holds for all distributions \( \mathcal{P} \) consistent with \( \mathcal{G} \). By Lemma 1, where the distribution \( \mathcal{P} \) is modified to \( \mathcal{P}' \) without affecting term (ii)<(iv), then graphical optimality does not hold.

This concludes the proof of Theorem 3. \( \square \)

B.11 Proof of Corollary B.1

**Corollary** (Minimality and minimum cardinality). Given Assumptions \( \square \) assume that graphical optimality holds, and, hence, \( O \) is optimal. Further it holds that:

1. If \( O \) is not minimal, then \( J_O > J_Z \) for all minimal valid \( Z \neq O \).
2. If \( O \) is minimal valid, then \( O \) is the unique set that maximizes the adjustment information \( J_Z \) among all minimal valid \( Z \neq O \).
3. \( O \) is of minimum cardinality, that is, there is no subset of \( O \) that is still valid and optimal.

**Proof.** We again define disjunct sets \( R, B, A \) with \( A = \mathcal{O} \setminus \mathcal{Z}, R = \mathcal{Z} \setminus \mathcal{O}, \) and \( B = \mathcal{O} \cap \mathcal{Z}, \) where any of them can be empty, but not both \( R \) and \( A \) since then \( \mathcal{Z} = \mathcal{O} \). Hence \( \mathcal{O} = \mathcal{A} \mathcal{B} \) and \( \mathcal{Z} = \mathcal{B} \mathcal{R} \). Consider relation (S14) in this case,

\[
J_O = J_Z + I_{A,Y|B\mathcal{R}X} + I_{X,R|\mathcal{ABS}} - I_{R,Y|\mathcal{AB}\mathcal{X}} - I_{X:A|\mathcal{BRS}}. \tag{S14}
\]

Part 1 and 2: Since graphical optimality holds, we know that \( J_O = J_Z \) can only be achieved if term (i) = term (iii) and term (ii) = term (iv). From Eq. (S8) we know that term (i) = (iii) can only hold if \( I_{A,Y|B\mathcal{R}X} = 0 \). But this implies \( A = \emptyset \) by Faithfulness since, by construction, \( A \subset \mathcal{O} \) is always connected to \( Y \) (potentially through \( M \)) given \( \mathcal{X} \mathcal{S} \mathcal{O} \setminus \mathcal{A} \). Then term (iv) = 0 and, by optimality, \( I_{X,R|\mathcal{ABS}} = 0 \). But the latter would imply that \( Z = \mathcal{B} \mathcal{R} \) is either not minimal anymore since \( R \) is not connected to \( X \) and, hence, does not block any non-causal path not already blocked by \( B \). Then \( J_O > J_Z \) among all minimal valid \( Z \) (Part 1). Or \( Z \) is minimal and \( R = \emptyset \), for which \( Z = \mathcal{O} \) is the unique set maximizing \( J_Z \) among all minimal valid \( Z \neq \mathcal{O} \) (Part 2).

Part 3, i.e., that removing any subset from \( O \) decreases \( J_O \) follows directly from setting \( R = \emptyset \) and considering \( A \neq \emptyset \) (since otherwise nothing would be removed). Then term (ii) and term (iii) are both zero and by optimality term (iv), which must be smaller or equal to term (ii), is zero. Since \( A \) is connected to \( Y \) (see Part 1) by Faithfulness we have \( J_O > J_{O\setminus A} \). \( \square \)
C Algorithms

Algorithm C.1 Construction of \( O \)-set and test for backdoor-identifiability.

**Require:** Causal graph \( G \), cause variable \( \mathcal{X} \), effect variable \( \mathcal{Y} \), mediators \( \mathcal{M} \), conditioned variables \( \mathcal{S} \)

1: Initialize \( P = \emptyset \), \( C = \emptyset \) and \( P_C = \emptyset \)
2: for \( W \in \mathcal{YM} \) do
3: \( P = P \cup pa(W) \setminus \text{forb} \)
4: for \( W \in \mathcal{YM} \) do
5: Initialize nodes in this level \( \mathcal{L} = \{W\} \)
6: Initialize ignorable nodes \( \mathcal{N} = \emptyset \)
7: while \( |\mathcal{L}| > 0 \) do
8: Initialize next level \( \mathcal{L}' = \emptyset \)
9: for \( C \in \text{sp}(\mathcal{L}) \setminus \mathcal{N} \) do
10: if \( C = \mathcal{X} \) then
11: return No valid backdoor adjustment set exists.
12: if \( C \notin C \) and Def.\( \text{(1)} \) \( C \notin \text{forb} \) and ((2a) \( C \in \text{vancs} \) or (2b) \( C \perp \perp \mathcal{X} | \text{vancs} \))
then
13: \( C = C \cup \{C\} \)
14: \( \mathcal{L}' = \mathcal{L}' \cup \{C\} \)
15: else
16: if \( C \notin C \) then
17: \( \mathcal{N} = \mathcal{N} \cup \{C\} \)
18: \( \mathcal{L} = \mathcal{L}' \setminus \mathcal{N} \)
19: for \( C \in C \) do
20: if \( X \in \text{pa}(C) \) then
21: return No valid backdoor adjustment set exists.
22: \( P_C = P_C \cup \text{pa}(C) \)
23: return \( O = PC_P C \)

Algorithm C.2 Construction of \( O_{\text{min}} \) and \( O_{C_{\text{min}}} \)-sets. The relevant code for \( O_{C_{\text{min}}} \) is indicated in parentheses.

**Require:** Causal graph \( G \), cause variable \( \mathcal{X} \), effect variable \( \mathcal{Y} \), mediators \( \mathcal{M} \), conditioned variables \( \mathcal{S} \), \( O = PC_P C \)-set

1: Initialize \( O_{\text{min}} = O \) (\( C_{\text{min}} = CP_C \setminus P \))
2: for \( Z \in O_{\text{min}} \) (\( Z \in C_{\text{min}} \)) do
3: if \( Z \) has no active path to \( X \) given \( \mathcal{S} \setminus \{Z\} \) then
4: Mark \( Z \) for removal
5: Remove marked nodes from \( O_{\text{min}} \) (\( C_{\text{min}} \))
6: for \( Z \in O_{\text{min}} \) (\( Z \in C_{\text{min}} \)) do
7: if \( Z \) has no active path to \( Y \) given \( X \mathcal{S} \setminus \{Z\} \) (given \( X \mathcal{S}PC_{\text{min}} \setminus \{Z\} \)) then
8: Mark \( Z \) for removal
9: Remove marked nodes from \( O_{\text{min}} \) (\( C_{\text{min}} \))
10: return \( O_{\text{min}} \) (\( O_{C_{\text{min}}} = PC_{\text{min}} \))
D Further details and figures of further numerical experiments

D.1 Setup

We compare the following adjustment sets (see definitions in Section 2.3):

- O
- Adjust
- O_{C_{\text{min}}}
- O_{\text{min}}
- Adjust_{X_{\text{min}}}
- Adjust_{\text{min}}

To investigate the applicability of different estimators, we use above adjustment sets together with the following estimators from sklearn (version 0.24.2) and the doubleml (version 0.4.0) package (see instantiated class for parameters):

- Linear ordinary least squares (LinReg) regressor `LinearRegression()`
- k-nearest-neighbor (kNN) regressor `KNeighborsRegressor(n_neighbors=3)`
- Multilayer perceptron (MLP) regressor `MLPRegressor(max_iter=2000)`
- Random forest (RF) [Breiman, 2001] regressor `RandomForestRegressor()`
- Double machine learning for partially linear regression models (DML) [Chernozhukov et al., 2018] `DoubleMLPLR(data, ml_g, ml_m)` from doubleml with `ml_g=ml_g=MLPRegressor(max_iter=2000)` from sklearn

Sklearn [Pedregosa et al., 2011] and doubleml [Bach et al., 2021] are both available under an MIT license.

As data generating processes we consider linear and nonlinear experiments generated with the following generalized additive model:

\[
V_j = \sum_{i} c_i f_i(V^i) + \eta_j \quad \text{for} \quad j \in \{1, \ldots, \tilde{N}\}. \tag{S15}
\]

To generate a structural causal model among \( \tilde{N} \) variables we randomly choose \( L \) links whose functional dependencies are linear for linear experiments and one half is \( f_i(x) = (1 + 5xe^{-x^2/20})x \) for nonlinear experiments. Coefficients \( c_i \) are drawn uniformly from \( \pm [0.1, 2] \). For linear experiments we use normal noise \( \eta_j \sim \mathcal{N}(0, \sigma^2) \) and, in addition, for nonlinear models \( \frac{1}{2} \) of the noise terms is Weibull-distributed, both with standard deviation \( \sigma \) drawn uniformly from \( [0.5, 2] \). From the \( \tilde{N} \) variables of each dataset we randomly choose a fraction \( \lambda \) as unobserved and denote the number of observed variables as \( N \). For each combination of \( N \in \{5, 10, 15, 20\} \), \( L \in \{2\tilde{N}, 3\tilde{N}\} \), and \( \lambda \in \{30\%, 40\%, 50\%\} \) we randomly create a structural causal model and then randomly pick an observed pair \((X = V^i, Y = V^j)\) connected by a causal path, set \( S = \emptyset \), and consider the intervention \( \text{do}(V^i = V^i + 1 = x) \) relative to the unperturbed data \((x')\) as ground truth, which corresponds to the linear regression coefficient in the linear case. We further assert that the following criteria hold: (1) the effect is identifiable, (2) the minimal adjustment cardinality is \( |\text{vancs}_{\text{min}}(X, Y)| > 0 \), and (3) the (absolute) causal effect is \( \geq 10^{-3} \) to make sure that Faithfulness holds (if these criteria cannot be fulfilled, another model is generated). We create 500 models for each combination of \( N, L, \lambda \). Surprisingly, among in total 12,000 randomly created configurations 93% fulfill the optimality conditions in Thm. This may indicate that also in many real-world scenarios graphical optimality actually holds. Here we do not consider the effect of a selected conditioning variable \( S \) since it would have a similar effect on all methods considered.

For the considered graphs the computation time to construct adjustment sets is very short and arguably negligible to the actual cost of fitting methods that use these adjustment sets. The results were evaluated on Intel Xeon Platinum 8260.
D.2 Figures for linear least squares estimator

93% of configurations are optimal

Opt: 72 / 34  
Adjust: 10 / 7

percentage of best-ranked among optimal / non-optimal configurations:

Opt: 72 / 34  
Adjust: 10 / 7
OptCmin: 7 / 46  
Optmin: 5 / 4  
AdjustXmin: 2 / 7  
Adjustmin: 0 / 0

Figure S2: Results of linear experiments with linear estimator and sample size \( n = 30 \). The diagonal depicts letter-value plots [Hofmann et al., 2017] of adjustment set cardinalities and the off-diagonal shows pairs of RMSE ratios for all combinations of (O, Adjust, O\(_{\text{Cmin}}\), O\(_{\text{min}}\), Adjust\(_{\text{Xmin}}\), Adjust\(_{\text{min}}\)) for optimal configurations (left in blue) and non-optimal configurations (right in green). Values above 200% are not shown. The dashed horizontal line denotes the median of the RMSE ratios, and the white plus their average. The letter-value plots are interpreted as follows: The largest box shows the 25%–75% range. The next smaller box above (below) shows the 75%–87.5% (12.5%–25%) range and so forth. The numbers on best-ranked methods at the top indicate the percentage of the 12,000 randomly created configurations where the method had the lowest variance. The highest percentage is marked in bold. Note that the highest ranked method may outperform others only by a small margin. The results in the letter-value plots provide a more quantitative picture. See also Fig. S7 where the ranks are further distinguished by the O-set cardinality.
Figure S3: As in Fig. S2 but for $n = 50$. 

93% of configurations are optimal

LinReg | random_linear | $n = 50$ | $d_x = 1$

percentage of best-ranked among optimal / non-optimal configurations:

Opt: 78 / 39   Adjust: 10 / 6   OptCmin: 5 / 45   Optmin: 3 / 3   AdjustXmin: 2 / 5   Adjustmin: 0 / 0
93% of configurations are optimal | random_lineargaussian | \( n = 100 \) | \( d_k = 1 \)
percentage of best-ranked among optimal / non-optimal configurations:
Opt: 81 / 45  Adjust: 9 / 6  Opt\(_{C_{\min}}\): 4 / 38  Opt\(_{\min}\): 2 / 2  Adjust\(_{\min}\): 1 / 6  Adjust\(_{\min}\): 0 / 0

Figure S4: As in Fig. S2 but for \( n = 100 \).
LinReg | random_lineargaussian | \( n = 1000 \) | \( d_x = 1 \)

93% of configurations are optimal | percentage of best-ranked among optimal / non-optimal configurations:
Opt: 85 / 50  Adjust: 8 / 6  Opt\(_{Cmin}\): 2 / 36  Opt\(_{min}\): 1 / 0  Adjust\(_{Xmin}\): 1 / 5  Adjust\(_{min}\): 0 / 0

Figure S5: As in Fig. S2 but for \( n = 1000 \).
Figure S6: As in Fig. S2 but for $n = 10000$. 

LinReg | random_lineargaussian | $n = 10000$ | $d_x = 1$

93% of configurations are optimal | percentage of best-ranked among optimal / non-optimal configurations:
Opt: 86 / 53 Adjust: 8 / 4 OptCmin: 1 / 36 Optmin: 1 / 0 AdjustXmin: 1 / 5 Adjustmin: 0 / 0

...
Figure S7: Percentage of configurations where each method has the lowest variance for linear experiments, stratified by the cardinality of the O-set (x-axis) for $n = 30$ (top) to $n = 10,000$ (bottom).
D.3 Figures for non-parametric estimators
Figure S8: As in Fig. S2 but with kNN estimator ($k = 3$) and $n = 1000$. See also Figs. S15, S16 where the ranks are further distinguished by the $O$-set cardinality.
93% of configurations are optimal | percentage of best-ranked among optimal / non-optimal configurations:

Figure S9: As in Fig. S2 but for kNN estimator ($k = 3$), the nonlinear model, and $n = 1000$. 
93% of configurations are optimal

Opt: 31 / 32  Adjust: 13 / 4  OptCmin: 18 / 26  Optmin: 10 / 6  Adjustmin: 8 / 5

Figure S10: As in Fig. S2 but for MLP estimator and $n = 1000$. 
Figure S11: As in Fig. S2 but for MLP estimator, the nonlinear model, and $n = 1000$. 

93% of configurations are optimal | percentage of best-ranked among optimal / non-optimal configurations:

MLP | random_nonlinearmixed | $n = 1000$ | $d_k = 1$
RF | random_lineargaussian | $n = 1000$ | $d_s = 1$

93% of configurations are optimal | percentage of best-ranked among optimal / non-optimal configurations:

Opt: 14 / 10  Adjust: 10 / 5  Opt$C_{\min}$: 14 / 17  Opt$min$: 24 / 24  Adjust$X_{\min}$: 13 / 19  Adjust$min$: 22 / 22

Figure S12: As in Fig. S2 but for RF estimator and $n = 1000$. 
RF | random_nonlinearmixed | $n = 1000$ | $d_k = 1$

93% of configurations are optimal | percentage of best-ranked among optimal / non-optimal configurations: Opt: 21/25 Adjust: 14/10 OptCmin: 16/18 Optmin: 15/12 AdjustXmin: 16/23 Adjustmin: 14/10

Figure S13: As in Fig. S2 but for RF estimator, the nonlinear model, and $n = 1000$. 

26
93% of configurations are optimal | percentage of best-ranked among optimal / non-optimal configurations:
Opt: 24 / 19  Adjust: 13 / 4  Opt_{Cmin}: 20 / 29  Opt_{min}: 12 / 8  Adjust_{min}: 19 / 28  Adjust_{min}: 10 / 9

Figure S14: As in Fig. S2 but for DML estimator and $n = 1000$. 

DML | random_lineargaussian | $n = 1000$ | $d_x = 1$
Figure S15: As in Fig. S7 but including non-parametric estimators for $n = 1000$. 
Figure S16: As in Fig. S15 but for nonlinear experiments.
References


