Research Article

Adarsh Subbaswamy*, Bryant Chen, Suchi Saria

A unifying causal framework for analyzing dataset shift-stable learning algorithms

https://doi.org/10.1515/jci-2021-0042
received August 19, 2021; accepted March 12, 2022

Abstract: Recent interest in the external validity of prediction models (i.e., the problem of different train and test distributions, known as *dataset shift*) has produced many methods for finding predictive distributions that are invariant to dataset shifts and can be used for prediction in new, unseen environments. However, these methods consider different types of shifts and have been developed under disparate frameworks, making it difficult to theoretically analyze how solutions differ with respect to stability and accuracy. Taking a causal graphical view, we use a flexible graphical representation to express various types of dataset shifts. Given a known graph of the data generating process, we show that all invariant distributions correspond to a causal hierarchy of graphical operators, which disable the edges in the graph that are responsible for the shifts. The hierarchy provides a common theoretical underpinning for understanding when and how stability to shifts can be achieved, and in what ways stable distributions can differ. We use it to establish conditions for minimax optimal performance across environments, and derive new algorithms that find optimal stable distributions. By using this new perspective, we empirically demonstrate that that there is a tradeoff between minimax and average performance.

Keywords: dataset shift, transportability, invariance, stability

MSC 2020: 68T01, 68T30, 68T37, 62C20

1 Introduction

Statistical and machine learning (ML) predictive models are being deployed in a number of high impact applications, including healthcare [1], law enforcement [2], and criminal justice [3]. These safety-critical applications have a high cost of failure – model errors can lead to incorrect decisions that have a profound impact on the quality of human lives – which makes it important to ensure that systems being developed and deployed for these problems behave *reliably* (i.e., they perform to their specification). To do so, developers are forced to reason in advance about likely sources of failure and address them prior to deployment (i.e., during model training). A key source of failure is due to *dataset shifts* [4,5]: differences between the environment in which training data were collected and the environment in which the model will be deployed that manifest as changes in the data distribution. These differences can arise due to deploying a model at a new site from which data were unavailable during training, or due to natural variations that occur over time. Failing to account for these differences can result in model predictions with worse performance (i.e., expected loss) than anticipated.

*** Corresponding author: Adarsh Subbaswamy,** Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, United States, e-mail: asubbaswamy@jhu.edu
Bryant Chen: Brex Inc, San Francisco, California, United States, e-mail: bryant@brex.com
Suchi Saria: Department of Computer Science, Johns Hopkins University and Bayesian Health, Baltimore, MD 21218, United States, e-mail: ssaria@cs.jhu.edu

Across a number of application domains, the recent COVID-19 pandemic has demonstrated ways in which dataset shifts can induce model failures. For example, the pandemic resulted in a drastic shift in online retail and the consumer packed goods industries: during the onset of the pandemic, the predictive algorithms powering Amazon's supply chain failed due to the sudden increased demand for household supplies (e.g., bottled water and paper products), resulting in unprecedented item shortages and delivery delays [6].

Beyond changes to customer behavior, dataset shift has been identified as a key challenge to ensuring reliability in safety-critical domains such as healthcare (see, e.g., the example ways in which dataset shift can occur in medical applications in ref. [5]). Consider the following examples: Long-term (e.g., 3 years) patient mortality prediction models are used to help determine which patients may need long-term support after being discharged from the hospital. In one study, the authors trained a model to predict 3-year patient mortality from electronic health record (EHR) data at a single hospital. The authors found that, for 68% of laboratory tests, the timing of the laboratory test orders was more predictive of mortality for the model than the corresponding values of those tests [7]. As a result, the model learned predictive dependencies between the time of day when a lab test was ordered and patient mortality. These dependencies are brittle: they are highly variant across hospitals because the timing of lab tests is determined by hospital-specific policies and physician-specific preferences [8,9]. Models that have learned these brittle dependencies can experience significant deterioration in performance and become unsafe to use [10].

As another example, consider [11], in which the authors trained a model to diagnose pneumonia from chest X-rays. While the model was found to be very accurate on new patients at the medical center where it was developed, this performance deteriorated significantly when applied at new, but similar, medical centers. Their analysis showed that the model had learned dependencies between stylistic features (e.g., text, orientation, coloring) present in the X-ray and pneumonia. These associations varied widely across hospitals because the choice of stylistic features depended on the X-ray equipment, hospital policies, and technician preferences.

These examples demonstrate that dataset shifts can arise from a variety of changes (i.e., interventions) in the underlying data generating process (DGP) such as changes in behavior (e.g., shifts in clinician treatment patterns) or changes in data acquisition (e.g., new X-ray machines and settings). Preventing failures due to these kinds of shifts requires a *causal* understanding of the parts of the data generating process that can shift, and learning models of *stable* distributions that are invariant to these shifts.[1]

Given that dataset shifts can happen in nearly every domain where predictive models are used, and given how serious the consequence of failures due to these shifts can be, it is critical to be able to ensure the reliability of models under such shifts: That is, we need to understand under a given set of dataset shifts, how will a model's behavior change? What can be said about a model's *stability* (i.e., are the model's predictions still accurate after the shifts)? In the X-ray example, a model developer wants to know: what shifts can lead to model instability? Would shifts in color encoding schemes or upgrades to X-ray equipment lead to instability and deteriorate model accuracy? Has the model learned any dependencies (e.g., between pneumonia and choice of equipment) that will lead to this instability? For models trained using different algorithms, what guarantees do they give about stability to shifts in color schemes? How do the accuracies of these models differ under these dataset shifts? What guarantees can be made about a model's worst-case performance under such shifts? Lacking common footing for framing stability and dataset shift, it is difficult to begin to answer these questions and compare algorithms.

A common framing of dataset shift is to assume limited data from a clearly defined "target" environment or distribution of interest is used (along with more plentiful data from a "source" environment) to make inferences about the target environment. This framing allows an analyst to "*reactively*"[2] adjust to target data samples. Reactive approaches to addressing dataset shift exist across multiple fields of study.

1 We will refer to distributions that are stable to dataset shifts as "shift-stable" or simply "stable."

2 The term "reactive" to describe approaches that use (possibly unlabeled) data from the target domain during learning was coined in ref. [21].

Some examples include methods for domain adaptation in machine learning (for an overview, see ref. [4]), generalizability and transportability in causal inference (e.g., [12–15]), and sample selection bias in statistics and econometrics (e.g., [16–19]). Reactive approaches to model training require data from the target environment, which makes it difficult to learn a model from source data alone, which will perform well in a new, unseen environment. In this case, it is important to instead use *proactive* learning approaches which learn models that are stable to any anticipated problematic shifts.

One common class of proactive learning methods is *declarative* in nature. These methods allow users to specify dependencies (i.e., causal relationships) between variables in the dataset that are likely to experience a dataset shift. That is, the data generating process is expected to differ between source and target environments due to an unknown intervention on the specified causal relationships. For instance, in the previously discussed X-ray example, a user might want to specify that the model should be stable to changes in scanner manufacturer, the choice of X-ray orientation (front-to-back vs back-to-front), or the color encoding scheme, since all of these are problematic shifts that are likely to occur. Example learning methods include approaches which find stable subsets of features [20,21], approaches which learn models under hypothetically stable data generating processes [22,23], and "counterfactual" approaches which compute counterfactual features [10,21,24] or perform data augmentation [25] to remove unstable dependencies. A key feature of declarative learning methods is that they can give guarantees about the stability of model predictions to changes in the specified shifts. When a user specifies that they desire invariance to the choice of X-ray orientation, then the declarative method will find a stable solution satisfying this specification. However, this requires domain expertise to be able to specify the likely problematic shifts to which stability is desired. A notable exception is the method of [23], which learns candidate shifts that occurred across datasets and allows users to choose which invariances to enforce. In addition, while different declarative methods can guarantee stability, it is unknown what tradeoffs exist between methods with respect to accuracy. For example, models trained using stable feature subsets vs counterfactuals can both satisfy stability to, e.g., the choice of X-ray orientation, but we currently struggle to answer how their accuracy will compare and differ under shifts in X-ray orientation preferences. While both are stable, is one more accurate under shifts than the other?

A second class of proactive methods are *imperative* in nature. These methods take in datasets collected from multiple, heterogeneous environments and automatically extract invariant predictors from the data without user input [26–29]. Examples of these methods are those that compute features sets [26] or representations [27,28] that yield invariant predictors. In the X-ray example, imperative methods would require datasets collected from a large number of health centers, which diversely represented the sets of shifts that could be observed (e.g., the datasets differ in terms of scanner manufacturer, X-ray orientation, and encoding schemes). An advantage of such approaches is that they do not require domain expertise in order to determine invariances. These methods often provide theoretical guarantees about minimax optimal performance (i.e., that they have the smallest worst-case error) across the input distributions. Thus, by using an imperative approach, a model developer can guarantee good worst-case performance at new hospitals that "look like" a mixture of the training hospitals. However, they generally do not provide guarantees about stability to a set of specified shifts: we do not know the ways in which the datasets differ (or by how much), so we cannot answer if the model is stable to shifts in scanner manufacturers, X-ray orientations, or encoding schemes.

The difficulties of understanding model behavior within and across each thread of work prevent rigorous analysis of the reliability of models under dataset shifts. In reality, we are presented with a prediction problem in which the data have been collected and generated under some DGP. Dataset shifts can then lead to changes to arbitrary pieces of the DGP. Thus, there is a need for a framework that enables us to answer the fundamental questions about model behavior under changes to the DGP. This would provide common ground to compare algorithms that address dataset shift, and to generate generalizable insights from methods that address particular instances of shifts.

1.1 Contributions

In this article, we provide a unifying framework for specifying dataset shifts that can occur, analyzing model stability to these shifts, and determining conditions for achieving the lowest worst-case error (i.e., minimax optimal performance) across environments produced by these shifts. This provides common ground so that we can begin to answer fundamental questions such as: To what dataset shifts are the model's predictions stable vs unstable? Has the model learned a predictive relationship that is stable to a set of prespecified shifts of interest? How will the model's performance be affected by these shifts? For models trained using different methods, what guarantees do they provide about stability and accuracy?

The framework centers around two key requirements: First, a known causal graphical representation of the environment data generating process which includes specifications of what can shift. These specifications take the form of marked unstable edges in the graph which represent causal dependencies between variables that can shift across environments. We consider arbitrary shifts (i.e., interventions) to causal mechanisms in the graph, as opposed to, e.g., constraining shifted distributions to be within bounded norm-balls of the training data. This specification entails commonly studied instances of dataset shift (such as label shifts, covariate shifts, and conditional shifts), but also handles the more general, unnamed shifts that we expect to see in practice. Second, we restrict our analysis to methods whose target distribution can be expressed graphically. This entails algorithms that do not learn intermediate feature representations, but instead operate directly on the observed variables. For models that do not induce a graphical representation, we discuss how our results might be used to probe these models for their stability properties and discuss opportunities for future work to bridge this gap.

Our main contribution is the development of a causal hierarchy of stable distributions, in which distributions at higher levels of the hierarchy guarantee lower worst-case error. The levels of the hierarchy provide insight into how stability can be achieved: the levels correspond to three operators on the graphical representation of the environment, which modify the graph to produce *stable distributions* – the learning targets for stable learning algorithms (Definitions 6, 7, and 8). We further show that the operators have different stability properties: higher-level operators more precisely remove unstable edges from the graph when producing stable distributions (Corollary 2). By using this graphical characterization of stability, we provide a simple graphical criterion for determining if a distribution is stable to a set of prespecified shifts: a distribution is stable if it modifies the graph to remove the corresponding unstable edges (Theorem 1). We then address questions about the accuracy of different stable solutions by showing how the hierarchy provides a causal characterization of the minimax optimal predictor: the predictor that achieves the lowest worst-case error across shifted environments (Proposition 7). Surprisingly, we find that frequently studied intervention-invariant solutions generally do not achieve this minimax optimal performance. Finally, we demonstrate through a series of semisynthetic experiments that there is a tradeoff between minimax and average performance. Through these contributions, we provide a common theoretical underpinning for understanding model behavior under dataset shifts: when and how stability to shifts can be achieved, in what ways stable distributions can differ, and how to achieve the lowest worst-case error across shifted environments.

2 Related work

While the focus of this work is on statistical and machine learning models, we briefly discuss concepts related to dataset shift that has been studied in other fields. In particular, *external validity*, or the ability of experimental findings to generalize beyond a single study, has long been an important goal in the social and medical sciences [30]. For example, practitioners (such as clinicians) who want to assess the results of a randomized trial must consider how the results of the trial relate to the target population of interest. This need has led to much discussion and work on assessing the *generalizability* of randomized trials (see, e.g., [31–33]). More recently, methodological work in causal inference has focused on *transportability* (see [15]

for a review). For example, researchers have developed causal graphical methods for determining when and how experimental findings can be transported from one population or setting to a new one (see, e.g., [12,34, 14,35]). Generalizability is also of importance in economics research [36], with much methodological work focusing on the problem of *sample selection bias* resulting from nonrandom data collection (see, e.g., [16–18]). Selection bias leads to systematic differences between the observed data and the general population, and thus is related to problems of external validity, which consider different environments or populations.

Returning to the focus of this work, the problem of differing train and test distributions in predictive modeling is known as *dataset shift* in machine learning [4]. Classical approaches, such as domain adaptation, assume access to unlabeled samples from the target distribution which they use to reweight training data during learning or extract invariant feature representations (e.g., [37–40]). More recently, work on *statistical transportability* has produced sound and complete algorithms for determining how data from multiple domains can be synthesized to compute a predictive conditional distribution in the target environment of interest [12,41]. These methods *reactively* adjust to target data. In many practical scenarios, however, it is not possible to get samples from all possible target distributions. Instead, this requires *proactive* methods that make assumptions about the set of possible target environments in order to learn a model from source data that can be applied elsewhere [21]. Work on proactive methods has primarily focused on either bounded or unbounded shifts. Bounded shifts have been studied through the lens of *distributional robustness*, often assuming shifts within a finite radius divergence ball (e.g., [42,43]). Ref. [44] consider robustness to bounded magnitude interventions in latent style features. Ref. [45] considered bounded and unbounded mean-shifts in mechanisms (in which the means of certain variables vary by environment). Ref. [46] build on this to allow for stochastic mean-shifts. In this article, we focus on unbounded shifts: in safety-critical domains the cost of failure is high, and it can be difficult to accurately specify bounds.[3]

Many proactive methods use datasets from multiple source environments to train invariant models to predict in new, unseen environments. Ref. [47] propose a kernel method to find a data transformation minimizing the difference between feature distributions across environments while preserving the predictive relationship. Invariant risk minimization (IRM) learns a representation such that the optimal classifier is invariant across the input environments [27,48]. Similar works learn invariant predictors by seeking derivative invariance across environments [28,29]. These methods learn representations, so the predictors do not induce a graphical representation. Despite this, we discuss how our graphical results can be applied to probe these methods for their stability properties. IRM establishes its ability to generalize to new environments using the framework of *invariant causal prediction* (ICP) [49]. Under ICP, the minimax optimal performance of a causally invariant predictor is tied to assuming shifts occur in all variables except the target prediction variable. In reality, however, specific shifts occur and we want to determine which ones to protect against (and how). In this work, we take this approach by starting with the data generating process, and derive stable solutions to prespecified sets of shifts.

Other proactive methods make explicit use of connections to causality. Related to ICP, [26] uses multiple source environment datasets to find a feature subset that yields an invariant conditional distribution. This has also been extended to the reactive case in which unlabeled target data are available (see also [20]). For discrete variable settings in which data from only one source environment are available and there is no unobserved confounding, covariate balancing techniques have been used to determine the causal features that yield a stable conditional distribution [50,51]. Other causal methods assume explicit knowledge of the graph representing the DGP instead of requiring multiple datasets. Explicitly assuming no unobserved confounders, [10] protects against shifts in continuous-time longitudinal settings by predicting *counterfactual* outcomes. Ref. [21] finds a stable feature set that can include counterfactual variables, assuming linear mechanisms. Ref. [24] regularizes toward counterfactual invariance so that a model learns

3 Though the focus of this article is on arbitrary shifts, we note that a "stability accuracy tradeoff" has also been observed in works on bounded distributional robustness. See [86] for an example.

to predict only using causal associations. Other works consider using counterfactuals generated by human annotation [52], data augmentation (see [25,53] for discussion of the relationship between causality and data augmentation), or active learning [54] to improve model robustness. Ref. [22] uses *selection diagrams* [12] to identify mechanisms that can shift and find a stable *interventional* distribution to use for prediction. More recently, end-to-end approaches have been developed, which relax the need for the graph to be known beforehand, instead learning it from data [23,55]. In this article, we provide a common ground for understanding the different types of stable solutions and for finding stable predictors with the best worst-case performance.

3 A hierarchy of shift-stable distributions

In this section, we present a causal hierarchy of stable distributions that are invariant to different types of dataset shifts. First, we will introduce a general graphical representation for specifying dataset shifts that can occur (Section 3.2). We use this representation to give a simple graphical criterion for determining if a distribution is stable to a set of prespecified shifts. Then, we present the hierarchy and show that the levels of the hierarchy correspond to three operators on the graphical representation, which modify the graph to produce stable distributions (Section 3.3). This allows different stable distributions to be compared in terms of how they modify the graphical representation. We further show that the hierarchy is nested, and thus, it has implications on the existence of stable distributions (Section 3.3.2). We begin by introducing the necessary background on causal graphs (Section 3.1). Proofs of results are present in Appendix B.

3.1 Preliminaries

3.1.1 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 (subscript $\mathcal{G}$ omitted when obvious from context). For an edge e, $\mathrm{He}(e)$ and $\mathrm{Ta}(e)$ will refer to the head and tail of the edge, respectively.

3.1.2 Structural causal models

We represent the data generating process (DGP) underlying a prediction problem using acyclic-directed mixed graphs (ADMGs), $\mathcal{G}$, which consists of a set of vertices $\mathbf{O}$ corresponding to observed variables and sets of directed and bidirected edges such that there are no directed cycles. Directed edges indicate direct causal relations, while bidirected edges indicate the presence of an unobserved confounder (common cause) of the two variables. ADMGs are able to represent directed acyclic graph (DAG) models that contain latent variables. However, the latent variables do not need to be known. For example, if an ADMG has an edge $X \leftrightarrow Y$, then this means that there is some, possibly unknown, mechanism by which X and Y are confounded (i.e., there exists some unobserved variable U such that $X \leftarrow U \rightarrow Y$, but the variable U may be unknown). Thus, using ADMGs, a modeler can reason about the effects of unobserved confounding even if the mechanism or the confounder itself is unknown.

The graph $\mathcal{G}$ defines a Structural Causal Model (SCM) [56] in which each variable $V_i \in \mathbf{O}$ is generated as a function of its parents and a variable-specific exogenous noise variable U_i: $V_i = f_i(pa(V_i), U_i)$. The

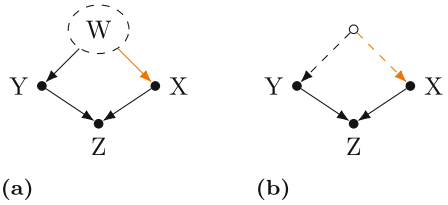

Figure 1: (a) Posited DAG for the pneumonia example of ref. [11]. Y represents the target condition, pneumonia. X represents the X-ray style features (e.g., orientation and color encoding scheme). Z represents the x-ray itself. W represents the hospital department the patient visited. The orange edge represents the unstable style feature mechanism, while the dashed node represents an unobserved variable. (b) The corresponding ADMG for the DAG in (a). The unobserved confounder W has been replaced by a bidirected edge.

prediction problem associated with the graph consists of a target output variable Y and the remaining observed variables as input features.

As an example, consider the DAG in Figure 1a. This DAG corresponds to a simple version of the pneumonia example in ref. [11]. The goal is to diagnose pneumonia Y from chest X-rays Z and stylistic features (i.e., orientation and coloring) of the image X. The latent variable W represents the hospital department the patient visited. The corresponding ADMG is shown in Figure 1b. The unobserved confounder, W, has been replaced by a bidrected edge.

3.2 Stability and types of dataset shifts

In this section, we introduce types of dataset shifts that have been previously studied. Then, we graphically characterize instability in terms of edges in the graph of the data generating process. This will be key to the development of the hierarchy in Section 3.3.

To define the types of dataset shifts, assume that there is a set of environments such that a prediction problem maps to the same graph structure $\mathcal{G}$. However, each environment is a different instantiation of that graph such that certain mechanisms differ. Thus, the factorization of the data distribution is the same in each environment, but terms in the factorization corresponding to shifts will vary across environments. As an example, consider again the graph shown in Figure 1a. In the pneumonia example, each department has its own protocols and equipment, so the style preferences $P(X|W)$ vary across departments. In this example, a *mechanism shift* in the style mechanism $P(X|W)$ leads to differences across environments.

Definition 1. (Mechanism shift) A shift in the mechanism generating a variable V corresponds to arbitrary changes in the distribution $P(V|pa(V))$.

Causal mechanism shifts produce many previously studied instances of dataset shift. Consider, for example, *label shift*, a well-studied mechanism shift in which the distribution of the features X given the label Y ($P(X|Y)$) is stable, but $P(Y)$ varies across environments. Label shift corresponds to a causal graph $Y \rightarrow X$ in which the features are caused by the label, and the mechanism that generates Y varies across environments, resulting in changes in the prevalence $P(Y)$ (see, e.g., [38,57]).

More generally, mechanism shifts are the most common and general type of shift considered in prior work on proactive approaches for addressing dataset shift [20,22,26, 49,50]. However, special cases of mechanism shifts have also been studied. For example, [45,46,58] considered parametric *mean-shifted mechanisms*, in which the means of variables in linear SCMs can vary by environment.

Definition 2. (Mean-shifted mechanisms) A mean shift in the mechanism generating a variable V corresponds to an environment-specific change in the intercept of its linear structural equation $V = \text{intercept}_{\text{env}} + \sum_{X \in pa(V)} \lambda_{xv} X + u_v$. Nonlinear generalizations are possible.

Another special case considered by ref. [21] is *edge-strength shifts*, in which the relationship encoded by a subset of edges into a variable may vary. Variation along an individual edge corresponds to the *natural direct effect* [56, Chapter 4]. Thus, an edge-strength shift is a mechanism shift that changes the natural direct effect associated with the edge.

Definition 3. (Edge-strength shift) An edge-strength shift in edge $X \to V$ corresponds to a change in the *natural direct effect*: for $Y = pa(V)\backslash X$ we have that $E[V(x', Y(x)) - V(x)]$ changes, where $V(x)$ is the counterfactual value of V had X been x, and $V(x', Y(x))$ is the counterfactual value of V had X been x' and had Y been counterfactually generated under $X = x$.

Key result: *All of these shifts can be expressed in terms of edges*. First, edge-strength shifts directly correspond to particular edges. Next, since the mechanism generating a variable V is encoded graphically by all of the edges into V, shifts in mechanism can be represented by marking all edges into V as unstable. For shifts in mechanism to an exogenous variable V with no parents in the graph, one might imagine adding an explicit mechanism variable M_V to the graph and considering the edge $M_V \to V$ to be unstable. Finally, mean-shifts correspond to an edge $A \to V$, where the mean of V is shifted in each environment A (also referred to as an "anchor," see [45] for a discussion of anchor variables). Thus, mean-shifts are an example of a specific type of edge shift. While the edge representation of shifts is more general, we note that it cannot differentiate between specific instances of shifts (e.g., a mean-shift and a shift in the natural direct effect of an "anchor" variable will have the same graphical representation).

We denote the set of *unstable edges* that can vary across environments by $E_u \subseteq E$, where E is the set of edges in $\mathcal{G}$. Graphically, unstable edges will be colored.

Definition 4. (Unstable edge) An edge is said to be unstable if it is the target of an edge-strength shift or a mechanism shift.

The concept of unstable edges provides a flexible and extensible way to graphically represent dataset shifts.

3.2.1 Extensions to new types of shifts

We note that defining shifts in terms of unstable edges makes it possible to tackle new problems determined by shifts in sets or paths of unstable edges. For example, DAGs can be used to represent noni.i.d. network data in which certain edges represent *interference* between units (e.g., friendship ties in social networks) [59,60]. Thus, one can define dataset shifts pertaining to networks (e.g., deleting, adding, or changing the strength of friendships). Similarly, dataset shifts due to changing path-specific effects [61] are another interesting avenue for future exploration (e.g., reductions in side effects of a drug while maintaining its efficacy).

While the focus of this article is on predictive modeling, we note that the shifts we describe have the opportunity to interact with causal inference work on *transportability* and the "*data fusion problem*" [14]. There has been much methodological work on causal graphical methods for transporting causal effect estimates (see, e.g., [12,62–65]). These works have primarily considered transporting causal effects under mechanism shifts. The proposed edge-based definitions of shifts can help frame transportability problems under new types of edge-based shifts such as those motivated earlier.

3.2.2 Stable distributions

We can now define *stable distributions*, which are the target sought by methods addressing instability due to shifts. We will refer to a model of a stable distribution as a *stable predictor*.

Definition 5. (Stable distribution) Consider a graph $\mathcal{G}$ with unstable edges E_u defining a set of environments (different data distributions that factorize with respect to $\mathcal{G}$ that have been generated by differences in the mechanisms associated with E_u). A distribution $P(Y|\mathbf{Z})$ is said to be stable if for any two environments, $\mathcal{E}_1, \mathcal{E}_2$, that are instantiations of $\mathcal{G}$, $P_{\mathcal{E}_1}(Y|\mathbf{Z}) = P_{\mathcal{E}_2}(Y|\mathbf{Z})$ holds. The distribution $P(Y|\mathbf{Z})$ is not restricted to being an observational distribution.

Having established a common graphical representation for arbitrary shifts of various types, we provide a graphical definition of stable distributions. First, define an active *unstable path* to be an active path (as determined by the rules of d-separation [66]) that contains at least one unstable edge.

Key result: *The nonexistence of active unstable paths is a graphical criterion for determining a distribution's stability.*

Theorem 1. *$P(Y|\mathbf{Z})$ is stable if there is no active unstable path from $\mathbf{Z}$ to Y in $\mathcal{G}$ and the mechanism generating Y is stable.*

Intuitively, Theorem 1 means that a stable distribution cannot capture a statistical association that relies on the information encoded by an unstable edge. In the pneumonia example of Figure 1a, the $W \rightarrow X$ edge that denotes the X-ray style mechanism was determined to be unstable. Because W is unobserved, a model of $P(Y|X, Z)$ will learn an association between Y and X through W. Thus, $P(Y|X, Z)$ contains an active unstable path, and this distribution is unstable to shifts in the style mechanism. This means that $P(Y|X, Z)$ is different in each environment. By contrast, if W were observed and we could condition on it, then $P(Y|X, Z, W)$ is stable to shifts in the style mechanism because all paths containing the unstable edge are blocked by W. Thus, $P(Y|X, Z, W)$ is invariant across environments.

In the next section, we use this edge-based graphical characterization to show that all stable distributions, including those found by existing methods, can be categorized into three levels. Thus, this hierarchy defines the ways in which it is possible to achieve stability to shifts.

3.3 Hierarchy of shift-stable distributions

Many works seek stable distributions in order to make predictions that are stable or invariant to dataset shifts. However, because these methods have been developed in isolation, there has been little discussion of whether these methods find the same stable distributions, or how these distributions differ from one another. As a main contribution of this article, we show in this section that there exists a hierarchy of stable distributions, in which stable distributions at different levels have distinct graphical properties. Thus, the development of this hierarchy provides a common theoretical underpinning for understanding when and how stability to shifts can be achieved, and in what ways stable distributions can differ. In this section, we will define the levels of the hierarchy and show that they correspond to different operators that can remove unstable edges from the graph. Then, in the next section, we will further study how differences between levels of the hierarchy affect worst-case performance across environments.

3.3.1 Levels of the hierarchy

Armed with the graphical characterization of stability from the previous section, we now introduce a hierarchy of the three categories of stable distributions. The levels of the hierarchy are as follows: (1) observational conditionals, (2) conditional interventionals, and (3) counterfactuals. This hierarchy is related to the hierarchy of causal queries, which defines three levels of causal study questions an investigator can have: association, intervention, and counterfactuals [56]. Also relatedly, in ref. [67], the authors connect the identification of different types of causal effects to a hierarchy of graphical interventions: node, edge, and path interventions. While these works develop hierarchies that relate different types of causal

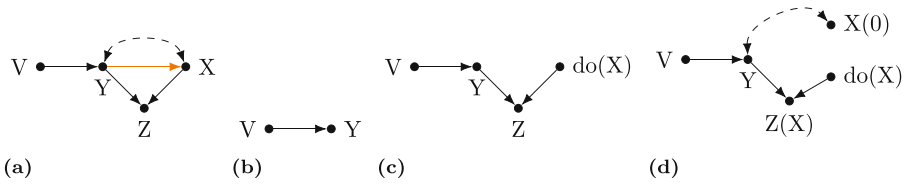

Figure 2: (a) Example graph for a data generating process in which the orange $Y \to X$ edge is unstable. (b) The level 1 operator applied to the graph in (a). The stable level 1 distribution over observed variables is $P(Y|V)$, which ignores all information from X and Z. (c) Level 2 operator applied to (a), deleting the edges into X. The level 2 stable distribution is $P(Y|V, Z, do(X))$. The level 2 operator deletes the stable $Y \leftrightarrow X$ edge. (d) Level 3 operator applied to (a). The stable level 3 distribution is $P(Y|V, Z(X = x), X(Y = 0))$, where $Z(X = x)$ is the value of Z had X been set to its observed value x, and $X(Y = 0)$ is the counterfactual value of X had Y been set to 0. The counterfactual operator retains the stable $Y \leftrightarrow X$ edge.

queries and effects, in this article, we develop a hierarchy of shift-stable distributions that connects different types of stable distributions to interventions, which remove unstable parts of the data generating process from the underlying graph of the DGP.

Each level of the hierarchy of stable distributions corresponds to graphical operators that differ in the precision with which they can remove edges in the graph (Corollary 2, main result of this subsection). By using the graph in Figure 2(a) as a common example, we discuss each level in detail below. Note that in Figure 2(a), the goal is to predict Y from V, X, Z, and the $Y \to X$ edge is unstable.

Definition 6. (Stable level 1 distribution) Let $\mathcal{G}$ be an ADMG with unstable edges E_u defining a set of environments $\mathcal{E}$. A stable level 1 distribution is an observational conditional distribution of the form $P(Y|\mathbf{Z})$ such that, for any two environments $\mathcal{E}_1, \mathcal{E}_2 \in \mathcal{E}$, $P_{\mathcal{E}_1}(Y|\mathbf{Z}) = P_{\mathcal{E}_2}(Y|\mathbf{Z})$ holds.

Level 1: Methods at level 1 of the hierarchy seek invariant conditional distributions of the form $P(Y|\mathbf{Z})$ that use a subset of observed features for prediction [20,26]. These distributions only have conditioning (i.e., the standard rules of d-separation) as a tool for disabling unstable edges. Hence, *the conditioning operator is coarse and removes large pieces of the graph*. Consider Figure 2a, in which the maximal stable level 1 distribution is $P(Y|V)$, since conditioning on either X or Z activates the path through the unstable (orange) edge. The conditioning operator disables all paths from X and Z to Y to produce Figure 2b. While the operator successfully removes the unstable edge, many stable edges were removed as well.

Definition 7. (Stable level 2 distribution) Let $\mathcal{G}$ be an ADMG with unstable edges E_u defining a set of environments $\mathcal{E}$. A stable level 2 distribution is a conditional interventional distribution of the form $P(Y|do(\mathbf{W}), \mathbf{Z})$ such that, for any two environments $\mathcal{E}_1, \mathcal{E}_2 \in \mathcal{E}$, $P_{\mathcal{E}_1}(Y|do(\mathbf{W}), \mathbf{Z}) = P_{\mathcal{E}_2}(Y|do(\mathbf{W}), \mathbf{Z})$ holds.

Level 2: Methods at level 2 [22] find conditional interventional distributions [56] of the form $P(Y|do(\mathbf{W}), \mathbf{Z})$. In addition to conditioning, level 2 distributions use *the do operator, which deletes all edges into an intervened variable* [56]. Figure 2c shows the result of $do(X)$ applied to Figure 2a: the edges into X (including the unstable edge) are removed. Thus, $P(Y|Z, V, do(X))$ is stable and retains statistical information along stable paths from Z and X that the level 1 distribution $P(Y|V)$ did not. However, the stable $Y \leftrightarrow X$ edge was also removed by the operator. Intervening interacts with the factorization (according to the graph) of the joint distribution of the observed variables $P(\mathbf{O})$ by deleting the terms corresponding to mechanisms of the intervened variable: $P(Y|Z, V, do(X)) \propto P(Y|V)P(Z|X, Y)$ in Figure 2a. The term $P(Z|X, Y)$ corresponds to the stable information we retain by intervening that we could not capture by conditioning.

Definition 8. (Stable level 3 distribution) Let $\mathcal{G}$ be an ADMG with unstable edges E_u defining a set of environments $\mathcal{E}$, and let $Z(w)$ denote the counterfactual value of Z had W been set to w for variables $Z, W \subseteq \mathbf{O}$. A stable level 3 distribution is a counterfactual distribution of the form $P(Y(\mathbf{W}, \mathbf{Z}(\mathbf{W}'))|\mathbf{Z}(\mathbf{W}))$ such that, for any two environments $\mathcal{E}_1, \mathcal{E}_2 \in \mathcal{E}$, $P_{\mathcal{E}_1}(Y(\mathbf{W}, \mathbf{Z}(\mathbf{W}'))|\mathbf{Z}(\mathbf{W})) = P_{\mathcal{E}_2}(Y(\mathbf{W}, \mathbf{Z}(\mathbf{W}'))|\mathbf{Z}(\mathbf{W}))$ holds.

Level 3: Finally, level 3 methods [21,24] seek counterfactual distributions, which allow us to consider conflicting values of a variable, or to replace a mechanism with a new one. For example, let Y and Z denote two children of a variable X. If we hypothetically set X to x' for $X \to Y$ but left X as its observed value x for $X \to Z$, this corresponds to counterfactual $Y(x')$ and factual $Z(x) = Z$. By setting a variable to a reference value (e.g., 0) for one edge but not others, *computing counterfactuals effectively removes* (*or replaces*) *a single edge*. In Figure 2c, we saw that $P(Y|V, Z, do(X))$ is stable and deletes both edges into X, including the stable $Y \leftrightarrow X$ edge. However, if we compute the counterfactual $X(Y = 0)$, depicted in Figure 2d, then the level 3 distribution $P(Y|X(Y = 0), V, Z(X = x))$ is stable and only deletes the unstable $Y \to X$ edge, retaining information along the $Y \leftrightarrow X$ path. More generally, level 3 distributions allow us to counterfactually replace mechanisms (and thus replace the influence along unstable edges) with new ones. We will exploit this fact in Section 4 when we investigate accuracy. The effects of the three operators produce the following result:

Corollary 2. *Distributions at increasing levels of the hierarchy of stability grant increased precision in disabling individual edges* (*and thus paths*).

Key result: *Thus, the difference between operators associated with the different levels of stable distributions is the precision of their ability to disable edges into a variable.* Level 1, conditioning, must remove large amounts of the graph to disable edges. Level 2, intervening, deletes all edges into a variable. Level 3, computing counterfactuals, can precisely disable a single edge into a variable. Since paths encode statistical influence, this also provides a natural definition for a *maximally stable distribution* as one which deletes the unstable edges, and only the unstable edges. Thus, given a stable distribution found by any method, we can compare to the maximally stable distribution to see which, and how many, stable paths were removed.

Another important fact is that the hierarchy is *nested*. This means that a level 1 distribution can be expressed as a level 2 distribution (and a level 2 distribution can be expressed as a level 3 distribution):

Lemma 3. ([22], Corollary 1) *A stable level* 1 *distribution of the form* $P(Y|\mathbf{Z})$ *can be expressed as a stable level* 2 *distribution of the form* $P(Y|\mathbf{Z}', do(\mathbf{W}))$ *for* $\mathbf{Z}' \subseteq \mathbf{Z} \subseteq \mathbf{O}$, $\mathbf{W} \subseteq \mathbf{O}$.

Lemma 4. *A stable level* 2 *distribution of the form* $P(Y|\mathbf{Z}', do(\mathbf{W}))$ *can be expressed as a stable level* 3 *distribution of the form* $P(Y(\mathbf{W})|\mathbf{Z}'(\mathbf{W}))$.

3.3.2 Consequences

There are a number of practical consequences of the hierarchy of shift-stable distributions: First, level 1 distributions can always be learned from the available data because conditional distributions are *observational* quantities. This means that we can simply fit and learn a model of $P(Y|\mathbf{Z})$ from the training data. However, because the conditioning operator throws away large parts of the graph, including many stable paths, models of level 1 distributions will generally have higher error compared to models of levels 2 and 3 distributions. A tradeoff exists, though, since levels 2 and 3 distributions are not always *identifiable* – they cannot always be estimated as a function of the observational training data. Level 2 distributions model the effects of hypothetical interventions, and, just as in causal inference, unobserved confounding can lead to identifiability issues (for more detail on identification and level 2 stable distributions see ref. [22]). In addition to identifiability challenges, level 3 counterfactual distributions require further assumptions about the functional form of the causal mechanisms in the SCM. Under a fully specified SCM (i.e., the functions defining mechanisms and their parameters are all known), counterfactual inference can be performed using a three step abduction, action, prediction procedure described in [56, Chapter 7]. For example, the method of [21] assumes linear causal mechanisms to compute level 3 distributions. However, we often have limited information about functional forms and the distribution of the exogenous noise variables in an SCM. If we

want to make counterfactual queries with fewer or no parametric assumptions, then identifiability becomes even more difficult: In general, not all counterfactual queries will be testable. That is, *experimental* data cannot be used to uniquely verify the result of a counterfactual query (where as experimental data can verify the result of any interventional query). For nonparametric SCMs, [68] provided an algorithm for determining if a counterfactual query is empirically testable. Thus, one must balance strong parametric assumptions about the form of causal mechanisms against the possibility of untestable counterfactuals.

The nested nature of the hierarchy means that it has consequences on the existence of stable distributions: If there is no stable level 3 distribution, then no stable level 1 or level 2 distributions exist. Considering the other direction, if we find that no stable level 1 distribution exists, there may still be a stable levels 2 or 3 distribution. This is an important consideration as more methods for finding stable distributions are developed. For example, [22] developed a sound and complete algorithm for finding stable level 2 distributions in a graph. This means that the algorithm returns a distribution *if and only if* a stable level 2 distribution exists. *An open problem is to develop a sound and complete algorithm for finding stable level 3 distributions*. Such a result would be very powerful: If a complete algorithm failed to find a stable level 3 distribution, then that would mean no stable distributions (level 1, 2, or 3) exist.

We have shown that the hierarchy of stable distributions defines graphical operators that can be used to construct stable distributions by disabling edges in the the underlying graph. Next, we show how the ability of counterfactual level 3 distributions to replace edges can be used to achieve minimax optimal performance under dataset shifts.

4 Worst-case performance of shift-stable distributions

We now compare stable distributions with respect to their minimax performance under dataset shifts. Specifically, we show that there is a hypothetical environment in which counterfactually training a model would yield minimax optimal performance across environments. We further show that this level 3 counterfactual distribution is not, in general, a level 2 interventional distribution. Counter to the increasing interest in invariant interventional solutions like Invariant Risk Minimization and its related follow-ups (e.g., [27–29]), these results motivate the development of counterfactual (as opposed to level 2) learning algorithms.

4.1 A decision theoretic view

We now present our result characterizing the stable distribution that achieves minimax optimal performance. First, recall that dataset shifts result in a set of hypothetical environments $\mathcal{E}$ generated from the same graph $\mathcal{G}$ such that the mechanisms associated with unstable edges in $\mathcal{G}$ differ in each environment. For simplicity, we will assume that the mechanism of a single variable $W \in \mathbf{X}$ is subject to shifts, while the mechanisms of all other variables $\mathbf{V} = \{\mathbf{X}, Y\}\backslash\{W\}$ remain stable across environments. Each distribution in the set of data distributions $\mathcal{U}$ corresponding to each environment factorizes according to $\mathcal{G}$, but differs only in the term $P(W|pa_{\mathcal{G}}(W))$, which corresponds to the mechanism for generating W.

Now consider the following game: Suppose the data modeler (DM) wishes to pick the distribution $B \in \mathcal{U}$ such that the corresponding Bayes predictor h_B^* (i.e., the true $P_B(Y|\mathbf{X})$) minimizes the worst-case expected loss (i.e., worst-case risk) across all distributions in $\mathcal{U}$. This can be written as follows:

$$\inf_{B\in\mathcal{U}} \sup_{Q\in\mathcal{U}} E_Q[\ell(h_B^*, \mathbf{O})]. \tag{1}$$

Following a game theoretic result [69, Theorem 6.1], this game has a solution for bounded loss functions ℓ (e.g., the Brier score but not the log loss):

Theorem 5. *Consider a classification problem and suppose ℓ is a bounded loss function. Then equation* (1) *has a solution, and the maximum generalized entropy distribution $Q^* \in \mathcal{U}$ satisfies* $B^* = \arg\inf_{B\in\mathcal{U}} \sup_{Q\in\mathcal{U}} E_Q[\ell(h^*_B, \mathbf{O})] = \arg\sup_{Q\in\mathcal{U}} \inf_{B\in\mathcal{U}} E_Q[\ell(h^*_B, \mathbf{O})] = Q^*$.

Key result: *That this game has a solution means that B^* is the "optimal training environment" such that counterfactually training a predictor in B^* to learn the true $P_{B^*}(Y|\mathbf{X})$ would produce the minimax optimal predictor*. Importantly, this optimal environment B^* depends on the choice of loss function. There are two consequences of this result: First, $P_{B^*}(Y|\mathbf{X})$ is not, in general, a level 2 distribution (and thus, level 2 distributions are not, in general, minimax optimal). Second, there is a level 3 distribution, which corresponds to $P_{B^*}(Y|\mathbf{X})$ and thus is minimax optimal.

Proposition 6. *The level 2 stable distribution $P(Y|do(W), \mathbf{X}) = P_Q(Y|\mathbf{X})$, where Q is the member of $\mathcal{U}$ such that W has a uniform distribution, i.e., $P(W, pa(W)) = cP(pa(W))$ for $c \in \mathbb{R}^+$.*

In Appendix B, we provide a counterexample in which $Q \neq B^*$. This shows that the level 2 stable distribution $P(Y|do(W), \mathbf{X})$ is not minimax optimal.

Proposition 7. *The level 3 distribution $P(Y(W_{B^*})|\mathbf{X}(W_{B^*}))$ equals $P_{B^*}(Y|\mathbf{X})$ and is minimax optimal, where W_{B^*} is the counterfactual W generated under the mechanism associated with the environment B^*.*

Thus, given training data from $P_0 \in \mathcal{U}$, if we could counterfactually learn $P(Y|\mathbf{X})$ in the environment associated with B^*, then the resulting predictor would be minimax optimal. This means the stable level 3 distribution $P(Y(W_{B^*})|\mathbf{X}(W_{B^*}))$ produces the best, worst-case performance across environments out of all distributions that could be used for prediction.

4.2 A simple learning algorithm

Algorithm 1: Gradient descent ascent

input: # of steps T, Step size η, Data $\mathbf{O}$

output: Robust model parameters $\hat{\theta}$

Initialize $\phi^{(1)}$, $\theta^{(1)}$;

for $t \in 2 \dots T$ **do**

$\phi^{(t)} = \phi^{(t-1)} + \eta\nabla_\phi g(\mathbf{O}, \theta^{(t-1)}, \phi^{(t-1)})$;

$\theta^{(t)} = \theta^{(t-1)} - \eta\nabla_\theta g(\mathbf{O}, \theta^{(t-1)}, \phi^{(t-1)})$;

return $\frac{1}{T}\sum_{t=1}^{T}\theta^{(t)}$

We now consider a simple distributionally robust likelihood reweighting algorithm for learning the minimax optimal level 3 predictor. This approach can serve as a starting point for developing new stable learning algorithms that achieve minimax optimal performance under dataset shift.[4]

4 Alternatively, one could try to directly compute the maximum generalized entropy distribution. See, e.g., [84] for a simple example.

For simplicity, suppose there are no unobserved confounders (i.e., $\mathcal{G}$ has no bidirected edges). We relax this condition in Appendix C. Then, learning in the environment B^* using training data from P_0 can be done by reweighting the training data:

$$E_{B^*}[\ell(f(\mathbf{x}), y)] = E_{P_0}\left[\frac{P_{B^*}(\mathbf{O})}{P_{P_0}(\mathbf{O})}\ell(f(\mathbf{x}), y)\right] = E_{P_0}\left[\frac{P_{B^*}(W|pa(W))}{P_{P_0}(W|pa(W))}\ell(f(\mathbf{x}), y)\right],$$

assuming full shared support (i.e., overlap between $P_{B^*}(W|pa(W))$ and $P_{P_0}(W|pa(W))$ for all values of W, $pa(W)$).

Because the minimax optimal training environment B^* is unknown, we now seek to train the minimax optimal predictor by parameterizing environments and iteratively finding the worst-case environment. Let $h(W, pa(W); \phi) = \frac{P_Q(W \mid pa(W))}{P_{P_0}(W \mid pa(W))}$ s.t. $h \in [0, \infty)$ and $E[h|pa(W)] = 1$ be a reweighting function parameterized by ϕ. Note that different values of ϕ correspond to different hypothetical training environments Q. The learning problem becomes

$$\min_{\theta} \max_{\phi} g(\mathbf{O}, \theta, \phi) \tag{2}$$

$$\text{s.t.} g(\mathbf{O}, \theta, \phi) = E_{P_0}[h(W, pa(W); \phi)\ell(f(\mathbf{x}); \theta), y] \tag{3}$$

with model parameters θ. This objective resembles those of distributionally robust methods (e.g., [43]) without restrictions on the density ratio h or the divergence between P_Q and P_{P_0}.

While many possibilities exist, perhaps the simplest version of h is to explicitly learn a parametric density model (e.g., logistic regression for discrete W) $\hat{P}$ for $P_{P_0}(W|pa(W))$ and use the same density model class to model $P_Q(W|pa(W)) = \hat{Q}(W|pa(W); \phi)$. Algorithm 1 describes a gradient descent ascent learning procedure (GDA) for this case, which alternates between finding environmental parameters $\phi^{(t)}$ that maximize the risk of the previous prediction model (with parameters $\theta^{(t-1)}$), and finding model parameters $\theta^{(t)}$ that minimize risk under the previously found worst-case environment (with parameters $\phi^{(t-1)}$). It is important to note that this general minimax learning problem is often very challenging with complicated convergence and equilibrium dynamics (see, e.g., [70–72]). Thus, Algorithm 1 only serves as a starting point for designing counterfactual level 3 learning algorithms.

5 Experiments

We turn to semisynthetic experiments on a real medical prediction task to demonstrate practical performance implications of the hierarchy. To carefully study model behavior under dataset shifts, we posit a graph of the DGP for this dataset and reweight the data to simulate a large number of dataset shifts. We investigate how the performance of models of stable distributions at different levels of the hierarchy behave as test environments differ from the training environment. Our results show that though level 3 models can produce the best worst-case performance (i.e., minimax optimal), level 2 models may perform better on average. This further highlights that model developers need to carefully choose how they achieve stability.

5.1 Motivation and data

One prominent application of machine learning is patient risk stratification in healthcare. It has been widely noted that developing reliable clinical decision support models is difficult due to changes in clinical practice patterns [7]. The resulting behavior-related associations are often brittle – policies change over time and differ across hospitals – and can cause models to make dangerous predictions if left unaccounted [10]. We investigate practical implications of the hierarchy on this important risk prediction challenge.

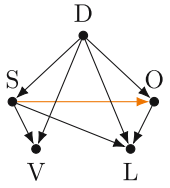

Figure 3: DAG for the sepsis prediction task. The orange edge denotes the unstable edge: lab test ordering policies vary across hospitals.

Below we describe our setup, which loosely follows the setup of [73] for predicting patient risk of sepsis, a life-threatening response to infection. We use electronic health record data collected over four years at our institution's hospital. The dataset consists of 278,947 patient encounters that began in the emergency department. The prevalence of sepsis (S) is 2.3%. Three categories of variables were extracted: vital signs (V) (heart rate, respiratory rate, and temperature), lab tests (L) (lactate), and demographics (D) (age and gender). For encounters that resulted in sepsis, physiologic data available prior to sepsis onset time were used. For nonsepsis encounters, all data available until the time the patient was discharged from the hospital was used. Min, max, and median features were derived for each time-series variable. Unlike vitals, lab measurements are not always ordered (O), so a binary missingness indicator was given. The graph of the DGP is shown in Figure 3.

5.1.1 Shifts in lab test ordering patterns

Different lab test ordering policies correspond to shifts in the conditional $P(O|s, d)$. As a result, missingness patterns vary across datasets derived from different hospitals, because the lab test rate can vary from one institution to another [74]. To compare across datasets corresponding to differing lab testing patterns, we simulated one hundred datasets as follows: For a given test split, we fit a (logistic regression) model of the ordering policy $P(O|s, d)$ (i.e., the P model). Then, for a new ordering policy $Q(O|s, d)$, we reweight the test samples by $\frac{Q}{P}$ to mimic data from a new hospital, which differs only in the ordering policy. Reweighting the examples makes it such that the overall distribution of the reweighted test set is different from the distribution of the original test set without perturbing the feature values of individual examples. Thus, the reweighted datasets consist entirely of examples that were observed in the original dataset.

To simulate an edge shift, we created new ordering policies Q by perturbing the coefficient of sepsis in the P model. This corresponds to changing the log odds ratio for sepsis of a patient receiving a lab test. A log odds <0 mean that lab test orders are more likely for nonsepsis patients than for sepsis patients, while a log odds >0 means that lab test orders are more likely for sepsis patients than for nonsepsis patients. To simulate a mechanism shift, we perturbed all coefficients in the P model.

5.2 Experimental setup

Train/test splits were generated via 5-fold cross-validation. Full experimental details are presented in Appendix A. Models were fit using the Brier score (which for binary classification is $(y - \hat{y})^2$, where y is the true label and $\hat{y}$ is the predicted probability of class $y = 1$) as the loss since it is a bounded loss function (required by Theorem 5).

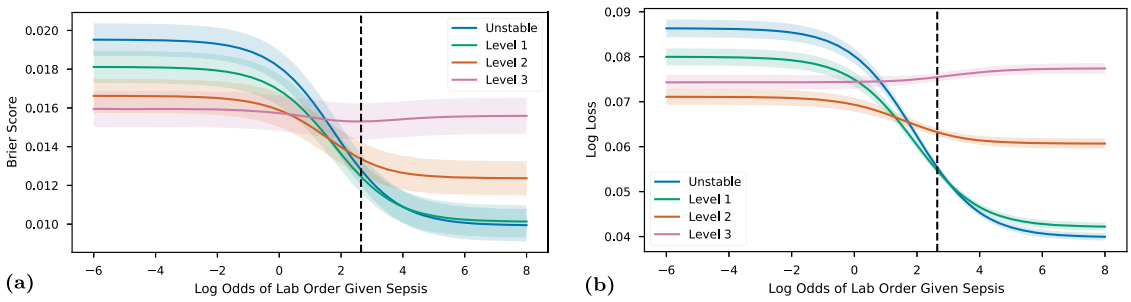

Figure 4: (a) Performance (Brier score) of different models vs log-odds of lab order for the shift in Figure 3. Vertical dashed line denotes training value. Shaded regions denote 95% confidence intervals. (b) Performance (log loss) of different models vs log-odds of lab order for the shift in Figure 3. Note that the level 3 model is minimax optimal in (a) while it is not in (b).

5.2.1 Models

We consider the four possible models: stable models for each level of the hierarchy and an *unstable* baseline that does not adjust for shifts. In fitting models, any model structure (e.g., random forests, neural networks) can be used to fit the marginal/conditional distributions. The choice of model does not impact the study conclusions drawn here. For simplicity, we used logistic regression for all models. The level 1 model excludes the lab-derived and lab order features. With respect to the graph in Figure 3, this effectively deletes the O and L nodes (and all edges into these nodes). The level 2 model is of $P(S|d, v, l, do(o))$, and is an implementation of the "graph surgery estimator" [22]. In Figure 3, the *do* operator deletes both edges into the O node. Finally, the level 3 model was trained using Algorithm 1, with a logistic regression counterfactual reweighting model. In Figure 3, this deletes and then replaces the mechanism for O with a new ordering policy mechanism. All models were implemented in `JAX` [75]. Full details of how the level 2 and 3 models were fit are in Appendix A.

5.3 Results

When test environments differ from the training environment, stable models have more robust performance than unconstrained, unstable models. An unconstrained model uses all dependencies present in the training data; in other words, the model captures correlations due to all paths in the underlying graph. As we impose invariance constraints (by disabling edges), stable models show performance improvements over the unstable model as the test distribution deviates further from the training distribution. We see this, for example, in Figure 4a when, due to the edge shift, the correlation flips from being negative to positive: the level 1, 2, and 3 models outperform the unstable model for log odds ratios <1.

As desired, the level 3 model achieves the best worst-case performance amongst the four models, indicating that training using Algorithm 1 was successful. Further, the performance of the level 3 model is nearly constant across the shifts. This is encouraging evidence, because constant risk is a sufficient condition for a Bayes estimator to be minimax optimal [76]. The results are largely consistent in Figure 5, in which we consider a mechanism shift in the lab test ordering policy. We see that irrespective of the KL divergence between the training and shifted distributions, the level 3 model still has almost constant performance.

Finally, from Theorem 5, we know that the optimal training distribution depends on the choice of loss function (Theorem 5). Thus, we do not expect a minimax optimal predictor under one loss to be optimal when measured under a different loss. Indeed, in Figure 4b, when the four models are evaluated with respect to the log loss, the level 3 model is no longer minimax optimal. In fact, its performance is strictly worse than that of the level 2 model. Even when evaluated using the Brier score (Figure 4a), the worst-case performance of the level 3 model is only slightly better than the worst-case performance of the level 2

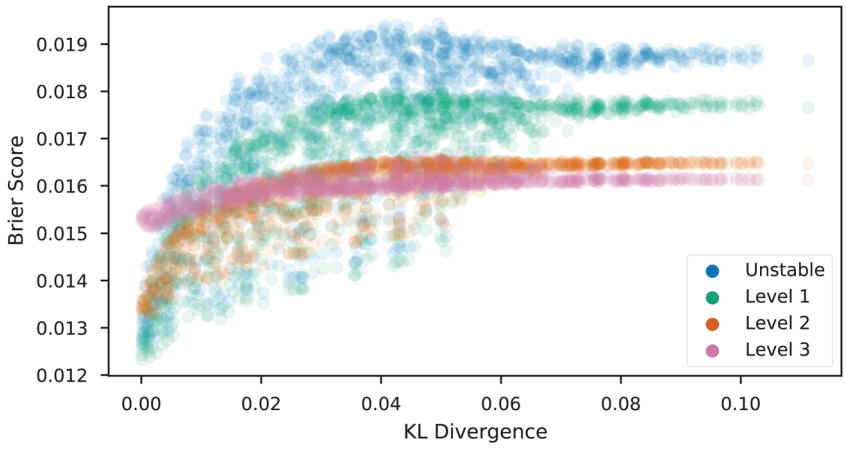

Figure 5: Performance (Brier score) of different models vs KL-divergence of new environment distribution from training distribution under a mechanism shift to the lab ordering policy. The level 3 model is minimax optimal.

model. Further, the level 2 model sees performance improvements when the log odds increase that the level 3 model does not (loss drops noticeably for x-axis values >0). Thus, on average, the level 2 model might be preferable on these data, and a conservative objective like worst-case performance may not be desirable. This illustrates a classic problem in statistical decision theory: while minimax objectives can be too conservative, it may be difficult to characterize the "average" environment or to specify a reasonable prior over environments.

6 Limitations

The primary limitation of the framework presented in this article is its reliance on a known causal graph of the data generating process. Correct specification of the graph is important because the addition of an edge, or the change of orientation of an edge, can change the stability of a distribution. Adding an edge can open new active unstable paths, while a change in orientation of an edge can cause an inactive path to become active (e.g., conditioning on a chain $X \to Y \to Z$ we have $X \perp\!\!\!\perp Z|Y$ vs conditioning on a collider $X \to Y \leftarrow Z$ we have $X \not\perp\!\!\!\perp Z|Y$). As the number of variables increases, it becomes difficult to manually specify an entire causal graph with confidence. In this section, we discuss options for addressing the limitation of misspecification of (or inability to specify) the graph.

When domain knowledge is insufficient to specify a causal graph, one can try to learn the structure of the graph from data, a problem known as *causal discovery* or *structure learning* (see [77] for an overview). *Constraint-based* structure learning algorithms work by using (conditional) independence tests to determine edge adjacencies. Thus, by testing compatibility with the data, it is possible to learn the structure of the graph up to an *equivalence class*: a set of fully specified graphs which imply the same independences. Notably, some constraint-based structure learning algorithms tolerate and account for the possibility of unknown confounding variables (see [78] for a recent review of structure learning algorithms).

Using structure learning, it is possible to learn the range of causal structures that are compatible with the data. Given this range of causal structures, there are two main approaches one could use to find stable distributions. One approach is to enumerate each member of the equivalence class, find and fit models of stable distributions in the fully specified member, and compare across the members of the equivalence class. This approach is akin to sensitivity analysis approaches for finding the range of causal effect estimates in the equivalence class (see, e.g., [79,80]). The challenge with this approach is that it does not produce a single model that is guaranteed to be stable, but rather a range of candidate "possibly stable" models. One would require data from a new environment to test the stability of the candidate models.

A second approach is to find a distribution which is stable in every member of the equivalence class. Such a distribution is guaranteed to be stable, regardless of which member of the equivalence class represents the "true" data generating process. While this could be done through enumeration of each member of the equivalence class (as in the previously outlined sensitivity analysis approach), recent approaches allow us to find stable distributions in graphical representations of the equivalence class [23,55]. The output of many constraint-based structure learning algorithms is a *partial* graph in which edges may be partially directed (i.e., edge endpoints may be an arrowhead, an arrow tail, or ∘ representing that either is possible). One can then consider extensions of graphical operators from the hierarchy to partial graphs. As one example, [23] proposed a method for finding stable level 1 and level 2 distributions in partial graphs. More generally, a promising direction for future work is to extend results from the proposed framework to partial graphs. Because partial graphs can be learned from data, this would relax the requirement of a fully specified graph as the starting point for this graphical framework.

7 Contrast with invariant risk minimization

The discussion in this article has focused on a graphical perspective – explicitly starting with the knowledge of the data generating process and using this to determine when and how stability to shifts is achievable. An alternative emerging paradigm in machine learning has focused on *invariant risk minimization* (IRM) [27–29]. IRM is applicable when multiple datasets from different environments are available, and the goal is to learn a representation that produces an optimal predictor which is invariant across these environments. In this section, we discuss an important limitation of the invariant risk minimization paradigm that highlights a key advantage of graphical approaches. We also discuss how graphical analyses can guide the future work to address this.

A critical question that determines the usefulness of an invariant predictor is: To what set of shifts is the predictor stable? The answer to this question defines the set of new environments to which an invariant predictor can be safely applied. In the graphical approach, the answer is transparent by design. Shifts are defined as (arbitrary) changes to particular causal mechanisms in the graph, so an invariant predictor is exactly one which is stable to the specified shifts in mechanisms. Further, the graph allows model developers to choose the set of shifts to which a predictor should be stable and provide guarantees about shifts that are protected against.

In contrast, IRM methods currently struggle to answer this critical question. First, existing IRM methods do not identify the differences that exist across the observed environments. Thus, they are unable to provide guarantees about the nature of the shifts in environment (i.e., the causal mechanisms) against which they protect. This also means, it is difficult to state the set of new environments to which the invariant predictor can be safely applied. Further, because IRM automatically determines invariance from datasets, there is no opportunity for developers to specify particular invariances that they want to hold.

Outside of invariant risk minimization, there are opportunities to leverage ideas from other works on invariant learning and ideas from the proposed graphical framework to improve IRM-type methods. For example, [81] shows a relationship between invariant predictors and calibration across environments. This suggests a possible approach for probing an invariant predictor for stability to particular mechanisms shifts. First, using *structure learning* [77], it is possible to detect particular mechanism shifts that occur across environments [23,55,82]. Then, when mechanisms of interest have been identified, one can test for stability to particular mechanisms shifts by examining how the calibration of the predictor changes as evaluation data are reweighted according to the distribution associated with the mechanism shift. This would provide a *post hoc* way to verify the integrity of a trained invariant predictor.

As another example, [24] show that counterfactual invariances leave observable distributional signatures that can be used to design regularizers to enforce the given invariance. This motivates the combination of IRM-type objectives with regularizers that explicitly capture desired invariances at different levels of the hierarchy of shift-stability. This would allow developers to specify particular invariances they want to

guarantee while also automatically learning other invariances from the data. In the context of image classification, [44] showed how multiple views of an image and data augmentation can be used to learn models that are invariant to shifts in known and unknown style features. This provides ideas for learning specified invariances in settings with unstructured data (e.g., images and text).

8 Conclusion

The use of machine learning in production represents a shift from applying models to static datasets to applying them in the real world. As a result, aspects of the underyling DGP are almost certain to change. Many methods have been developed to find distributions that are stable to dataset shift, but as a field we have lacked common underlying theory to characterize and relate different stable distributions. To address this, we developed a common framework for expressing the different types of shifts as unstable edges in a graphical representation of the DGP. We further showed that stable distributions belong to a causal hierarchy in which stable distributions at different levels have distinct operators that can remove unstable edges in the graph. This provides a new, but natural, way to characterize and construct stable models by only removing unstable edges. This also motivates a new paradigm for future work developing methods that can modify individual edges. We also showed that popular invariant solutions (level 2; invariant under intervention) do not, in general, achieve minimax optimal performance across environments. Our experiments showed that there is a tradeoff between worst-case average performance. Thus, model developers need to carefully determine when and how they achieve invariance.

Acknowledgments: The authors gratefully acknowledge support from the Sloan Foundation (FG-2018–10877).

Conflict of interest: Authors state no conflict of interest.

Data availability statement: This study was approved by the Johns Hopkins University internal review board. Under this agreement, the data cannot be shared with outside investigators. The authors can provide more details about the data upon request.

References

[1] Strickland E. Hospitals roll out AI systems to keep patients from dying of sepsis. IEEE Spectrum. 2018;19. https://spectrum.ieee.org/hospitals-roll-out-ai-systems-to-keep-patients-from-dying-of-sepsis.
[2] Winston A. Palantir has secretly been using New Orleans to test its predictive policing technology. The Verge. 2018;27. https://www.theverge.com/2018/2/27/17054740/palantir-predictive-policing-tool-new-orleans-nopd.
[3] Angwin J, Larson J, Mattu S, Kirchner L. Machine bias. ProPublica. May 2016;23(2016):139–59.
[4] Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND. Dataset shift in machine learning. Cambridge, MA, USA: The MIT Press; 2009.
[5] Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med. 2021;385(3):283–6.
[6] Dickson B. How the coronavirus pandemic is breaking artificial intelligence and how to fix it. Gizmodo; 2020. Available from: https://gizmodo.com/how-the-coronavirus-pandemic-is-breaking-artificial-int-1844544143.
[7] Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. Bmj. 2018;361:k1479.
[8] Grytten J, Sørensen R. Practice variation and physician-specific effects. J Health Econom. 2003;22(3):403–18.
[9] Cutler D, Skinner JS, Stern AD, Wennberg D. Physician beliefs and patient preferences: a new look at regional variation in health care spending. Am Econ J Econ Policy. 2019;11(1):192–221.
[10] Schulam P, Saria S. Reliable decision support using counterfactual models. In: Advances in neural information processing systems. Long Beach, CA, USA: Neural Information Processing Systems Foundation, Inc.;2017. p. 1697–708.

[11] Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. 2018;15(11):e1002683.
[12] Pearl J, Bareinboim E. Transportability of causal and statistical relations: a formal approach. In: Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence. San Francisco, CA, USA: AAAI Press; 2011. p. 247–54.
[13] Stuart EA, Bradshaw CP, Leaf PJ. Assessing the generalizability of randomized trial results to target populations. Prevention Sci. 2015;16(3):475–85.
[14] Bareinboim E, Pearl J. Causal inference and the data-fusion problem. Proc Nat Acad Sci. 2016;113(27):7345–52.
[15] Degtiar I, Rose S. A review of generalizability and transportability. 2021. arXiv: http://arXiv.org/abs/arXiv:210211904.
[16] Heckman J. Shadow prices, market wages, and labor supply. Econometrica J Econom Soc. 1974;42(4):679–94.
[17] Heckman JJ. Sample selection bias as a specification error. Econometrica J Econom Soc. 1979;47(1):153–61.
[18] Winship C, Mare RD. Models for sample selection bias. Annual Rev Sociol. 1992;18(1):327–50.
[19] Vella F. Estimating models with sample selection bias: a survey. J Human Res. 1998;33(1):127–69.
[20] Magliacane S, van Ommen T, Claassen T, Bongers S, Versteeg P, Mooij JM. Domain adaptation by using causal inference to predict invariant conditional distributions. In: Advances in neural information processing systems. Montreal, Canada: Neural Information Processing Systems Foundation, Inc.; 2018. p. 10869–79.
[21] Subbaswamy A, Saria S. Counterfactual normalization: proactively addressing dataset shift using causal mechanisms. In: Uncertainty in artificial intelligence. Monterey, CA, USA: AUAI Press; 2018. p. 947–57.
[22] Subbaswamy A, Schulam P, Saria S. Preventing failures due to dataset shift: learning predictive models that transport. In: Artificial intelligence and statistics (AISTATS). Naha, Okinawa, Japan: PMLR; 2019. p. 3118–27.
[23] Subbaswamy A, Saria S. I-SPEC: An End-to-End Framework for Learning Transportable, Shift-Stable Models. 2020. arXiv: http://arXiv.org/abs/arXiv:200208948.
[24] Veitch V, D'Amour A, Yadlowsky S, Eisenstein J. Counterfactual invariance to spurious correlations in text classification. In Advances in neural information processing systems. La Jolla, CA, USA: Neural Information Processing Systems Foundation, Inc.; 2021. p. 34.
[25] Ilse M, Tomczak JM, Forré P. Selecting data augmentation for simulating interventions. In: International Conference on Machine Learning. San Diego, CA, USA: PMLR; 2021. p. 4555–62.
[26] Rojas-Carulla M, Schölkopf B, Turner R, Peters J. Invariant models for causal transfer learning. J Mach Learn Res. 2018;19(1):1309–42.
[27] Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D. Invariant risk minimization. 2019. arXiv: http://arXiv.org/abs/arXiv:190702893.
[28] Bellot A, van der Schaar M. Generalization and invariances in the presence of unobserved confounding. 2020. arXiv: http://arXiv.org/abs/arXiv:200710653.
[29] Koyama M, Yamaguchi S. Out-of-distribution generalization with maximal invariant predictor. 2020. arXiv: http://arXiv.org/abs/arXiv:200801883.
[30] Campbell DT, Stanley JC, Gage NL. Experimental and quasi-experimental designs for research. Houghton: Mifflin and Company; 1963.
[31] Rothwell PM. Commentary: External validity of results of randomized trials: disentangling a complex concept. Int J Epidemiol. 2010;39(1):94–6.
[32] Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. Am J Epidemiol. 2010;172(1):107–15.
[33] Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. J R Statist Soc A (Statist Soc). 2011;174(2):369–86.
[34] Pearl J, Bareinboim E. External validity: From do-calculus to transportability across populations. Statist Sci. 2014;29(4):579–95.
[35] Dahabreh IJ, Robins JM, Haneuse SJ, Hernán MA. Generalizing causal inferences from randomized trials: counterfactual and graphical identification. 2019. arXiv: http://arXiv.org/abs/arXiv:190610792.
[36] Camerer C. The promise and success of lab-field generalizability in experimental economics: a critical reply to Levitt and List. Available at SSRN 1977749. 2011.
[37] Huang J, Gretton A, Borgwardt K, Schölkopf B, Smola AJ. Correcting sample selection bias by unlabeled data. In: Advances in neural information processing systems. Vancouver, B.C., Canada: Neural Information Processing Systems Foundation, Inc.; 2007. p. 601–8.
[38] Zhang K, Schölkopf B, Muandet K, Wang Z. Domain adaptation under target and conditional shift. In: International Conference on Machine Learning. Atlanta, USA: PMLR; 2013. p. 819–27.
[39] Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, et al. Domain-adversarial training of neural networks. J Machine Learn Res. 2016;17(1):2096–30.
[40] Gong M, Zhang K, Liu T, Tao D, Glymour C, Schölkopf B. Domain adaptation with conditional transferable components. In: International Conference on Machine Learning. New York, NY, USA: PMLR; 2016. p. 2839–48.
[41] Correa JD, Bareinboim E. From statistical transportability to estimating the effect of stochastic interventions. In: IJCAI. Macao, China: International Joint Conferences on Artificial Intelligence; 2019. p. 1661–7.
[42] Sinha A, Namkoong H, Duchi J. Certifying some distributional robustness with principled adversarial training. 2017. arXiv: http://arXiv.org/abs/arXiv:171010571.

[43] Duchi J, Namkoong H. Variance-based regularization with convex objectives. 2016. arXiv: http://arXiv.org/abs/arXiv:161002581.
[44] Heinze-Deml C, Meinshausen N. Conditional variance penalties and domain shift robustness. Mach Learn. 2020;110:1–46.
[45] Rothenhäusler D, Meinshausen N, Bühlmann P, Peters J. Anchor regression: heterogeneous data meets casuality. 2018. arXiv: http://arXiv.org/abs/arXiv:180106229.
[46] Oberst M, Thams N, Peters J, Sontag D. Regularizing towards causal invariance: linear models with proxies. 2021. arXiv: http://arXiv.org/abs/arXiv:210302477.
[47] Muandet K, Balduzzi D, Schölkopf B. Domain generalization via invariant feature representation. In: International Conference on Machine Learning; 2013. Atlanta, GA, USA: PMLR; p. 10–18.
[48] Ahuja K, Shanmugam K, Varshney K, Dhurandhar A. Invariant risk minimization games. In: International Conference on Machine Learning. Vienna, Austria: PMLR; 2020. p. 145–55.
[49] Peters J, Bühlmann P, Meinshausen N. Causal inference by using invariant prediction: identification and confidence intervals. J R Statist Soc Ser B (Statist Methodol). 2016;78(5):947–1012.
[50] Kuang K, Cui P, Athey S, Xiong R, Li B. Stable prediction across unknown environments. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2018. p. 1617–26.
[51] Kuang K, Xiong R, Cui P, Athey S, Li B. Stable prediction with model misspecification and agnostic distribution shift. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34. New York, NY, USA: AAAI Press; 2020. p. 4485–92.
[52] Kaushik D, Hovy E, Lipton ZC. Learning the difference that makes a difference with counterfactually-augmented data. 2019. arXiv: http://arXiv.org/abs/arXiv:190912434.
[53] Kaushik D, Setlur A, Hovy EH, Lipton ZC. Explaining the efficacy of counterfactually augmented data. In: International Conference on Learning Representations. Addis Ababa, Ethiopia: OpenReview; 2020.
[54] Sundin I, Schulam P, Siivola E, Vehtari A, Saria S, Kaski S. Active learning for decision-making from imbalanced observational data. 2019. arXiv: http://arXiv.org/abs/arXiv:190405268.
[55] Zhang K, Gong M, Stojanov P, Huang B, Glymour C. Domain adaptation as a problem of inference on graphical models. 2020. arXiv: http://arXiv.org/abs/arXiv:200203278.
[56] Pearl J. Causality. Cambridge, England: Cambridge University Press; 2009.
[57] Schölkopf B, Janzing D, Peters J, Sgouritsa E, Zhang K, Mooij J. On causal and anticausal learning. In: Proceedings of the 29th International Coference on International Conference on Machine Learning. Madison, WI, USA: Omnipress; 2012. p. 459–66.
[58] Meinshausen N. Causality from a distributional robustness point of view. In: 2018 IEEE Data Science Workshop (DSW). Lausanne, Switzerland: IEEE; 2018. p. 6–10.
[59] Ogburn EL, VanderWeele TJ. Causal diagrams for interference. Statist Sci. 2014;29(4):559–78.
[60] Sherman E, Shpitser I. Intervening on network ties. In: Uncertainty in artificial intelligence. Toronto, Canada: PMLR; 2020. p. 975–84.
[61] Avin C, Shpitser I, Pearl J. Identifiability of path-specific effects. In: IJCAI International Joint Conference on Artificial Intelligence; 2005. p. 357–63.
[62] Bareinboim E, Pearl J. Transportability of causal effects: completeness results. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 26. Toronto, Ontario, Canada: AAAI Press; 2012. p. 698–704.
[63] Bareinboim E, Pearl J. Meta-transportability of causal effects: a formal approach. In: Artificial intelligence and statistics. Scottsdale, AZ, USA: PMLR; 2013. p. 135–43.
[64] Lee S, Correa J, Bareinboim E. General transportability-synthesizing observations and experiments from heterogeneous domains. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34. New York, NY, USA: AAAI Press; 2020. p. 10210–7.
[65] Lee S, Correa JD, Bareinboim E. Generalized transportability: Synthesis of experiments from heterogeneous domains. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, NY, USA: AAAI Press; 2020.
[66] Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1988.
[67] Shpitser I, Tchetgen ET. Causal inference with a graphical hierarchy of interventions. Annals of Statistics. 2016;44(6):2433.
[68] Shpitser I, Pearl J. What counterfactuals can be tested. In: 23rd Conference on Uncertainty in Artificial Intelligence, UAI 2007; 2007. p. 352–9.
[69] Grünwald PD, Dawid AP. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. Annals Statist. 2004;32(4):1367–433.
[70] Daskalakis C, Ilyas A, Syrgkanis V, Zeng H. Training gans with optimism. 2017. arXiv: http://arXiv.org/abs/arXiv:171100141.
[71] Daskalakis C, Panageas I. The limit points of (optimistic) gradient descent in min-max optimization. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada: Neural Information Processing Systems Foundation, Inc.; 2018. p. 9256–66.
[72] Lin T, Jin C, Jordan M. On gradient descent ascent for nonconvex-concave minimax problems. In: International Conference on Machine Learning. Vienna, Austria: PMLR; 2020. p. 6083–93.

[73] Giannini HM, Ginestra JC, Chivers C, Draugelis M, Hanish A, Schweickert WD, et al. A machine learning algorithm to predict severe sepsis and septic shock: development, implementation, and impact on clinical practice. Critical Care Med. 2019;47(11):1485–92.
[74] Rhee C, Dantes R, Epstein L, Murphy DJ, Seymour CW, Iwashyna TJ, et al. Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009–2014. Jama. 2017;318(13):1241–9.
[75] Bradbury J, Frostig R, Hawkins P, Johnson MJ, Leary C, Maclaurin D, et al. Google, editor. JAX: composable transformations of Python.NumPy programs. GitHub; 2018. Available from: http://github.com/google/jax.
[76] Berger JO. Statistical decision theory and Bayesian analysis. New York, NY, USA: Springer Science and Business Media; 2013.
[77] Spirtes P, Glymour CN, Scheines R, Heckerman D, Meek C, Cooper G, et al. Causation, prediction, and search. Cambridge, MA, USA: MIT Press; 2000.
[78] Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models. Front Genetic. 2019;10:524.
[79] Maathuis MH, Kalisch M, Bühlmann P. Estimating high-dimensional intervention effects from observational data. Annal Statist. 2009;37(6A):3133–64.
[80] Malinsky D, Spirtes P. Estimating bounds on causal effects in high-dimensional and possibly confounded systems. Int J Approximate Reason. 2017;88:371–84.
[81] Wald Y, Feder A, Greenfeld D, Shalit U. On calibration and out-of-domain generalization. In Advances in neural information processing systems. 2021. La Jolla, CA, USA: Neural Information Processing Systems Foundation, Inc.; p. 34.
[82] Zhang K, Huang B, Zhang J, Glymour C, Schölkopf B. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In: IJCAI: Proceedings of the Conference. Vol. 2017. NIH Public Access; 2017. p. 1347.
[83] Delahanty RJ, Alvarez J, Flynn LM, Sherwin RL, Jones SS. Development and evaluation of a machine learning model for the early identification of patients at risk for sepsis. Annals Emergency Med. 2019;73(4):334–44.
[84] van Ommen T. Robust causal domain adaptation in a simple diagnostic setting. In: International Symposium on Imprecise Probabilities: Theories and Applications. Ghent, Belgium: PMLR; 2019. p. 424–9.
[85] Tian J. Studies in causal reasoning and learning [Ph.D. thesis]. University of California: Los Angeles; 2002.
[86] Duchi JC, Namkoong H. Learning models with uniform performance via distributionally robust optimization. Annals Statist. 2021;49(3):1378–406.

Appendix

A Medical risk prediction experiment

A.1 Data

Our experimental setup follows that of [83]. The dataset contains electronic health record data collected over four years at our institution's hospital. The dataset consists of 278,947 emergency department patient encounters. The prevalence of the target disease, sepsis (S), is 2.1%. Features pertaining to vital signs (V) (heart rate, respiratory rate, temperature), lab tests (L) (lactate), and demographics (D) (age, gender) were extracted. For encounters that resulted in sepsis (i.e., positive encounters), physiologic data available up until sepsis onset time were used. For nonsepsis encounters, all data available until discharge were used. For each of the time-series physiologic features (V and L), min, max, and median summary features were derived. Unlike vitals, lab measurements are not always ordered (O) and are subject to missingness (lactate 89% missing). To model lab missingness, missingness indicators (O) for the lab features were added, and lab value-missingness interactions terms were used in place of lab value features.

A.2 Experimental details

Logistic regression models were fit using a custom JAX[75] implementation. L_2 regularization with regularization coefficient 0.1 was used (hyperparameter chosen via grid search using the performance of the

unstable model on a hold-out 10% of the initial dataset). These same hyperparameters were used to train the levels 1–3 models. For the predictive models, a b-spline basis feature expansion was used for continuous features (lab values and vital signs). Following the standards in [83] for accounting for missingness, the missingness feature and the missingness-lab value interaction features were added.

The specific shift in lab test ordering patterns considered was a shift in lactate ordering, as these patterns have seen great variation across hospitals and are known to be associated with sepsis [74]. Lactate missingness has a correlation of −0.36 with sepsis in this dataset (i.e., the presence of the measurement is predictive of the target variable).

Thus, to simulate the edge shift, in each test fold, we first fit a logistic regression model (no b-spline basis expansion, with default scikit-learn hyperparameters) to the test fold's lactate missingness ($O = 0$) given S, D. That is, a logistic regression model of $P(O = 0|s, d)$. Then, to simulate the edge shifted lactate ordering policies, we replaced the coefficient for sepsis in the logistic regression model with 100 values on a grid from −6 to 8. The resulting logistic regression model is of the hypothetical shifted hospital's ordering policy $Q(O = 0|s, d)$. Evaluating the loss under each shift was then done by using sample weights computed as $\frac{Q_i}{P_i}$ for each test sample using the two models.

The mechanism shift was simulated in a similar manner to the edge shift. However, instead of only perturbing the coefficient of sepsis in the $P(O|s, d)$ model, all coefficients and the intercept were perturbed. Specifically, for a single test fold, 1,000 new coefficients were sampled as follows: Let w denote the weight in the P model. The new coefficients/intercepts were drawn from Unif$(-|w| - 0.1, |w| + 0.1)$. Because all weights of the logistic regression model changed, we plotted the shifts according to the estimated (using the test set) KL-divergence between the Q and P logistic regression models: $E_{P_{S,D}}[KL(P(O|s, d)||Q(O|s, d)]$.

The level 1 model was fit using a reduced feature set that excluded the lactate features (min, max, median) and lactate missingness indicator. The level 2 model $P(S|d, v, l, do(o))$, an instance of the "graph surgery estimator" [22], was fit by inverse probability weighting (IPW). The term corresponding to O in the factorization of the DAG in Figure 4a is $P(O|s, d)$, so we fit a logistic regression model of this distribution using the training data. Then, the main logistic regression prediction model with the full feature set was trained using sample weights $\frac{1}{P(o_i \mid s_i, d_i)}$. The resulting model was the level 2 model. The level 3 model is similar to the level 2 model, but instead corresponds to a counterfactual ordering policy $Q(O|s, d)$. The level 3 logistic regression model was trained using the gradient ascent descent procedure in Algorithm 1. As noted in the main article, this procedure has complicated dynamics, and we found that it was quite sensitive to the choice of step size (or learning rate η). Through grid search using performance on a hold-out 10% of the initial dataset, the value $\eta = 5$ was selected. The model parameters θ were initialized using the learned level 2 model parameters, and the reweighting ϕ parameters were initialized via random draws from $\mathcal{N}(0, 0.1^2)$. The resulting model was the level 3 model.

B Proofs

Theorem 1. *$P(Y|\mathbf{Z})$ is stable if there is no active unstable path from $\mathbf{Z}$ to Y in $\mathcal{G}$ and the mechanism generating Y is stable.*

Proof. We first recall that all of the shifts considered in Section 3.2 are types of arbitrary shifts in mechanism: mean-shifted mechanism are a special parametric case, and edge-strength shifts correspond to a constrained class of mechanism shifts in which the natural direct effect associated with the mechanism has changed. Thus, if a distribution is stable to arbitrary shifts in mechanisms, then it will also be stable to mean-shifts and edge-shifts. Hence, in our proof, we will prove a distribution is stable by leveraging previous graphical results on stability under shifts in mechanisms (and stability to specific cases follows).

To do so, we will leverage results from *transportability*, which uses a graphical representation called *selection diagrams* (see [12, 22] for details). A selection diagram is a a graph augmented with selection variables $\mathbf{S}$ (which each have at most one child) that point to variables whose mechanisms may vary across

environments. Prior results have shown that a distribution $P(Y|\mathbf{Z})$ is stable if $Y \perp\!\!\!\perp \mathbf{S}|\mathbf{Z}$ in the selection diagram (see [12, Theorem 2] and [22, Definition 3]). Thus, to prove the theorem, we will first translate our unstable edge representation of the graph to a selection diagram. Then, we will show that if Y is not d-separated from the selection variables that this implies, there is an unstable active path to Y.

We first translate our unstable edge representation of the graph to a selection diagram. For an edge e, let $\mathrm{He}(e)$ denote the variables that e points into. Now for each $e \in E_u$, add a unique selection variable that points to each $V = \mathrm{He}(e)$. This indicates that the mechanism that generates V is unstable. We now consider the cases in which there could be an active path from a selection variable to Y (which would make a distribution $P(Y|\mathbf{Z})$ unstable) and also show that this corresponds to an active path that contains an unstable edge.

There are two possible ways there can be an active path from a variable $S \in \mathbf{S}$ to Y. If there is an active forward path from S to Y (e.g., $S \to ch(S) \to \ldots Y$), then there is a corresponding active path from $Ta(e)$ to Y that contains the unstable edge e: e.g., a path $Ta(e) - e \to ch(S) \to \ldots Y$. Alternatively, an active forward path indicates that the mechanism that generates Y is unstable.

The other case is if there is an active path beginning with a collider from S to Y (e.g., $S \to ch(S) \leftarrow \ldots Y$). Then there is a corresponding active path from $Ta(e)$ to Y that contains e: e.g., $Ta(e) - e \to ch(S) \leftarrow \ldots Y$. Thus, in a selection diagram if $P(Y|\mathbf{Z})$ is unstable, then there is an active unstable path to Y in the original unstable edge-denoted graph. Taking the contrapositive of this statement proves the theorem. □

Lemma 3. ([22], Corollary 1). *A stable level* 1 *distribution of the form* $P(Y|\mathbf{Z})$ *can be expressed as a stable level* 2 *distribution of the form* $P(Y|\mathbf{Z}', do(\mathbf{W}))$ *for* $\mathbf{Z}' \subseteq \mathbf{Z} \subseteq \mathbf{O}$, $\mathbf{W} \subseteq \mathbf{O}$.

Proof. This is a restatement of Corollary 1 in [22]. □

Lemma 4. *A stable level* 2 *distribution of the form* $P(Y|\mathbf{Z}', do(\mathbf{W}))$ *can be expressed as a stable level* 3 *distribution of the form* $P(Y(\mathbf{W})|\mathbf{Z}'(\mathbf{W}))$.

Proof. Consider the (level 2) intervention $do(X) = x$. For a variable V letting $V(x)$ denote the value V would have taken had X been set to x, we have that $P(V(x)) = P(V|do(x))$. When interventions are consistent (i.e., for $x \neq x'$, there are no conflicting interventions $do(X = x)$ and $do(X = x')$) counterfactuals reduce to the *potential responses* of interventions expressible with the *do* operator [56, Definition 7.1.4]. □

For completeness, we restate the following result from [69]. For the present article, both the action space $\mathcal{A}$ and the set of distributions Γ are $\mathcal{U}$ (the DM is picking a training distribution and the action a from $\mathcal{U}$ and nature is picking the test distribution P from $\mathcal{U}$).

Theorem 8. (Theorem 6.1, [69]) *Let* $\Gamma \subseteq \mathcal{P}$ *be a convex, weakly closed, and tight set of distributions. Suppose that for each* $a \in \mathcal{A}$ *the loss function* $L(x, a)$ *is bounded above and upper semicontinuous in* x. *Then the restricted game* $\mathcal{G}^\Gamma = (\Gamma, \mathcal{A}, L)$ *has a value. Moreover, a maximum entropy distribution* P^*, *attaining*

$$\sup_{P \in \Gamma} \inf_{a \in \mathcal{A}} L(P, a),$$

exists.

Theorem 5. *Consider a classification problem and suppose* ℓ *is a bounded loss function. Then equation* 1 *has a solution, and the maximum generalized entropy distribution* $Q^* \in \mathcal{U}$ *satisfies* $B^* = \arg\inf_{B \in \mathcal{U}} \sup_{Q \in \mathcal{U}} E_Q[\ell(h_B^*, \mathbf{O})] = \arg\sup_{Q \in \mathcal{U}} \inf_{B \in \mathcal{U}} E_Q[\ell(h_B^*, \mathbf{O})] = Q^*$.

Proof. This result follows directly from [69, Theorem 6.1].

The preconditions are trivially satisfied: The set of all distributions over W is convex, closed, and tight. We consider bounded loss functions, which for finite discrete Y (i.e., for classification problems) are continuous. Thus, the game has a solution.

Further, by [69, Corollary 4.2], the maximum generalized entropy distribution Q^* is also the distribution minimizing the worst-case expected loss. □

Proposition 6. *The level 2 stable distribution* $P(Y|do(W), \mathbf{X}) = P_Q(Y|\mathbf{X})$, *where* Q *is the member of* $\mathcal{U}$ *such that* W *has a uniform distribution, i.e.,* $P(W, pa(W)) = cP(pa(W))$ *for* $c \in \mathbb{R}^+$.

Proof. We know that every distribution in $\mathcal{U}$ factorizes according to the graph $\mathcal{G}$, and that they only differ in the term corresponding to the mechanism for W, $P(W|pa(W))$. Thus, for any $D \in \mathcal{U}$, $P_D(W, pa(W)) = P_D(W|pa(W))P_D(pa(W)) = P_D(W|pa(W))P(pa(W))$, noting that $P(pa(W))$ is the same across all members of $\mathcal{U}$. It suffices to show, then, that $P(\mathbf{O}\backslash\{W\}|do(W)) \propto P_Q(\mathbf{O})$ (within a constant factor), such that $P_Q(W, pa(W)) = cP(pa(W))$.

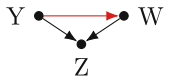

Figure A1: Graph for the counterexample.

Recall that, by definition, performing $do(W)$ deletes the W term from the factorization (or equivalently sets $P(W = w|do(w), pa(W)) = P(W = w|do(w)) = 1$), resulting in the so-called truncated factorization. Further, the resulting distribution $P(\mathbf{O}\backslash\{W\}|do(W))$ is a proper distribution (sums to 1) over $\mathbf{O}\backslash\{W\}$. Consider two cases: 1) That W is a discrete variable or 2) W is a continuous variable. With slight abuse of notation, for continuous variables, the results will be with respect to the pdf.

(1) Suppose W is discrete and that across environments it is observed to take k distinct values for $k \in \mathbb{N}, k < \infty$. $P(\mathbf{O}\backslash\{W\}|do(W))$ is not a proper distribution over $\mathbf{O}$ because $\sum_w P(W = w|do(w), pa(W)) = k$. However, this can be made proper by by normalizing it such that $P(W = w|do(w), pa(W)) = \frac{1}{k}$. Thus, $P(\mathbf{O}\backslash\{W\}|do(W))$ is within a constant factor of $P_Q(\mathbf{O})$, where Q is the member of $\mathcal{U}$ such that $P(W|pa(W)) = P(W) = \frac{1}{k}$ (i.e., where W has a discrete uniform distribution). With respect to the theorem statement, $c = \frac{1}{k}$.

(2) This case follows similarly. Suppose W is continuous and that across environments it is observed to be bounded in the interval $[-M, M]$, $0 < M < \infty$. Then, $P(\mathbf{O}\backslash\{W\}|do(W))$ is not a proper density over $\mathbf{O}$ because $\int_{-M}^{M} f(W = w|do(w), pa(W))dw = 2M$, but this can be made proper by normalizing the pdf of $P(W|pa(W))$ to be $\frac{1}{2M}$. Thus, the level 2 density is within a constant factor of Q, the member of $\mathcal{U}$ where W has a continuous uniform distribution over the interval $[-M, M]$. With respect to the theorem statement, $c = \frac{1}{2M}$. □

Corollary 9. *Stable level two distributions are not, in general, minimax optimal.*

Proof. The following counterexample is adapted from an example in [84].

Consider the DAG $\mathcal{G}$ in Figure A1 in which the goal is to predict Y from W and Z, and the mechanism for generating W (i.e., $P(W|Y)$) varies across environments. The distribution factorizes as $P(Z, W, Y) = P(Z|W, Y)P(W|Y)P(Y)$.

Let all variables be binary, and assume that $P(Y = 1) = \frac{1}{2}$ and $P(Z = 1|W, Y) = \frac{1}{2}$ if $Y = X$ and $P(Z = 1|W, Y) = 1$ otherwise. Finally, we will parameterize $P(W|Y)$ as follows: $P(W = 1|Y = 0) = 1 - \alpha_0$ and $P(W = 1|Y = 1) = \alpha_1$ for $\alpha_0, \alpha_1 \in [0, 1]$. For the Brier score, [84] computed the maximum generalized entropy parameter values to be $\alpha_0 = 2 - \sqrt{2}$ and $\alpha_1 = 2 - \sqrt{2}$.

Thus, the minimax optimal $P(W|Y)$ that yields the maximum generalized entropy $P(Y|W, Z)$ is $P(W = 1|Y = 1) = 2 - \sqrt{2} \approx 0.586$ and $P(W = 1|Y = 0) = \sqrt{2} - 1 \approx 0.414$. This is different than the $P(W|Y)$

that yields the $P(Y|W,Z)$ equivalent to the stable level 2 solution $P(Y|do(W),Z)$, which is $P(W|Y) = P(W) = 0.5$ (by Proposition 6). Thus, the level 2 solution $P(Y|do(W),Z)$ is not optimal for this graph using the Brier score (this also holds for the log loss; see the computations in ref. [84]). □

Proposition 7. *The level* 3 *distribution* $P(Y(W_{B^*})|\mathbf{X}(W_{B^*}))$ *equals* $P_{B^*}(Y|\mathbf{X})$ *and is minimax optimal, where* W_{B^*} *is the counterfactual* W *generated under the mechanism associated with the environment* B^*.

Proof. Given that our training data were generated from $P_0 \in \mathcal{U}$, we are interested in $P(Y|X)$ if counterfactually W had been generated using the mechanism (i.e., we edited the structural equation in the SCM) $g_{B^*}(pa(W), \varepsilon_w)$, the mechanism for W associated with the environment $B^* \in \mathcal{U}$ identified in Theorem 5. This mechanism change produces a new distribution associated with W, $P_{B^*}(W|pa(W))$.

We can represent this counterfactually by letting $Z(W_{B^*}) = Z(g_{B^*}(pa(W))$ be the potential outcome of Z had W been generated according to g_{B^*} for some variable Z (the rhs notation is sometimes used to express policy interventions, see, e.g., [60]). Thus, the counterfactual distribution can be expressed as $P(Y(W_{B^*})|W_{B^*}, pa(W), \mathbf{X}\backslash\{W, pa(W)\}(W_{B^*})) = P(Y(W_{B^*})|\mathbf{X}(W_{B^*}))$ (noting that $pa(W)(W_{B^*}) = pa(W)$ because changing the mechanism of W does not affect its parents). Because P_0 and B^* differ only with respect to the mechanism generating W, the counterfactual distribution associated with this mechanism change yields $P_{B^*}(Y|X)$. Thus, $P(Y(W_{B^*})|\mathbf{X}(W_{B^*}))$ is minimax optimal because $P_{B^*}(Y|X)$ was shown to be minimax optimal in Theorem 5. □

C Likelihood reweighting in the presence of unobserved confounders

In Section 4, we developed a likelihood reweighting formulation for a minimax optimal predictor by assuming that the graph has no bidirected edges (no unobserved confounders). We now relax this condition.

First, note that the *c-component* (or *district*) of a variable in an ADMG is the set of nodes reachable via purely bidirected paths (i.e., paths of the form $V_1 \leftrightarrow \cdots \leftrightarrow V_2$). An ADMG over variables $\mathbf{O}$ factorizes as follows:

$$Q[\mathbf{O}] = P(\mathbf{O}) = \sum_{\mathbf{U}} \prod_{O_i \in \mathbf{O}} P(O_i|pa(O_i), U_i)P(\mathbf{U}),$$

where $\mathbf{U}$ are the exogenous noise variables. Note that an ADMG factorizes as a product of Q-factors over the c-components. That is, if $\mathbf{O}$ is partitioned into c-components $\{S_1, S_2, \ldots, S_k\}$, then $P(\mathbf{O}) = \prod_1^k Q[S_k]$ [85, Lemma 7].[5] Finally, let $O_1 < O_2 < \cdots < O_n$ be a topological order over $\mathbf{O}$. Then, each c-factor is identifiable and given by $Q[S_j] = \prod_{i|O_i \in S_j} P(O_i|pa(T_i)\backslash\{O_i\})$, where T_i is the c-component of $\mathcal{G}_{O_1 \ldots O_i}$ that contains O_i.

Now we can see that the term $P(O_i|pa(T_i)\backslash\{O_i\})$ is the ADMG generalization of $P(O_i|pa(O_i))$ in DAGs and is the term associated with the mechanism for generating O_i. Thus, if $P(W|do(pa(W)))$ is identifiable in the ADMG, then we we need to perform likelihood reweighting with respect to $P(W|pa(T_W)\backslash\{W\})$. That is, let $B^* \in \mathcal{U}$ be the minimax optimal training distribution/environment. Then,

$$E_{B^*}[\ell(f(\mathbf{x}, y))] = E_{P_0}\left[\frac{P_{B^*}(\mathbf{O})}{P_{P_0}(\mathbf{O})}\ell(f(\mathbf{x}, y))\right] = E_{P_0}\left[\frac{P_{B^*}(W|pa(T_W)\backslash\{W\})}{P_{P_0}(W|pa(T_W)\backslash\{W\})}\ell(f(\mathbf{x}, y))\right]$$

and we can define a reweighting function as mentioned earlier.

5 When the graph has no bidirected edges, each node is its own c-component.