Preventing Failures Due to Dataset Shift: Learning Predictive Models That Transport

Adarsh Subbaswamy
Johns Hopkins University

Peter Schulam
Johns Hopkins University

Suchi Saria
Johns Hopkins University

Abstract

Classical supervised learning produces unreliable models when training and target distributions differ, with most existing solutions requiring samples from the target domain. We propose a proactive approach which learns a relationship in the training domain that will generalize to the target domain by incorporating prior knowledge of aspects of the data generating process that are expected to differ as expressed in a causal selection diagram. Specifically, we remove variables generated by unstable mechanisms from the joint factorization to yield the Surgery Estimator—an interventional distribution that is invariant to the differences across environments. We prove that the surgery estimator finds stable relationships in strictly more scenarios than previous approaches which only consider conditional relationships, and demonstrate this in simulated experiments. We also evaluate on real world data for which the true causal diagram is unknown, performing competitively against entirely data-driven approaches.

1 INTRODUCTION

As machine learning systems are increasingly deployed in practice, system developers are being faced with deployment environments that systematically differ from the training environment. However, models are typically evaluated by splitting a single dataset into train and test subsets such that training and evaluation data are, by default, drawn from the same distribution. When evaluated beyond this initial dataset, say in the deployment environment, model performance may significantly deteriorate and potentially cause harm in safety-critical applications such as healthcare (see e.g., Schulam and Saria (2017); Zech et al. (2018)). Because access to deployment environment data may not be available during training, it is not always feasible to employ *domain adaptation* techniques to directly optimize the model for the target domain. This motivates the need for *proactive* approaches which anticipate and address the differences between training and deployment environments without using deployment data (Subbaswamy and Saria, 2018). As a step towards building more reliable systems, in this paper we address the problem of proactively training models that are robust to expected changes in environment.

In order to ensure robustness, we must first be able to identify the sources of the changes. One way to do this is to reason about the differences in the underlying data generating processes (DGPs) that produce the data. For example, suppose we wish to diagnose a target condition T, say lung cancer, using information about patient chest pain symptoms C and whether or not they take aspirin A. From our prior knowledge of the DGP we know that lung cancer leads to chest pain and that aspirin can relieve chest pain. We also know that smoking K (unrecorded) is a risk factor for both lung cancer and heart disease, and aspirin is prescribed to smokers as a result. A diagnostic tool for this problem will be trained from one dataset before being deployed in hospitals that may not be represented in the data. Still, a modeler can reason about which aspects of the DGP are likely to differ across hospitals. For example, while the effects of lung cancer or aspirin on chest pain will not vary across hospitals, the policy used to prescribe aspirin to smokers (i.e., $P(A|K)$) is practice dependent and will vary.

What can the modeler do after identifying potential sources of unreliability in the data? Because the modeler does not know which prescription policies will be in place at deployment locations or by how much the deployment DGP will differ from the training DGP, the modeler should design the system to be *stable*

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

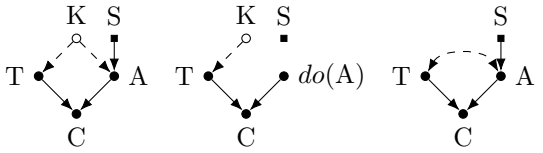


Figure 1: (a) Selection diagram for the diagnosis example. K is unobserved. (b) DAG after performing graph surgery. (c) The ADMG yielded by taking the latent projection of (a).

(i.e., invariant) to the differences in prescription policy. This means the model should predict using only the pieces of the DGP that are expected to stay the same across environments while not learning relationships that make use of the varying parts of the DGP. If the model predictions somehow depend on the prescription policy, then when the deployment policy strongly differs from the training policy model performance will significantly degrade and aspirin-taking subpopulations will be systematically misclassified.

To ensure that a model does not make use of unreliable relationships in its predictions (i.e., relationships involving prescription policy), it helps to have a representation of the DGP that makes explicit our assumptions about the DGP and what we expect will vary. A natural representation is to use *selection diagrams* (Pearl and Bareinboim, 2011), which consist of two types of nodes: nodes representing variables relevant to the DGP (e.g., smoking K or lung cancer T) and auxiliary *selection variables* (denoted by square S nodes) which identify sources of unreliability in the DGP. For example, the selection diagram in Figure 1a represents the DGP underlying the diagnosis example, with the selection variable S pointing to the piece of the DGP we expect to vary: the aspirin prescription policy $P(A|K)$. The selection variables point to *mutable* variables that are generated by mechanisms that are expected to differ across environments such that a selection diagram represents a family of DGPs which differ only in the mechanisms that generate the mutable variables.

Checking model stability graphically using selection diagrams is straightforward: a model is stable if its predictions are independent of the selection variables. If predictions are not independent of the selection variables then they depend on the environment-varying mechanisms that do not generalize. To illustrate, suppose smoking status K is not recorded in the data (denoted by the dashed edges in Figure 1a). A discriminative model of $P(T|A, C)$ which conditions on all recorded features will be dependent on the selection variable: $P(T|A, C) \neq P(T|A, C, S)$, which indicates this distribution is unstable (i.e., it differs by environment). One solution (see e.g., Subbaswamy and Saria (2018); Magliacane et al. (2018)), which we term *graph pruning*, is to perform *feature selection* to find a stable subset of features $\mathbf{Z}$ such that the conditional distribution transfers: $P(T|\mathbf{Z}) = P(T|\mathbf{Z}, S)$. However, for the problem in Figure 1a, the only stable set is $\mathbf{Z} = \emptyset$ because by d-separation (Koller and Friedman, 2009) conditioning on either A or C activates the path $T \leftarrow K \rightarrow A \leftarrow S$, inducing a dependence between T and S. Further, in cases where the mechanism that generates T varies across environments (i.e., the *target shift* scenario in which S is a parent of T), no stable feature set exists. While without more assumptions or external data there is no stable discriminative model for such problems, in this paper we relax these limitations and propose an approach which can recover stable predictive relationships in cases where graph pruning fails.

The proposed solution, which we term the *Graph Surgery* estimator,[1] is to directly remove any possible dependence on environment-varying mechanisms by using *interventional* (Pearl, 2009) rather than observational distributions to predict. Specifically, we consider a hypothetical intervention in which for each individual the mutable variables are set to the values they were actually observed to be (i.e., $do(A)$ in our example). Graphically, the intervention $do(A)$ results in a mutilated graph (Figure 1b) in which the edges into A, including edges from selection variables, are "surgically" removed (Pearl, 1998). The resulting interventional distribution[2]

$$\begin{aligned} P_A(T|C) = P(T|C, do(A)) &\propto P(T, C|do(A)) \\ &= P(T)P(C|T, A) \end{aligned}$$

is invariant to changes in how A is generated, reflecting the "independence of cause and mechanism" (Peters et al., 2017), ensuring stability, and allowing us to use information about A and C that the graph pruning solution ($P(T)$) does not. Graph surgery can be seen as learning a predictor from an alternate, hypothetical DGP in which the mutable variables were generated by direct assignment rather than by the environment-specific mechanisms. This severs dependence on selection variables to yield a stable predictor. One challenge is that when the DAG contains hidden variables (as is common in reality), interventional distributions are not always uniquely expressible as a function of the observational training data (Pearl, 2009). To address this we use the previously derived ID algorithm (Tian and Pearl, 2002; Shpitser and Pearl, 2006b) for determining identifiability of interventional distributions.

[1] Henceforth graph surgery, surgery estimator, or surgery.

[2] We will use $p(Y|do(X))$ and $P_X(Y)$ interchangeably.

Contributions: We propose the graph surgery estimator, an algorithm for estimating stable predictive models that can generalize even when train and test distributions differ. Graph surgery depends on a causal DAG to encode prior information about how the distribution of data might change. Given this prior information, it produces a predictor that does not depend on these unreliable parts of the data generating process. We show that graph surgery relaxes limiting assumptions made by existing methods for learning stable predictors. In addition, we connect the optimality of graph surgery to recently proposed adversarial distributional robustness problems.

2 RELATED WORK

Differences between training and test distributions have been previously studied as the problem of *dataset shift* (Quiñonero-Candela et al., 2009). Many specific forms of dataset shift have been characterized by dividing the variables into the input features and the target prediction outcome. By reasoning about the causal relationship between the inputs and target, various forms of dataset shift can be defined (Storkey, 2009; Schölkopf et al., 2012) which has led to methods for tackling specific instances such as *covariate shift* (e.g., Sugiyama et al. (2007); Gretton et al. (2009)), *target shift* (Zhang et al., 2013; Lipton et al., 2018), *conditional shift* (Zhang et al., 2015; Gong et al., 2016), and *policy shift* (Schulam and Saria, 2017). Using selection diagrams we can consider complex dataset shift scenarios beyond these two variable-type settings.

One issue is that methods for addressing dataset shift have mainly been *reactive*: they make use of unlabeled data from the target domain to reweight training data during learning and optimize the model specifically for the target domain (e.g., Storkey (2009); Gretton et al. (2009)). However, if we do not have target domain data available during learning, we must instead use *proactive* approaches in which the target domain remains unspecified (Subbaswamy and Saria, 2018).

One class of proactive solutions considers bounded *distributional robustness*. These methods assume that the possible test distributions are in some way centered around the training distribution. For example, in adversarial learning Sinha et al. (2018) consider a Wasserstein ball around the training distribution. Rothenhäusler et al. (2018) assume that differences between train and test distributions are bounded magnitude shift perturbations. However, these methods fail to give robustness guarantees on perturbations that are beyond the prespecified magnitude used during training. In safety-critical applications where preventing failures is crucial, we require unbounded invariance to perturbations which motivates the use of causal-based methods (Meinshausen, 2018).

To achieve stable models with complete invariance to perturbations, *graph pruning* methods consider a feature selection problem in which the goal is to find the optimal subset that makes the target independent from the selection variables. Rojas-Carulla et al. (2018) and Magliacane et al. (2018) accomplish this by empirically determining a stable conditioning set by hypothesis testing the stability of the set across multiple source domains and assuming that the target variable is not generated by a varying mechanism (no $S \rightarrow T$ edge). Extending this, Subbaswamy and Saria (2018) consider also adding *counterfactual* variables to stable conditioning sets which allow the model to make use of more stable information than by using observed variables alone. However, this requires the strong parametric assumption that causal mechanisms are linear. By using interventional distributions rather than counterfactuals, graph surgery is able to relax this assumption and nonparametrically use more stable information than observational conditional distributions. Additionally, graph surgery allows for the target to be generated by a varying mechanism.

3 METHODS

Our goal is to find a predictive distribution that generalizes even when train and test distributions differ. Derivation of the surgery estimator requires explicitly reasoning about the aspects of the DGP that can change and results in an interventional distribution in which the corresponding terms have been deleted from the factorization of the training distribution. In Section 3.1 we introduce requisite prior work on identifying interventional distributions before presenting the surgery estimator in Section 3.2 and establishing its soundness and completeness in Section 3.3.

3.1 Preliminaries

Notation: Throughout the paper sets of variables are denoted by bold capital letters while their particular assignments are denoted by bold lowercase letters. We will consider graphs with directed or bidirected edges (e.g., $\leftrightarrow$). Acyclic will be taken to mean that there exists no purely directed cycle. The sets of parents, children, ancestors, and descendants in a graph $\mathcal{G}$ will be denoted by $pa_{\mathcal{G}}(\cdot)$, $ch_{\mathcal{G}}(\cdot)$, $an_{\mathcal{G}}(\cdot)$, and $de_{\mathcal{G}}(\cdot)$, respectively. Our focus will be causal DAGs whose nodes can be partitioned into sets $\mathbf{O}$ of observed variables, $\mathbf{U}$ of unobserved variables, and $\mathbf{S}$ of selection variables. $\mathbf{O}$ and $\mathbf{U}$ consist of variables in the DGP, while $\mathbf{S}$ are auxiliary variables that denote mechanisms of the DGP that vary across environments.

Interventional Distributions: We now build up to the Identification (ID) algorithm (Tian and Pearl, 2002; Shpitser and Pearl, 2006b), a sound and complete algorithm (Shpitser and Pearl, 2006b) for determining whether or not an interventional distribution is identifiable, and if so, its form as a function of observational distributions. The ID algorithm operates on a special class of graphs known as *acyclic directed mixed graph* (ADMGs). Any hidden variable DAG can be converted to an ADMG by taking its *latent projection* onto $\mathbf{O}$ (Verma and Pearl, 1991). In the latent projection $\mathcal{G}'$ of a DAG $\mathcal{G}$ over observed variables $\mathbf{O}$, for $O_i, O_j \in \mathbf{O}$ there is an edge $O_i \to O_j$ if there exists a directed path from O_i to O_j in $\mathcal{G}$ where all internal nodes are unobserved, and $O_i \leftrightarrow O_j$ if there exists a divergent path from O_i to O_j (e.g., $O_i \leftarrow U \to O_j$) in $\mathcal{G}$ such that all internal nodes are unobserved. The bidirected edges represent *unobserved confounding*. Figure 1c shows the latent projection of the DAG in Figure 1a. The joint distribution of an ADMG factorizes as:

$$P(\mathbf{O}) = \sum_{\mathbf{U}} \prod_{O_i \in \mathbf{O}} P(O_i|pa(O_i))P(\mathbf{U}). \quad (1)$$

An intervention on $\mathbf{X} = \mathbf{O} \setminus \mathbf{V}$ sets these variables to constants $do(\mathbf{x})$. As constants, $P(x|do(x)) = 1$ such that $\prod_{X_i \in \mathbf{X}} P(X_i|pa(X_i))$ are deleted from (1) to yield the interventional distribution:

$$P_{\mathbf{X}}(\mathbf{V}) = \sum_{\mathbf{U}} \prod_{V_i \in \mathbf{V}} P(V_i|pa(V_i))P(\mathbf{U}).$$

Graphically, the intervention results in the mutilated graph $\mathcal{G}_{\overline{\mathbf{X}}}$ in which the edges into $\mathbf{X}$ have been removed.[3] When ADMG $\mathcal{G}$ contains bidirected edges, interventional distributions are not always *identifiable*.

Definition 1 (Causal Identifiability)**.** *For disjoint variable sets $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{O}$, the effect of an intervention $do(\mathbf{x})$ on $\mathbf{Y}$ is said to be identifiable from P in $\mathcal{G}$ if $P_{\mathbf{x}}(\mathbf{Y})$ is (uniquely) computable from $P(\mathbf{O})$ in any causal model which induces $\mathcal{G}$.*

The ID algorithm (a version of it is shown in Appendix A) determines if a particular interventional distribution is identified. Specifically, given disjoint variable sets $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{O}$ and an ADMG $\mathcal{G}$, a function call to $\mathrm{ID}(\mathbf{X}, \mathbf{Y}; \mathcal{G})$ returns an expression (in terms of $P(\mathbf{O})$) for $P_{\mathbf{X}}(\mathbf{Y})$ if it is identified, otherwise it throws a failure exception. The ID algorithm is nonparametric, so the terms in the expression it returns can be learned from training data with arbitrary black box approaches.

In Shpitser and Pearl (2006a), the ID algorithm was extended to answer *conditional effect* queries of the form $P_{\mathbf{X}}(\mathbf{Y}|\mathbf{Z})$ for disjoint sets $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subset \mathbf{O}$ by showing that every conditional ID query can be reduced to an unconditional ID query using the procedure shown in Algorithm 1. This procedure finds the maximal subset of variables in the conditioning set $\mathbf{Z}$ to bring into the intervention set using Rule 2 of *do*-calculus (action/observation exchange) (Pearl, 2009, Chapter 3). The resulting conditional interventional distribution is then proportional to the joint distribution of $\mathbf{Y}$ and the remaining variables in the conditioning set. A call to ID can then determine the identifiability of the resulting unconditional query.

Algorithm 1: Unconditional Query: $\mathrm{UQ}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}; \mathcal{G})$

input : ADMG $\mathcal{G}$, disjoint variable sets $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subset \mathbf{O}$
output: Unconditional query $\propto P_{\mathbf{X}}(\mathbf{Y}|\mathbf{Z})$.
$\mathbf{X}' = \mathbf{X}$; $\mathbf{Y}' = \mathbf{Y}$; $\mathbf{Z}' = \mathbf{Z}$;
while $\exists Z \in \mathbf{Z}$ *s.t.* $(\mathbf{Y} \perp\!\!\!\perp Z|\mathbf{X}, \mathbf{Z} \setminus \{Z\})_{\mathcal{G}_{\overline{\mathbf{X}},\underline{\mathbf{Z}}}}$, **do**
  $\mathbf{X}' = \mathbf{X}' \cup Z$;
  $\mathbf{Z}' = \mathbf{Z}' \setminus \{Z\}$;
$\mathbf{Y}' = \mathbf{Y} \cup \mathbf{Z}'$;
return $\mathbf{X}', \mathbf{Y}'$ of unconditional query $P_{\mathbf{X}'}(\mathbf{Y}')$

Transportability: *Transportability* is a framework for the synthesis of experimental and observational data from multiple environments to answer a statistical or causal query in a prespecified target environment (Pearl and Bareinboim, 2011; Bareinboim and Pearl, 2012). In order to build safe and reliable models, we restrict our attention to learning predictive models that can be *directly transported* from the source environment to an unspecified target environment without any adjustment.

Definition 2 (Selection diagram)**.** *A selection diagram is a causal DAG or ADMG augmented with auxiliary selection variables $\mathbf{S}$ (denoted by square nodes) such that for $S \in \mathbf{S}, X \in \mathbf{O} \cup \mathbf{U}$ an edge $S \to X$ denotes the causal mechanism that generates X may vary arbitrarily in different environments. Selection variables may have at most one child.*

We refer to the children of $\mathbf{S}$ as *mutable* variables. Selection diagrams define a family of distributions over environments such that $P(X|pa(X)), \forall X \in ch(\mathbf{S})$ in (1) can differ arbitrarily in each environment. Constructing a selection diagram generally requires domain knowledge to specify the mechanisms and the placement of selection variables. Without prior knowledge *causal discovery* methods can potentially be used (Spirtes et al., 2000).

We now define *stability* as a predictive analog of *direct transportability* (Pearl and Bareinboim, 2011), in which a source environment relationship holds in the

[3] Similarly, $\mathcal{G}_{\underline{\mathbf{X}}}$ will denote a mutilated graph in which edges out of $\mathbf{X}$ are removed.

target environment without adjustment.

Definition 3 (Stable estimator)**.** *An estimator for predicting a variable T is said to be stable if it is independent of all $S \in \mathbf{S}$.*

Graph pruning and graph surgery can both produce stable estimators, but pruning estimators will always be observational conditional distributions while surgery estimators will be the identified form of an interventional distribution.

3.2 The Graph Surgery Estimator

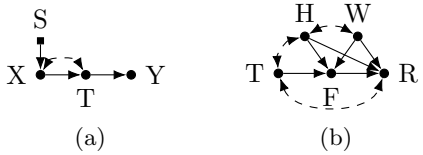
Graph surgery assumes the data modeler has constructed or been given a causal DAG of the DGP with target prediction variable T, observed variables $\mathbf{O}$, and unobserved variables $\mathbf{U}$ that has been augmented with selection variables $\mathbf{S}$ using prior knowledge about mechanisms that are expected to differ across environments (e.g., prescription policy). An overview of the procedure is as follows: The selection DAG is converted to a selection ADMG so it is compatible with the ID algorithm. Children of $\mathbf{S}$ in the selection ADMG form the set of mutable variables $\mathbf{M}$. The proposed algorithm then searches all possible interventional distributions (which intervene on $\mathbf{M}$) for the optimal (with respect to held-out source environment data) identifiable distribution, which is normalized and returned as the surgery estimator. We now cover each step in detail.

Only observed variables can be intervened on, so to determine $\mathbf{M}$, we take the latent projection of the selection DAG $\mathcal{H}$ to turn it into an ADMG $\mathcal{G}$. If a selection variable $S \in \mathbf{S}$ has multiple children in $\mathcal{G}$, then S should be split into multiple selection variables, one per child, with the new selection variables added to $\mathbf{S}$. Any disconnected variables in $\mathbf{S}$ can be removed. The mutable variables are then given by $\mathbf{M} = ch_{\mathcal{G}}(\mathbf{S}) \subseteq \mathbf{O}$. We now establish that intervening on (at least) $\mathbf{M}$ results in a stable estimator.

Proposition 1. *For $\mathbf{Y} \subseteq \mathbf{O}$, $\mathbf{X} \supseteq \mathbf{M}$ such that $\mathbf{Y} \cap \mathbf{X} = \emptyset$, the interventional distribution $P_{\mathbf{X}}(\mathbf{Y})$ is stable.*

Proof. The intervention $do(\mathbf{X})$ results in the graph $\mathcal{G}_{\overline{\mathbf{X}}}$ in which all edges into $\mathbf{X}$ are removed. Since $\mathbf{X} \supseteq \mathbf{M}$ and $\mathbf{M} = ch(\mathbf{S})$, the intervention removes all edges out of $\mathbf{S}$. This means $\mathbf{S}$ is disconnected (and thus d-separated) from $\mathbf{Y}$ in $\mathcal{G}_{\overline{\mathbf{X}}}$ which gives us stability. □

What interventional distribution should we use to predict T? A natural idea is to use the full conditional interventional distribution $P_{\mathbf{M}}(T|\mathbf{O}\setminus(\mathbf{M}\cup\{T\}))$ which can be turned into a corresponding unconditional query (so we can call ID) using a call to Algorithm 1: $\text{UQ}(\mathbf{M}, T, \mathbf{O} \setminus (\mathbf{M} \cup \{T\}); \mathcal{G})$. However, this has two issues. First, if the target variable is mutable itself ($T \in \mathbf{M}$) then the conditional interventional distribution is ill-defined since the three variable sets must be disjoint. If T is mutable, then we must intervene on it, graphically represented by deleting all edges into T. Variables related to T through edges out of T (e.g., children and their bidirected neighborhoods) can still be used to predict T. Thus, if $T \in \mathbf{M}$, we can generate an unconditional query of the form $P_{\mathbf{X}}(\mathbf{Y})$ from $\text{UQ}(\mathbf{M} \setminus \{T\}, T, \mathbf{O} \setminus \mathbf{M}; \mathcal{G}_{\overline{T}})$, noting that we are using the mutilated graph $\mathcal{G}_{\overline{T}}$. We must further modify the result to account for the fact that we are also intervening on T: $P_{\mathbf{X}\cup\{T\}}(\mathbf{Y} \setminus \{T\})$. Importantly, there is never a stable pruning estimator when $T \in \mathbf{M}$ which shows that graph surgery can provide stability in cases where existing pruning solutions cannot.



(a) (b)

Figure 2: (a) Selection ADMG which requires intervening on T. (b) Causal ADMG for Bike Sharing.

Second, the full conditional interventional distribution may not be identifiable. We propose an exhaustive search over possible conditioning sets: trying each $P_{\mathbf{M}}(T|\mathbf{Z})$ for $\mathbf{Z} \in \mathcal{P}(\mathbf{W})$, $\mathbf{W} = \mathbf{O} \setminus (\mathbf{M} \cup \{T\})$ where $\mathcal{P}(\cdot)$ denotes the power set. In the interest of identifiability, even if $T \notin \mathbf{M}$ we may want to consider intervening on T.[4] For example, in Figure 2(a), $P_X(T|Y)$ and $P_X(T)$ are not identifiable, but $P_{X,T}(Y)$ is. Thus, we should consider the unconditional query returned by Algorithm 1 in both $\mathcal{G}$ and $\mathcal{G}_{\overline{T}}$ (with the modification of moving T to the intervention set). The full procedure is given as Algorithm 2. Note that it returns the estimator that performs the best on held out source-environment validation data with respect to some loss function ℓ. If there is no identifiable interventional distribution the Algorithm throws a failure exception.

3.3 Soundness and Completeness

Algorithm 2 is sound in that it only returns stable estimators and complete in that it finds a stable surgery estimator if one exists. Proofs are in the supplement.

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

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

[4] Deleting edges in a graph generally helps identifiability (Pearl, 2009).

Algorithm 2: Graph Surgery Estimator

input : ADMG $\mathcal{G}$, mutable variables $\mathbf{M}$, target T
output: Expression for the surgery estimator or `FAIL` if there is no stable estimator.

Let $S_{ID} = \emptyset$; Let $Loss = \emptyset$;
for $\mathbf{Z} \in \mathcal{P}(\mathbf{O} \setminus (\mathbf{M} \cup \{\mathbf{T}\}))$ **do**
  if $T \notin \mathbf{M}$ **then**
    Let $\mathbf{X}, \mathbf{Y} = \text{UQ}(\mathbf{M}, \{T\}, \mathbf{Z}; \mathcal{G})$;
    try
      $P = \text{ID}(\mathbf{X}, \mathbf{Y}; \mathcal{G})$;
      $P_s = P / \sum_T P$;
      Compute validation loss $\ell(P_s)$;
      S_{ID}.append(P_s); $Loss$.append($\ell(P_s)$);
    catch
      pass;
  Let $\mathbf{X}, \mathbf{Y} = \text{UQ}(\mathbf{M}, \{T\}, \mathbf{Z}; \mathcal{G}_{\overline{T}})$;
  $\mathbf{X} = \mathbf{X} \cup \{T\}$; $\mathbf{Y} = \mathbf{Y} \setminus \{T\}$;
  if $\mathbf{Y} \cap (T \cup ch(T)) = \emptyset$ **then**
    continue;
  try
    $P = \text{ID}(\mathbf{X}, \mathbf{Y}; \mathcal{G})$;
    $P_s = P / \sum_T P$;
    Compute validation loss $\ell(P_s)$;
    S_{ID}.append(P_s); $Loss$.append($\ell(P_s)$);
  catch
    continue;
if $S_{ID} = \emptyset$ **then**
  return `FAIL`;
return $P_s \in S_{ID}$ with lowest corresponding $Loss$;

4 CONNECTIONS WITH EXISTING APPROACHES

We establish connections between graph surgery and existing proactive approaches, showing that graph pruning (which finds stable conditional relationships) is a special case of surgery and that surgery has an optimal distributionally robust interpretation.

4.1 Relationship with Graph Pruning

We show that graph pruning estimators are in fact surgery estimators, so graph surgery does not fail on problems graph pruning can solve.

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. We show $\exists \mathbf{W} \subseteq \mathbf{Z}$ s.t. $P(T|\mathbf{Z}) = P_{\mathbf{M}}(T|\mathbf{W})$. See supplement. □

In the proof of Lemma 1 we derived that graph pruning is a special case of graph surgery:

Corollary 1. *Graph pruning estimators are graph surgery estimators since they can be expressed as conditional interventional distributions.*

Lemma 2. *There exists a problem for which graph pruning cannot find a non-empty stable conditioning set but for which graph surgery does not fail.*

Proof. As one such example, see Figure 1(c). □

From the previous two Lemmas the following Corollary is immediate:

Corollary 2. *There exists a stable graph surgery estimator for a strict superset of the problems for which there exists a stable graph pruning estimator.*

We have now shown that graph surgery strictly generalizes graph pruning.

4.2 Surgery As Distributional Robustness

We now discuss the optimality of the surgery estimator for adversarial transfer problems in the presence of unstable mechanisms.

Suppose selection ADMG $\mathcal{G}$ defines a prediction problem with target variable T and input features $\mathbf{X} = \mathbf{O} \setminus \{T\}$. Further suppose all variables are continuous such that the prediction problem is regression and that we use the L_2 loss. Under classical assumptions that training and test distributions are the same, our goal is to learn a function $f(\mathbf{x})$ that minimizes the expected (squared) loss or *risk*: $E_{\mathbf{o} \sim P(\mathbf{O})}[(t - f(\mathbf{x}))^2]$.

In our setting, however, $P(\mathbf{O})$ varies across domains. Recall that a selection diagram defines a family of distributions Γ over $\mathbf{O}$ such that for any particular domain (i.e., setting of $\mathbf{S}$) there exists a $Q_{\mathbf{s}} \in \Gamma$ such that $P(\mathbf{O}|\mathbf{s})$ factorizes according to (1) and members of Γ differ in $\prod_{W \in \mathbf{M}} Q_{\mathbf{s}}(W|pa(W))$. As opposed to the classical setting, now our goal is to learn a predictor that optimizes for loss across the distributions in Γ. When constructing reliable models for safety-critical applications in which model failures can be dangerous, a natural choice is to minimize the worst-case or minimax risk across the environments. This can be written as the following game in which we seek the optimal f from the set of continuous functions $\mathcal{C}^0$:

$$\min_{f \in \mathcal{C}^0} \sup_{Q_{\mathbf{s}} \in \Gamma} E_{Q_{\mathbf{s}}}[(t - f(\mathbf{x}))^2]. \tag{2}$$

We now give two sufficient conditions under which using the surgery estimator is optimal in that $f_s(\mathbf{x}) =$

$E[T|\mathbf{x} \setminus \mathbf{m}, do(\mathbf{m})]$ achieves (2). Proofs are in the supplement.

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 achieves (2).*

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 achieves (2).*

In the case of Theorem 3, the surgery estimator reduces to using an invariant subset of features to predict and the result follows from the optimality of graph pruning methods for distributions which share the invariant conditional (Rojas-Carulla et al., 2018, Theorem 4). Theorem 4, however, can correspond to certain cases in which $P_{\mathbf{M}}(T|\mathbf{X} \setminus \mathbf{M})$ is not an observational conditional distribution. One example is the so-called *front-door* graph (Pearl, 2009), in which there exists no stable pruning estimator while the surgery estimator is optimal across all predictors. Further discussion of this case is in the supplement.

We finally discuss the surgery estimator in the context of stable estimators. Stability is a particularly desirable property because, assuming the estimator is identified, stable estimators can be learned using data from any environment in Γ. Further, we have the following (proof in supplement):

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

We have discussed cases in which the surgery estimator is minimax optimal across the environments in Γ and established that the surgery estimator is the optimal stable predictor. In the context of Theorem 5, this means that without additional assumptions (e.g., parametric assumptions about forms of causal mechanisms or assumptions about the magnitude of differences across environments) the surgery estimator provides the best method for prediction that can be trained from and used in any environment.

5 EXPERIMENTS

We evaluate the graph surgery estimator in proactive transfer settings in which data from the target distribution is unavailable. The goal of our experiments is to demonstrate that the surgery estimator is stable in situations in which existing methods are either not applicable or suboptimal. To this end, we first consider a simulated experiment for which the true selection diagram is known. Then we apply the surgery estimator to real data to demonstrate its practical utility even when the true selection diagram is unknown. We compare against a naive pooled ordinary least squares baseline (OLS) and causal transfer learning (CT), a state-of-the-art pruning approach (Rojas-Carulla et al., 2018).[5] If CT fails to find a non-empty subset, we predict using the pooled source data mean. On real data we also compare against Anchor Regression (AR), a distributionally robust method for bounded magnitude shift interventions (Rothenhäusler et al., 2018) which requires a causal "anchor" variable with no parents in the graph. All performance is measured using mean squared error (MSE).

5.1 Simulated Data

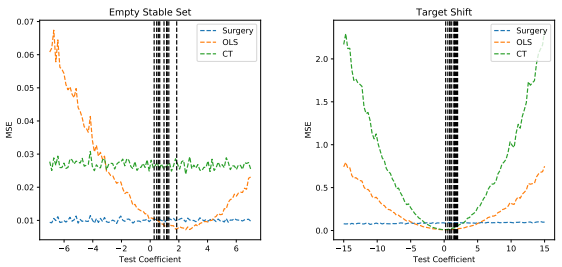
We simulate data from zero-mean linear Gaussian systems using the DAG in Figure 1(a) considering two variations (full details in supplement).[6] The first considers the selection problem in Figure 1(a) in which A is a mutable variable, defining a family of DGPs which vary in the coefficient of K in the structural equation for A. We generate 10 source environments and apply on test environments in which we vary the coefficient on a grid. Recall that in this DAG the empty set is the only stable conditioning set and CT should model $P(T)$. While this is stable, we expect the performance to be worse than that of the surgery estimator: $P_s \propto P_A(T, C) = P(C|T, A)P(T)$ which is able to use additional stable factors.

The MSE as we vary the test coefficient of K is shown in Figure 3a. As expected, the stable models CT and Surgery are able to generalize beyond the training environments (vertical dashed lines), while the unstable OLS loss grows quickly. However, for small deviations from the training environment OLS outperforms the stable methods which shows that there is a tradeoff between stability and performance in and near the training environment. The gap between CT and Surgery is expected since Surgery models an extra stable, informative factor: $P(C|T, A)$.

We repeat this experiment but consider the target shift scenario in which T is the mutable variable, and the DGPs across environments differ in the coefficient of K in the structural equation for T. Now there is no stable conditioning set which violates the assumption of CT. Again, CT used the empty conditioning set $P(T)$ but in this case is unstable so the loss grows quickly in Figure 3b. As before, OLS is unstable but performs best near the source environments. The surgery estimator $P_s \propto P_{TA}(C) = P(C|T, A)$ is stable and the loss appears constant compared to the unstable alternatives. These experiments demonstrate that stability is an important property when differences in mechanisms can

[5] `https://github.com/mrojascarulla/causal_transfer_learning`

[6] This DAG contains no anchor so we cannot apply AR.



(a) (b)

Figure 3: (a) MSE in test environments for the Fig 1a scenario. (b) MSE in test environment for target shift scenario. Vertical lines denote training environments.

be arbitrarily large. In the supplement we aggregate results for many repetitions of the simulations.

5.2 Real Data: Bike Rentals

Following Rothenhäusler et al. (2018) we use the UCI Bike Sharing dataset (Fanaee-T and Gama, 2013; Dheeru and Karra Taniskidou, 2017) in which the goal is to predict the number of hourly bike rentals R from weather data including temperature T, feeling temperature F, wind speed W, and humidity H. As in Rothenhäusler et al. (2018), we transform R from a count to continuous variable using the square root. The data contains 17,379 examples with temporal information such as season and year. We partition the data by season (1-4) and year (1-2) to create environments with different mechanisms. We posit the causal diagram in Figure 2(b) with confounding caused by unobserved temporal factors, and hypothesize that differences in season result in unstable mechanisms for weather: the mutable variables are $\mathbf{M} = \{H, T, W\}$. If this diagram is true, then no stable pruning estimator exists, so we expect surgery to outperform CT and OLS if the differences in mechanisms are large. The full conditional interventional distribution $P_{THW}(R|F) = P_{THWF}(R)$ is identified and the surgery estimator is given by $\hat{R}_s = \sum_T E[R|T, H, W, F]P(T|H, W)$. We posit linear Gaussians for each term and compute $\hat{R}_s$ using 10,000 Monte Carlo samples. Since AR and CT require data from multiple source environments, for each year (Y), we select one season as the target environment, using the other three seasons as source environments. Since OLS and Surgery do not make use of the season indicator, we simply pool the data for these methods.

We sample 80% of the training/test data 20 times and report the average MSE in Table 1 (intervals are one standard error). The surgery estimator performs competitively, achieving the lowest average MSE in 3 of 8 test cases. When the OLS MSE is high (seasons 3 and 4 in each year), Surgery tends to outperform it which we attribute to Surgery's stability. We also see that CT tends to perform poorly which lends some credibility to our hypothesized selection diagram which dictates that no stable pruning estimator exists. AR's very good performance is expected, since the shift-perturbation assumption seems reasonable in this problem. However, AR requires tuning of a hyperparameter for the maximum magnitude shift perturbation to protect against which is less preferable than stable estimators such as surgery in safety critical applications when the target environment is unknown and could be very different from the source.

Table 1: MSE on the Bike Sharing Dataset

Test Data	OLS	AR	CT	Surgery
(Y1) Season 1	20.8±0.10	**20.5**±0.10	42.2±2.04	20.7±0.36
Season 2	**23.2**±0.05	**23.2**±0.05	29.9±0.09	23.8±0.09
Season 3	32.2±0.14	31.4±0.13	32.2±0.14	**29.9**±0.26
Season 4	29.2±0.08	29.1±0.08	29.1±0.08	**28.2**±0.07
(Y2) Season 1	32.5±0.11	**32.2**±0.11	32.6±0.15	36.1±0.37
Season 2	39.3±0.11	**39.2**±0.11	46.1±0.12	39.5±0.13
Season 3	47.7±0.17	**46.7**±0.16	48.2±0.22	54.8±0.73
Season 4	46.2±0.16	46.0±0.16	46.1±0.16	**44.4**±0.16

6 CONCLUSION

Since the very act of deploying a system can result in shifts that bias the system in practice (e.g., Lum and Isaac (2016)), machine learning practitioners need to become increasingly aware of how deployment and training environments can differ. To this end, we have introduced a framework for identifying and expressing desired invariances to changes in the DGP, and the Surgery estimator as an approach for learning a model stable to such changes. The surgery estimator finds a stable and identifiable interventional distribution which is expressible as a function of the training data and can be fit using arbitrarily complex models. Further, the interventional distributions are strictly more applicable than the conditional distributions used by existing graph pruning approaches and are optimal from a distributionally robust perspective. In future work we wish to consider methods for when the selection diagram does not entail any identifiable stable predictors. In particular, some form of sensitivity analysis for dealing with uncertainty in the DGP such as infusing bounded-magnitude distributional robustness with prior knowledge of the DGP seems promising.

Acknowledgements

The authors thank Thijs van Ommen for helpful discussions about section 4.2.

References

Bareinboim, E. and Pearl, J. (2012). Transportability of causal effects: Completeness results. In *AAAI*, pages 698–704.

Dheeru, D. and Karra Taniskidou, E. (2017). UCI machine learning repository.

Fanaee-T, H. and Gama, J. (2013). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, pages 1–15.

Gong, M., Zhang, K., Liu, T., Tao, D., Glymour, C., and Schölkopf, B. (2016). Domain adaptation with conditional transferable components. In *International Conference on Machine Learning*, pages 2839–2848.

Gretton, A., Smola, A. J., Huang, J., Schmittfull, M., Borgwardt, K. M., and Schölkopf, B. (2009). Covariate shift by kernel mean matching. In Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D., editors, *Dataset shift in machine learning*, chapter 2, pages 131–160. The MIT Press.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Lipton, Z. C., Wang, Y.-X., and Smola, A. (2018). Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*.

Lum, K. and Isaac, W. (2016). To predict and serve? *Significance*, 13(5):14–19.

Magliacane, S., van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., and Mooij, J. M. (2018). Domain adaptation by using causal inference to predict invariant conditional distributions. In *Proceedings of the Thirty-Second Conference on Neural Information Processing Systems*.

Meinshausen, N. (2018). Causality from a distributional robustness point of view. In *2018 IEEE Data Science Workshop (DSW)*, pages 6–10. IEEE.

Pearl, J. (1998). Graphical models for probabilistic and causal reasoning. In *Quantified representation of uncertainty and imprecision*, pages 367–389. Springer.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J. and Bareinboim, E. (2011). Transportability of causal and statistical relations: a formal approach. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pages 247–254. AAAI Press.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. MIT press.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2009). Dataset shift in machine learning.

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

Rothenhäusler, D., Bühlmann, P., Meinshausen, N., and Peters, J. (2018). Anchor regression: heterogeneous data meets causality. *arXiv preprint arXiv:1801.06229*.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. (2012). On causal and anticausal learning. In *International Conference on Machine Learning*, pages 459–466.

Schulam, P. and Saria, S. (2017). Reliable decision support using counterfactual models. In *Advances in Neural Information Processing Systems*, pages 1697–1708.

Shpitser, I. and Pearl, J. (2006a). Identification of conditional interventional distributions. In *22nd Conference on Uncertainty in Artificial Intelligence, UAI 2006*, pages 437–444.

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

Sinha, A., Namkoong, H., and Duchi, J. (2018). Certifying some distributional robustness with principled adversarial training. In *ICLR*.

Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G., and Richardson, T. (2000). *Causation, prediction, and search*. MIT press.

Storkey, A. (2009). When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, pages 3–28.

Subbaswamy, A. and Saria, S. (2018). Counterfactual normalization: Proactively addressing dataset shift and improving reliability using causal mechanisms. In *Uncertainty in Artificial Intelligence*.

Sugiyama, M., Krauledat, M., and MÃžller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May):985–1005.

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

Verma, T. and Pearl, J. (1991). Equivalence and synthesis of causal models. In *Proceedings of Sixth Conference on Uncertainty in Artificial Intelligence*, pages 220–227.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS medicine*, 15(11):e1002683.

Zhang, K., Gong, M., and Schölkopf, B. (2015). Multi-source domain adaptation: A causal view. In *AAAI*, pages 3150–3157.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. (2013). Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827.