Appendix: Learning Predictive Models That Transport

A ID Algorithm

Algorithm 1: ID($\mathbf{X}$, $\mathbf{Y}$; $\mathcal{G}$)

input : ADMG $\mathcal{G}$, disjoint variable sets $\mathbf{X}, \mathbf{Y} \subset \mathbf{O}$
output: Expression for $P_{\mathbf{X}}(\mathbf{Y})$ if identified or `FAIL` if not identified.
1. $\mathbf{D} = an_{\mathcal{G}_{\mathbf{O}\setminus\mathbf{X}}}(\mathbf{Y})$;
2. Let c-components of $\mathcal{G}_{\mathbf{D}}$ be $\mathbf{D}_i$, $i = 1, \ldots, k$;
3. $P_{\mathbf{X}}(\mathbf{Y}) = \sum_{\mathbf{D}\setminus\mathbf{Y}} \prod_{i=1}^{k}$ `Identify`$(\mathbf{D}_i, \mathbf{O}, P(\mathbf{O}))$;

Function `Identify`$(\mathbf{A}, \mathbf{V}, Q = Q[\mathbf{V}])$**:**
- **if** $\mathbf{A} == \mathbf{V}$ **then**
  - **return** $Q[\mathbf{V}]$;
- /* $C_{\mathcal{G}_{\mathbf{V}}}(B)$ `is c-component of` B `in` $\mathcal{G}_{\mathbf{V}}$ */
- **if** $\exists B \in \mathbf{V} \setminus \mathbf{A}$ *such that* $C_{\mathcal{G}_{\mathbf{V}}}(B) \cap ch(B) = \emptyset$ **then**
  - Compute $Q[\mathbf{V} \setminus \{B\}]$ from Q (Corollary 1);
  - **return** `Identify`$(\mathbf{A}, \mathbf{V} \setminus \{B\}, Q[\mathbf{V} \setminus \{B\}])$;
- **else**
  - **return** `FAIL`$(\mathbf{A}, \mathcal{G}_{\mathbf{V}})$;

We now restate the identification algorithm (ID) (Tian and Pearl, 2002; Shpitser and Pearl, 2006) using the modified presentation in Jaber et al. (2018). When the interventional distribution of a set of variables is identified, the ID algorithm returns it in terms of observational distributions (i.e., if the intervention is represented using *do* notation, then the resulting expression contains no *do* terms). The ID algorithm is complete (Shpitser and Pearl, 2006), so if the interventional distribution is not identifiable, then the algorithm throws a failure exception. Note that $\mathcal{G}_{\mathbf{V}}$ denotes an *induced subgraph* which consists of only the variables in $\mathbf{V}$ and the edges between variables in $\mathbf{V}$.

We will need the following definition:

Definition 1 (C-component)**.** *In an ADMG, a c-component consists of a maximal subset of observed variables that are connected to each other through bidirected paths. A vertex with no incoming bidirected edges forms its own c-component.*

We also restate the following Corollary (Jaber et al., 2018, Corollary 1):

Corollary 1. *Given an ADMG $\mathcal{G}$ with observed variables $\mathbf{O}$ and unobserved variables $\mathbf{U}$, $V \in \mathbf{X} \subseteq \mathbf{O}$, and $P_{\mathbf{O}\setminus\mathbf{X}}$, if V is not in the same c-component with a child of V in $\mathcal{G}_{\mathbf{X}}$, then $Q[\mathbf{X} \setminus \{V\}]$ is identifiable and is given by*

$$Q[\mathbf{X} \setminus \{V\}] = \frac{P_{\mathbf{O}\setminus\mathbf{X}}}{Q[C(V)]} \sum_{V} Q[C(V)],$$

where $C(V)$ denotes the c-component of V in the induced subgraph $\mathcal{G}_{\mathbf{X}}$.

This Corollary allows us to derive the post-intervention distribution after intervening on V from the post-intervention distribution after intervening on the variables in $\mathbf{O} \setminus \mathbf{X}$. The modified presentation of Tian's ID algorithm given in Jaber et al. (2018) is in Algorithm 1, which computes the identifying functional for the post-interventional distribution of the variables in $\mathbf{Y}$ after intervening on the variables in $\mathbf{X}$ by recursively finding the identifying functional for each c-component in the post-intervention subgraph.

B Proofs

B.1 Soundness and Completeness of the Surgery Estimator

Theorem 1 (Soundness)**.** *When Algorithm 2 returns an estimator, the estimator is stable.*

Proof. Any query Algorithm 2 makes to ID considers intervening on a superset of the mutable variables $\mathbf{X} \supseteq \mathbf{M}$. By Proposition 1 this means the target interventional distribution is stable. From the soundness of the ID algorithm (Shpitser and Pearl, 2006, Theorem 5), the resulting functional of observational distributions that Algorithm 2 returns will be stable. □

Theorem 2 (Completeness)**.** *If Algorithm 2 fails, then there exists no stable surgery estimator for predicting T.*

Proof. Algorithm 2 is an exhaustive search over interventional distributions that intervene on supersets of $\mathbf{M}$ and are functions of T. Thus, by completeness of the ID algorithm (Shpitser and Pearl, 2006, Corollary 2), if there is a stable surgery estimator, the procedure will find one. □

B.2 Relationship with Graph Pruning

Lemma 1. *Let T be the target variable of prediction and $\mathcal{G}$ be a selection ADMG with selection variables $\mathbf{S}$. If there exists a stable conditioning set $\mathbf{Z}$ such that $P(T|\mathbf{Z}) = P(T|\mathbf{Z}, \mathbf{S})$, then Algorithm 2 will not fail on input $(\mathcal{G}, ch(\mathbf{S}), T)$.*

Proof. Assume that $P(T|\mathbf{Z})$ is a stable graph pruning estimator. Partition $\mathbf{Z}$ into $\mathbf{X}$ and $\mathbf{W}$ such that $\mathbf{X} \subseteq \mathbf{M}$ and $\mathbf{W} \cap \mathbf{M} = \emptyset$, and let $\mathbf{V} = \mathbf{M} \setminus \mathbf{X}$. It must be that $T \perp\!\!\!\perp \mathbf{X}|\mathbf{W}$ in $\mathcal{G}_{\underline{\mathbf{X}}}$. If this were not the case then there would be some $X \in \mathbf{X}$ such that there was a backdoor path from T to X, and since $X \in ch(\mathbf{S})$ there is a path $T \cdots \rightarrow X \leftarrow S$. Because X is conditioned upon, this collider path would be active and $S \not\perp\!\!\!\perp T$, implying $P(T|\mathbf{Z})$ is not stable (a contradiction). Now by Rule 2 of do-calculus, $P(T|\mathbf{X}, \mathbf{W}) = P_{\mathbf{X}}(T|\mathbf{W})$. Next consider the remaining mutable variables $\mathbf{V}$. Letting $\mathbf{V}(\mathbf{W})$ denote the subset of $\mathbf{V}$ nodes that are not ancestors of any $\mathbf{W}$ nodes in $\mathcal{G}_{\overline{\mathbf{X}}}$, we will show that $T \perp\!\!\!\perp \mathbf{V}|\mathbf{X}, \mathbf{W}$ in $\mathcal{G}_{\overline{\mathbf{X}},\overline{\mathbf{V}(\mathbf{W})}}$. First consider $V \in \mathbf{V}(\mathbf{W})$. For the independence to not hold, there must be an active forward path from V to T. But because $V \in ch(\mathbf{S})$, the path $S \rightarrow V \rightarrow \ldots T$ is active since V is not conditioned upon, implying contradictorily that $P(T|\mathbf{Z})$ was not stable. Now consider $V \in \mathbf{V} \setminus \mathbf{V}(\mathbf{W})$. For the independence to not hold, either there is an active forward path from V to T, or there is an active backdoor path from V to T. We previously showed the first case. In the second case, because V is an ancestor of some $W \in \mathbf{W}$ that is conditioned upon, the collider path $S \rightarrow V \leftarrow \ldots T$ is active, so $P(T|\mathbf{Z})$ is not stable (contradiction). Thus, by Rule 3 of do-calculus, we have that $P_{\mathbf{X}}(T|\mathbf{W}) = P_{\mathbf{M}}(T|\mathbf{W})$. This is one of the conditional interventional queries that Algorithm 2 considers, so Algorithm 2 will not fail. □

B.3 Optimality

Theorem 3. *If $\mathcal{G}$ is such that $P_{\mathbf{M}}(T|\mathbf{X} \setminus \mathbf{M})$ is identified and equal to $P(T|\mathbf{W})$ for some $\mathbf{W} \subseteq \mathbf{X}$, then $f_s(\mathbf{x}) = E[T|\mathbf{x} \setminus \mathbf{m}, do(\mathbf{m})]$ achieves (2):*

$$f_s \in \operatorname*{argmin}_{f \in \mathcal{C}^0} \sup_{Q_{\mathbf{s}} \in \Gamma} E_{Q_{\mathbf{s}}}[(t - f(\mathbf{x}))^2].$$

Proof. The structure of this proof follows that of Theorem 4 in Rojas-Carulla et al. (2018) which proves the optimality of using invariant conditional distributions to predict in an adversarial setting.

Consider a function $f \in \mathcal{C}^0$, possibly different from f_s. Now for each distribution $\mathbb{Q} \in \Gamma$ corresponding to an environment, we will construct a distribution $\mathbb{P} \in \Gamma$ such that

$$\int (t - f(\mathbf{x}))^2 d\mathbb{P} \geq \int (t - f_s(\mathbf{x}))^2 d\mathbb{Q}.$$

Denote the density of $\mathbb{Q}$ by $q(\mathbf{x}, t)$. Note that we have assumed that all distributions in Γ correspond to the same graph $\mathcal{G}$ in which $P_{\mathbf{M}}(T|\mathbf{X} \setminus \mathbf{M}) = P(T|\mathbf{W})$ for some $\mathbf{W} \subseteq \mathbf{X}$. Because $\mathbb{Q} \in \Gamma$, q factorizes according to (1) as a product of conditional densities (even when bidirected edges are present, the observational joint can be

factorized as a product of univariate conditionals using the c-component factorization (Tian, 2002)). To construct the density $p(\mathbf{x}, t)$ of $\mathbb{P}$ from q, for $M \in \mathbf{M}$ replace the $q(M|\cdot)$ terms with the marginal density $q(M)$. This is equivalent to removing the edges into $\mathbf{M}$ so notably $P(\mathbf{O} \setminus \mathbf{M}|\mathbf{M}) = P_{\mathbf{M}}(\mathbf{O} \setminus \mathbf{M})$ by rule 2 of *do*-calculus. Thus in $\mathbb{P}$ we have that $P_{\mathbf{M}}(T|\mathbf{X} \setminus \mathbf{M}) = P(T|\mathbf{X})$. But since the full conditional interventional distribution is stable it must be that $P(T|\mathbf{X}) = P(T|\mathbf{W})$. So, we know that $q(t|w) = p(t|w)$. Further, we have that $p(\mathbf{w}) = q(\mathbf{w})$ since we constructed $\mathbb{P}$ to have the same marginals of $\mathbf{M}$ as $\mathbb{Q}$ and the other terms remain stable across members of Γ. Thus $q(t, \mathbf{w}) = p(t, \mathbf{w})$. Letting $\mathbf{Z} = \mathbf{X} \setminus \mathbf{W}$, we note that $\mathbf{T} \perp\!\!\!\perp \mathbf{Z}|\mathbf{W}$ in $\mathbb{P}$. We now have that

$$
\begin{aligned}
\int (t - f(\mathbf{x}))^2 d\mathbb{P} &= \int_{t,\mathbf{x}} (t - f(\mathbf{x}))^2 p(\mathbf{x}, t) d\mathbf{x} dt \\
&\geq \int_{t,\mathbf{x}} (t - E[T|\mathbf{x}])^2 p(\mathbf{x}, t) d\mathbf{x} dt && \text{(Conditional mean minimizes MSE)} \\
&= \int_{t,\mathbf{x}} (t - E[T|\mathbf{x} \setminus \mathbf{m}, do(\mathbf{m})])^2 p(\mathbf{x}, t) d\mathbf{x} dt \\
& && \text{(Conditional and interventional distributions are equal by construction)} \\
&= \int_{t,\mathbf{w}} \int_{\mathbf{z}} (t - f_s(\mathbf{m}, \mathbf{x} \setminus \mathbf{m}))^2 p(\mathbf{z}|\mathbf{w}) p(\mathbf{w}, t) d\mathbf{z} d\mathbf{w} dt \\
&= \int_{t,\mathbf{w}} \int_{\mathbf{z}} (t - f_s(\mathbf{w}))^2 p(\mathbf{z}|\mathbf{w}) p(\mathbf{w}, t) d\mathbf{z} d\mathbf{w} dt && (E[T|\mathbf{x} \setminus \mathbf{m}, do(\mathbf{m})] = E[T|\mathbf{w}]) \\
&= \int_{t,\mathbf{w}} \int_{\mathbf{z}} (t - f_s(\mathbf{w}))^2 p(\mathbf{z}|\mathbf{w}) q(\mathbf{w}, t) d\mathbf{z} d\mathbf{w} dt && (q(t, \mathbf{w}) \text{ is stable}) \\
&= \int_{t,\mathbf{w}} \int_{\mathbf{z}} (t - f_s(\mathbf{w}))^2 q(\mathbf{z}|t, \mathbf{w}) d\mathbf{z} q(\mathbf{w}, t) d\mathbf{w} dt && ((t - f_s(\mathbf{w}))^2 \text{ is not a function of } \mathbf{z}) \\
&= \int_{\mathbf{z},t,\mathbf{w}} (t - f_s(\mathbf{w}))^2 q(\mathbf{w}, t, \mathbf{z}) d\mathbf{w} dt d\mathbf{z} \\
&= \int_{t,\mathbf{x}} (t - f_s(\mathbf{w}))^2 d\mathbb{Q}
\end{aligned}
$$

□

Theorem 4. *If $\mathcal{G}$ is such that $P_{\mathbf{M}}(T|\mathbf{X}\setminus\mathbf{M})$ is identified and not a function of $\mathbf{M}$, then $f_s(\mathbf{x}) = E[T|\mathbf{x}\setminus\mathbf{m}, do(\mathbf{m})]$ achieves (2):*

$$f_s \in \operatorname*{argmin}_{f \in \mathcal{C}^0} \sup_{Q_{\mathbf{s}} \in \Gamma} E_{Q_{\mathbf{s}}}[(t - f(\mathbf{x}))^2].$$

Proof. The structure of this proof closely follows the structure of the previous proof.

Consider a function $f \in \mathcal{C}^0$, possibly different from f_s. Now for each distribution $\mathbb{Q} \in \Gamma$ corresponding to an environment, we will construct a distribution $\mathbb{P} \in \Gamma$ such that

$$\int (t - f(\mathbf{x}))^2 d\mathbb{P} \geq \int (t - f_s(\mathbf{x}))^2 d\mathbb{Q}.$$

We shall again construct $\mathbb{P}$ from $\mathbb{Q}$ such that in $\mathbb{P}$ $P(T|\mathbf{X} \setminus \mathbf{M}, do(\mathbf{M})) = P(T|\mathbf{X})$. Note that we have assumed that $P(T|\mathbf{X} \setminus \mathbf{M}, do(\mathbf{M})))$ is not a function of $\mathbf{M}$. This usually corresponds to a *dormant independence* or *Verma constraint* (Shpitser and Pearl, 2008) in the graph: it means that $T \perp\!\!\!\perp \mathbf{M}|\mathbf{X} \setminus \mathbf{M}$ in $\mathcal{G}_{\overline{\mathbf{M}}}$ (the graph in which edges into $\mathbf{M}$ have been deleted). Further discussion of this can be found in the next subsection of the supplement.

Let $\mathbf{Z} = \mathbf{X} \setminus \mathbf{M}$. By Proposition 1 we have that $p(t, \mathbf{z}|do(\mathbf{m})) = q(t, \mathbf{z}|do(\mathbf{m})$ where p and q denote the densities of $\mathbb{P}$ and $\mathbb{Q}$, respectively. Note that recovering the joint density $p(t, \mathbf{z}, \mathbf{m})$ from $p(t, \mathbf{z}|do(\mathbf{m}))$ requires multiplying by a functional of the observational distribution $\mathbb{P}$ of the form $p'(\mathbf{m}|t, \mathbf{z})$ (that is, a product of *kernels* (Richardson et al., 2017) or conditional-like univariate densities of $\mathbf{m}$) where p' denotes that this is not an observational conditional density.

$$
\begin{aligned}
\int (t - f(\mathbf{x}))^2 d\mathbb{P} &= \int_{t,\mathbf{x}} (t - f(\mathbf{x}))^2 p(\mathbf{x}, t) d\mathbf{x} dt \\
&\geq \int_{t,\mathbf{x}} (t - E[T|\mathbf{x}])^2 p(\mathbf{x}, t) d\mathbf{x} dt && \text{(Conditional mean minimizes MSE)} \\
&= \int_{t,\mathbf{z},\mathbf{m}} (t - E[T|\mathbf{z}, do(\mathbf{m})])^2 p(\mathbf{z}, \mathbf{m}, t) d\mathbf{z} d\mathbf{m} dt \\
& && \text{(Conditional and interventional distributions are equal by construction)} \\
&= \int_{t,\mathbf{z},\mathbf{m}} (t - f_s(\mathbf{m}, \mathbf{z}))^2 p(\mathbf{z}, \mathbf{m}, t) d\mathbf{z} d\mathbf{m} dt \\
&= \int_{t,\mathbf{z},\mathbf{m}} (t - f_s(\mathbf{m}, \mathbf{z}))^2 p(\mathbf{z}, t|do(\mathbf{m})) p'(\mathbf{m}|t, \mathbf{z}) d\mathbf{z} d\mathbf{m} dt \\
&= \int_{t,\mathbf{z},\mathbf{m}} (t - f_s(\mathbf{m}, \mathbf{z}))^2 q(\mathbf{z}, t|do(\mathbf{m})) p'(\mathbf{m}|t, \mathbf{z}) d\mathbf{z} d\mathbf{m} dt && \text{(Stability of } q(\mathbf{z}, t|do(\mathbf{m})) \text{ by Prop 1)} \\
&= \int_{t,\mathbf{m},\mathbf{z}} (t - f_s(\mathbf{z}))^2 q(\mathbf{z}, t|do(\mathbf{m})) p'(\mathbf{m}|t, \mathbf{z}) d\mathbf{z} d\mathbf{m} dt && (E[T|do(\mathbf{m}, \mathbf{z}] \text{ is not a function of } \mathbf{m}) \\
&= \int_{t,\mathbf{m},\mathbf{z}} (t - f_s(\mathbf{z}))^2 q(\mathbf{z}, t|do(\mathbf{m})) q'(\mathbf{m}|t, \mathbf{z}) d\mathbf{z} d\mathbf{m} dt && ((t - f_s(\mathbf{z}))^2 \text{ is not a function of } \mathbf{m}) \\
&= \int_{t,\mathbf{m},\mathbf{z}} (t - f_s(\mathbf{z}))^2 q(\mathbf{z}, t, \mathbf{m}) d\mathbf{z} d\mathbf{m} dt \\
&= \int_{t,\mathbf{x}} (t - f_s(\mathbf{z}))^2 d\mathbb{Q}
\end{aligned}
$$

In the above derivation, note that $p(\mathbf{m}, t|do(\mathbf{m}))p'(\mathbf{m}|t, \mathbf{z})$ essentially represents a particular grouping of terms whose products equals $p(t, \mathbf{x})$ (e.g., in a DAG without hidden variables both terms would be products of conditionals of the form $p(v|pa(v))$). Since the integrand of the expectation is not a function of $\mathbf{M}$, we have the independence in the intervened graph, and we constructed $\mathbb{P}$ such that there are no backdoor paths from $\mathbf{M}$ to T, we can push the associated expectations inwards and replace them with the q' terms that recover the joint density q. To see this in the context of a particular graph see the front-door graph section of the supplement. □

Theorem 5. *The surgery estimator is optimal amongst the set of directly transportable statistical or causal relations for predicting T.*

Proof. First consider the set of directly transportable statistical relations for predicting T. These will be observational distributions (i.e., no *do* terms) of the form $P(T|\mathbf{Z})$ for $\mathbf{Z} \subseteq \mathbf{X}$. We have already shown that stable conditioning sets correspond to conditional interventional distributions of the form $P(T|\mathbf{W}, do(\mathbf{M})$ in Lemma 1 and Corollary 1. Thus, we only need to consider directly transportable causal relations (stable conditional interventional distributions). However, Algorithm 2 is exactly an exhaustive search over stable conditional interventional distributions that returns the optimal one, thus the surgery estimator is optimal amongst directly transportable statistical and causal relations. □

B.4 Front-door Graph

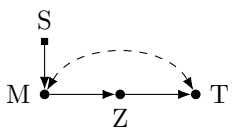

Figure 1: The front-door ADMG.

Consider the selection ADMG in Fig 1. Notably, for this graph there is no stable graph pruning estimator for predicting T. Conditioning on either $\mathbf{Z}$ or $\mathbf{M}$ activates the path $S \to M \leftrightarrow T$, and conditioning on nothing

leaves the path $S \to M \to Z \to T$ active, so there is no stable conditioning set (including the empty set). The full conditional surgery estimator, however, is identified and stable:

$$P(T|do(M), Z) = \sum_{m'} P(T|m', Z)P(m').$$

Note that this distribution is not a function of M as it has been marginalized out. This encodes the constraint that $T \perp\!\!\!\perp M|Z$ in $\mathcal{G}_{\overline{M}}$, the graph in which the edges into M are deleted. We see that in the front-door graph, after intervening on M the only relationship between M and T is via the directed chain $M \to Z \to T$. Thus Z mediates all of the effect of M on T, and the conditional interventional distribution, once computed, is not a function of M.

We can use this example to demonstrate how the proof of Theorem 4 works when interventional distributions are different from observational distributions. Now consider a distribution $\mathbb{Q}$ from the family Γ that corresponds to this graph. The density factorizes as $q(T, Z, M) = q(T|Z, M)q(Z|M)q(M)$. We will construct a new member of the family $\mathbb{P}$ such that $p(T, Z, M) = p'(T|Z)q(Z|M)q(M)$ where $p'(T|Z) = p'(T|Z, M) = \int_{m'} q(T|Z, m')q(m')dm'$. While the factorization looks different, $\mathbb{P}$ is simply a member of Γ that corresponds to the chain without unobserved confounding. Let $f_s(z) = f_s(z, m) = E[T|do(m), z]$ (not a function of M). Now consider some function $f(z, m) \in \mathcal{C}^0$:

$$\begin{aligned}
\int (t - f(z,m))^2 d\mathbb{P} &= \int_{t,z,m} (t - f(z,m))^2 p(t,z,m) dt dz dm \\
&\geq \int_{t,z,m} (t - E[T|z,m])^2 p'(t|z,m) p(z|m) p(m) dt dz dm \\
&= \int_{t,z,m} (t - E[T|z,do(m)])^2 p'(t|z) p(z|m) p(m) dt dz dm \\
&= \int_{t,z,m} (t - f_s(z,m))^2 q(t|z,do(m)) q(z|m) q(m) dt dz dm \\
&= \int_{t,z,m} (t - f_s(z))^2 q(t|z,do(m)) q(z|m) q(m) dt dz dm \\
&= \int_{t,z,m} (t - f_s(z))^2 q(t|z,do(m)) q(z) q(m|z) dt dm dz \\
&= \int_{t,z} (t - f_s(z))^2 q(z) \Big(\int_{m'} q(t|z,m') q(m') dm' \Big) \Big(\int_{m} q(m|z) dm \Big) dz dt \\
&= \int_{t,z} (t - f_s(z))^2 q(z) \Big(\int_{m'} q(t|z,m') q(m') dm' \Big) \Big(\int_{m} q(m|t,z) dm \Big) dz dt \\
&= \int_{t,z,m} (t - f_s(z))^2 q(t,m,z) dm dz dt \\
&= \int (t - f_s(z))^2 d\mathbb{Q}
\end{aligned}$$

C Experiment Details

C.1 Hyperparameters for Baselines

Causal transfer learning (CT) has hyperparameters dictating how much data to use for validation, the significance level, and which hypothesis test to use. In all experiments we set `valid split` = 0.6, `delta`=0.05, and `use hsic` = `False` (using HSIC did not improve performance and was much slower).

Anchor regression requires an "anchor" variable. In the real data experiment we use season as the anchor. It also has a hyperparameter which dictates the magnitude of perturbation shifts it protects against. We set this to twice the maximum standard deviation of any variable in the training data (including the target).

C.2 Simulated Experiment

We generate data from linear Gaussian structural equation models (SEMs) defined by the DAG in Figure 1a:

$$
\begin{aligned}
K &\sim \mathcal{N}(0, \sigma^2) \\
T &\sim \mathcal{N}(w_1 K, \sigma^2) \\
A &\sim \mathcal{N}(w_2, \sigma^2) \\
C &\sim \mathcal{N}(w_3 T + w_4 A, \sigma^2)
\end{aligned}
$$

We generate the coefficients $w_1, w_2, w_3, w_4 \sim \mathcal{N}(0, 1)$ and take $\sigma^2 = 0.1^2$.

In simulated experiment 1, A is the mutable variable so across source and target environments we vary the value of w_2. Similarly, in experiment 2 (target shift) T is the mutable variable so we vary the value of w_1.

We perform both experiments as follows: In each environment we sample 1000 examples. We generate coefficients $w_1, w_2, w_3, w_4 \sim \mathcal{N}(0, 1)$, and take 1000 samples. This is used as the training data for Graph Surgery. Then we generate 1000 samples for each of 9 other randomly generated values of w_2 or w_1 for experiments 1 and 2, respectively. The 10,000 total samples from 10 environments are used to train the OLS and CT baselines. Then we evaluate on 1000 samples from each of 100 test environments. The w_2 (or w_1) values are taken from an equally spaced grid. For experiment 1 we consider in $w_2 \in [-100, 100]$ while for experiment 2 we consider $w_1 \in [-10, 10]$. This process is repeated 500 times to yield results on 50,000 test environments.

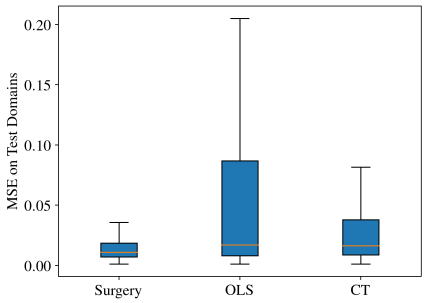

Figure 2: Boxplot of MSE in test environments for the Fig 1a scenario.

The boxplot of the test environment MSEs across the 50,000 test environments for Experiment 1 is shown in Figure 2. In this example, Surgery is the only consistently stable model. CT is stable when it selects the empty conditioning set, but in 70% of the 500 runs CT picks all features (i.e., it is equivalent to OLS). We see that the two (at least sometimes) stable methods have much lower variance in performance. Thus, stability implies less variance across environments which is desirable in the proactive transfer setting.

The boxplot of the test environment MSEs across the 50,000 test environments for Experiment 2 is shown in Figure 3. In this example, Surgery is the only consistently stable model. CT has no stable conditioning set. In 60% of runs CT conditioned on all features. The other times it tended to use the empty set. However, in this experiment $P(T)$ is not stable and uses less information than $P(T|A, C)$ (which OLS models) which is what causes it to have worse performance than OLS. Thus, even in the challenging target shift scenario, graph surgery allows us to estimate a stable model when no stable pruning or conditional model exists.

References

Jaber, A., Zhang, J., and Bareinboim, E. (2018). Causal identification under markov equivalence. In *Uncertainty in Artificial Intelligence*.

Richardson, T. S., Evans, R. J., Robins, J. M., and Shpitser, I. (2017). Nested markov properties for acyclic directed mixed graphs. *arXiv preprint arXiv:1701.06686*.

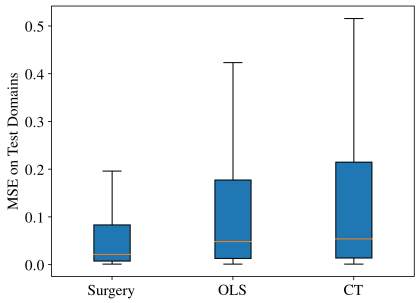

Figure 3: Boxplot of MSE in test environments for the target shift scenario.

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. (2018). Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36).

Shpitser, I. and Pearl, J. (2006). Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1219.

Shpitser, I. and Pearl, J. (2008). Dormant independence. In *Proceedings of the 23rd national conference on Artificial intelligence-Volume 2*, pages 1081–1087. AAAI Press.

Tian, J. (2002). *Studies in Causal Reasoning and Learning*. PhD thesis, University of California, Los Angeles.

Tian, J. and Pearl, J. (2002). A general identification condition for causal effects. In *AAAI*.