A Causal Framework for Distribution Generalization

Rune Christiansen, Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco, and Jonas Peters

Abstract—We consider the problem of predicting a response $Y$ from a set of covariates $X$ when test- and training distributions differ. Since such differences may have causal explanations, we consider test distributions that emerge from interventions in a structural causal model, and focus on minimizing the worst-case risk. Causal regression models, which regress the response on its direct causes, remain unchanged under arbitrary interventions on the covariates, but they are not always optimal in the above sense. For example, for linear models and bounded interventions, alternative solutions have been shown to be minimax prediction optimal. We introduce the formal framework of distribution generalization that allows us to analyze the above problem in partially observed nonlinear models for both direct interventions on $X$ and interventions that occur indirectly via exogenous variables $A$. It takes into account that, in practice, minimax solutions need to be identified from data. Our framework allows us to characterize under which class of interventions the causal function is minimax optimal. We prove sufficient conditions for distribution generalization and present corresponding impossibility results. We propose a practical method, NILE, that achieves distribution generalization in a nonlinear IV setting with linear extrapolation. We prove consistency and present empirical results.

Index Terms—Distribution generalization, causality, worst-case risk, distributional robustness, invariance, domain adaptation

1 INTRODUCTION

LARGE-SCALE learning systems, particularly those focusing on prediction tasks, have been successfully applied in various domains of application. Since inference is usually done during training time, any difference between training and test distribution poses a challenge for prediction methods [4], [15], [44], [51]. Dealing with these differences is of great importance in several fields such as environmental sciences, where methods need to extrapolate both in space and time. Tackling this problem requires restrictions on how the distributions may differ, since, clearly, generalization becomes impossible if the test distribution may be arbitrary. Given a response $Y$ and some covariates $X$, several existing procedures aim to find a minimax function $f$ which minimizes the worst-case risk $\sup_{P \in \mathcal{P}} \mathbb{E}_P[(Y - f(X))^2]$ across distributions contained in a small neighborhood $\mathcal{N}$ of the training distribution. The neighborhood $\mathcal{N}$ should be representative of the difference between the training and test distributions, and often mathematical tractability is taken into account, too [1], [61]. A typical approach is to define a $\rho$-ball of distributions $\mathcal{N}_\rho(P_0) := \{P : D(P, P_0) \leq \rho\}$ around the (empirical) training distribution $P_0$, with respect to some divergence measure $D$, such as the Kullback-Leibler divergence [6], [31]. While some divergence functions only consider distributions with the same support as $P_0$, the Wasserstein distance allows for a neighborhood of distributions around $P_0$ with possibly different supports [1], [10], [20], [61].

In our analysis, we do not start from a divergence measure, but instead model the difference between training and test distribution using the concept of interventions [45], [47]. We believe that for many problems this provides a useful description of distributional changes. We will see that, depending on the considered setup, this approach allows to find models that perform well even on test distributions which would be considered far away from the training distribution in any commonly used metric. For this class of distributions, causal regression models appear naturally because of the following well-known observation. A prediction model, which uses only the direct causes of the response $Y$ as covariates, is invariant under interventions on variables other than $Y$: the conditional distribution of $Y$ given its causes does not change (this principle is known, e.g., as invariance, autonomy or modularity) [2], [25], [45]. Such a causal regression model yields the minimal worst-case risk when considering all interventions on variables other than $Y$ [53, e.g., Theorem 1, Appendix]. It has therefore been suggested to use causal models in problems of distributional shifts [4], [28], [36], [39], [49], [53], [57]. In practice, however, not all relevant causal variables might be observed. One may further argue that causal methods are too conservative in that the interventions which induce the test distributions may not be arbitrarily strong. Instead, methods which focus on a trade-off between predictability and causality have been proposed for linear models [48], [54], see also Section 5.1. Anchor regression [54] is shown to be predictive optimal under a set of bounded interventions.

In this work, we introduce the general framework of distribution generalization, which permits a unifying perspective on the potentials and limitations of applying causal concepts...
to the problem of generalizing regression models from training to test distribution. In particular, we use it to characterize the relationship between a minimax optimal solution and the causal function, and to classify settings under which the minimax solution is identifiable from the training distribution.

1.1 Further Related Work
The field of distributional robustness or out-of-distribution generalization aims to develop procedures that are robust to changes between training and test distribution. This problem has been actively studied from an empirical perspective in machine learning research, for example, in image classification by using adversarial attacks, where small digital [23] or physical [21] perturbations of pictures can deteriorate the performance of a model. Arguably, these procedures are not yet fully understood theoretically. A more theoretical perspective is given by the previously mentioned minimization of a worst-case risk across distributions contained in a neighborhood of the training distribution, in our case, distributions generated by interventions.

Our framework includes the problems of multi-task learning, domain generalization and transfer learning [8], [13], [37], [50] (see Section 2.4 for more details), with a focus on minimizing the worst-case risk. In settings of covariate shift, e.g., [58], [62], [63], one usually assumes that the training and test distribution of the covariates are different, while the conditional distribution of the response given the covariates remains invariant [9], [17], [18], [41]. Sometimes, it is additionally assumed that the support of the training distribution covers that of the test distribution [58]. In this work, the conditional distribution of the response given the covariates is allowed to change between interventions, due to the existence of hidden confounders, and we consider settings where the test observations lie outside the training support.

Data augmentation methods have become successful techniques, e.g., in image classification, to adapt prediction procedures to such types of distribution shifts. These methods increase the diversity of the training data by changing the geometry and the color of the images (e.g., by rotation, cropping or changing saturation) [59], [67]. This allows the user to create models that generalize better to unseen environments, e.g., [65]. We view these approaches as ways to enlarge the support of the covariates, which, as our results show, comes with theoretical advantages, see Section 4.

Minimizing the worst-case risk is considered in robust methods [19], [34], too. It can also be formulated in terms of minimizing the regret in a multi-armed bandit problem [5], [7], [35]. In that setting, the agent can choose the distribution which generates the data. In our setting, though, we do not assume to have control over the interventions, and, hence, neither over the distribution of the sampled data.

1.2 Contribution and Structure
This work contains four main contributions: (1) A novel framework for analyzing the problem of generalization from training to test distribution, using the notion of distribution generalization (Section 2). (2) Results elucidating the relationship between a causal function and a minimax solution (Section 3). (3) Sufficient conditions which ensure distribution generalization, along with corresponding impossibility results (Section 4). (4) A practical method, called NILE (‘Non-linear Intervention-robust Linear Extrapolator’), which learns a minimax solution from i.i.d. observational data (Section 5).

Our framework describes how structural causal models can be used as technical devices for modeling plausible test distributions. It further allows us to formally define distribution generalization, which describes the ability to identify generalizing regression models (i.e., minimax solutions) from the observational distribution. While it is well known that the causal function is minimax optimal under the set of all interventions on the covariates, e.g., [53], we extend this result in several ways, for example, by allowing for hidden variables and by characterizing more general sets of interventions under which the causal function is minimax optimal. We further derive conditions on the model class, the observational distribution and the family of interventions under which distribution generalization is possible, and present impossibility results proving the necessity of some of these conditions. For example, we show that strong assumptions on the functional relationship between X and Y are needed whenever the interventions extend the training support of X. An example of such an assumption is to consider the class of differentiable functions that linearly extrapolate outside the support of X. For that model class, we propose the explicit method NILE, which obtains distribution generalization by exploiting a non-linear instrumental variables setup. We show that our method learns a minimax solution which corresponds to the causal function. We prove consistency and compare our algorithm to state-of-the-art approaches empirically.

We believe that our results shed some light on the potential merits of using causal concepts in the context of generalization. The framework allows us to make first steps towards answering when it can be beneficial to use non-causal functions for prediction under interventions, and what might happen under misspecification of the intervention class. Our results also formalize in which sense methods that generalize in the linear case — such as IV and anchor regression [54] — can be extended to nonlinear settings. Further, our framework implies impossibility statements for multi-task learning that relate to existing results [18].

Our code is available as an R-package at https://runesen.github.io/NILE; scripts generating all our figures and results can be found at the same url. Additional supporting material is given in the online appendix, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2021.3094760. Appendix A shows how to represent several causal models in our framework, available in the online supplemental material. Appendix B summarizes existing results on identifiability in IV models, available in the online supplemental material. Appendix C provides details on the test statistic that we use for NILE, available in the online supplemental material. Appendix D contains additional experiments, available in the online supplemental material. All proofs are provided in Appendix E, available in the online supplemental material.

2 Framework
For a real-valued response $Y \in \mathbb{R}$ and predictors $X \in \mathbb{R}^d$, we consider the problem of identifying a regression function
that works well not only on the training data, but also under perturbed distributions that we will model by interventions.

2.1 Modeling Intervention-Induced Distributions

We require a model that is able to model an observational distribution of \((X, Y)\) (as training distribution) and the distribution of \((X, Y)\) under a class of interventions on (parts of) \(X\) (as test distribution). We will do so by means of a structural causal model (SCM) \([11], [45]\). More precisely, denoting by \(H \in \mathbb{R}^d\) some additional (unobserved) variables, we consider the SCM

\[
H := \varepsilon_H, \quad X := h_2(H, \varepsilon_X), \quad Y := f(X) + h_1(H, \varepsilon_Y),
\]

where the assignments for \(H, X\) and \(Y\) consist of \(q, d\) and 1 coordinate(s), respectively. Here, \(f, h_1\) and \(h_2\) are measurable functions, and the innovation terms \(\varepsilon_X, \varepsilon_Y\) and \(\varepsilon_H\) are independent vectors with possibly dependent coordinates. Two comments are in order. First, the joint distribution of \((X, Y)\) is constrained only by requiring that \(X\) and \(h_1(H, \varepsilon_Y)\) enter the assignment for \(Y\) additively. This constraint affects the allowed conditional distributions of \(Y\) given \(X\), but does not make any restriction on the marginal distributions of either \(X\) or \(Y\). Second, we only use the above SCM as a technical device for modeling training and test distributions, by considering interventions on \(X\) or \(A\) (introduced in Section 2.3), for which we are analyzing the predictive performance of different models – similarly to how one could have considered a ball around the training distribution. We therefore only require the SCM to correctly (a) model the training-distribution, and (b) induce the test-distributions through interventions. Any other causal implications of the SCM, such as causal orderings between variables, causal effects or counterfactual statements, are not assumed to be correctly specified. As such, our framework includes a wide range of cases, including situations where training and test distribution come from interventions in an SCM with a different structure than (1), where, for example, some of the variables in \(X\) are not ancestors but descendants of \(Y\). To see whether our framework applies, one needs to check if the considered training and test distributions can be equivalently expressed as interventions in a model of our form. If the structure of the true data generating SCM is known, this can be done by directly transforming the SCM and the interventions. The following remark shows an example of such a transformation and may be interesting to readers with a special interest in causality. It can be skipped at first reading.

Remark 1 (Transforming causal models). Assume that

\[
X_1 := \varepsilon_1, \quad X_2 := k(Y) + \varepsilon_2, \quad Y := f(X_1) + \varepsilon_3,
\]

with \((\varepsilon_1, \varepsilon_2, \varepsilon_3) \sim Q\), and that we consider test distributions arising from shift interventions on \(X_2\). This set of training and test distributions can be equivalently modeled by the reduced SCM

\[
H := \varepsilon_3, \quad X := h_2(H, (\varepsilon_1, \varepsilon_2)), \quad Y := f(X_1) + H,
\]

with \((\varepsilon_1, \varepsilon_2, \varepsilon_3) \sim Q\), and where \(h_2\) is defined by \(h_2(H, (\varepsilon_1, \varepsilon_2)) := (\varepsilon_1, k(f(\varepsilon_1) + H) + \varepsilon_2)\). Both SCMs induce the same observational distribution over \((X_1, X_2, Y)\) and shift interventions on \(X_2\) in the original SCM correspond to shift interventions on \(X = (X_1, X_2)\) in the reduced SCM (where only the second coordinate is shifted). Our framework can then be used, for example, to give sufficient conditions under which generalization (formally defined below) is possible, see Proposition 4.3 and 4.4.

It is not always possible to transform an SCM into our reduced form, and it might also happen that the transformed interventions are not covered by our framework. For example, we do not allow for direct interventions on \(Y\) in the original model. In other cases, where the original SCM may contain additional hidden variables, even interventions on (parts of) \(X\) in the original SCM may translate into interventions on \(H\) in the reduced SCM, and are therefore not covered. Details and a more general treatment are provided in Appendix A, available in the online supplemental material.

Sometimes, the vector of covariates \(X\) contains variables, which are independent of \(H\), that enter into the assignments of the other covariates additively and cannot be used for the prediction (e.g., because they are not observed during testing). If such covariates exist, it can be useful to explicitly distinguish them from the remaining predictors. We will denote them by \(A\) and call them exogenous variables. Such variables are interesting for several reasons. (i) We will see that in general, interventions on \(A\) lead to intervention distributions with desirable properties for distribution generalization, see Section 4.4. (ii) Some of our results rely on the function \(f\) being identifiable from the observational distribution, see Assumption 1 below. The variables \(A\) can be used to state explicit conditions for identifiability. Under additional assumptions, for example, they can be used as instrumental variables, e.g., \([12], [24]\), a well-established tool for recovering \(f\) from the observational distribution of \((X, Y, A)\). (iii) The variable \(A\) can be used to model a covariate that is not observed under testing. It can also be used to index tasks (which we discuss at the end of Section 2.4). In the remainder of this work, we therefore consider a slightly larger class of SCMs that also includes exogenous variables \(A\). It contains the SCM (1) as a special case.\(^1\) We derive results for settings with and without exogenous variables \(A\).

2.2 Model

Formally, we consider a response \(Y \in \mathbb{R}^1\), covariates \(X \in \mathbb{R}^d\), exogenous variables \(A \in \mathbb{R}^r\), and unobserved variables \(H \in \mathbb{R}^l\). Let further \(\mathcal{F} \subseteq \{f: \mathbb{R}^d \rightarrow \mathbb{R}\}, \mathcal{G} \subseteq \{g: \mathbb{R}^r \rightarrow \mathbb{R}\}\), \(\mathcal{H}_1 \subseteq \{h_1: \mathbb{R}^{r+1} \rightarrow \mathbb{R}\}\) and \(\mathcal{H}_2 \subseteq \{h_2: \mathbb{R}^{rd} \rightarrow \mathbb{R}\}\) be fixed sets of measurable functions. Moreover, let \(Q\) be a collection of probability distributions on \(\mathbb{R}^{d+1+rd}\), such that for all \(Q \in Q\) it holds that if \((\varepsilon_X, \varepsilon_Y, \varepsilon_A, \varepsilon_H) \sim Q\), then \(\varepsilon_X, \varepsilon_Y, \varepsilon_A\) and \(\varepsilon_H\) are jointly independent, and for all \(h_1 \in \mathcal{H}_1\) and \(h_2 \in \mathcal{H}_2\) it holds that \(\varepsilon_Y := h_1(\varepsilon_H, \varepsilon_Y)\) and \(\varepsilon_X := h_2(\varepsilon_H, \varepsilon_X)\) have mean zero.\(^2\) Let \(\mathcal{M} := \mathcal{F} \times \mathcal{G} \times \mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{Q}\) denote the model class. Every model \(M = (f, g, h_1, h_2, Q) \in \mathcal{M}\) then specifies

1. This follows from choosing \(A\) as an independent noise variable and a constant \(g\).
2. This can be assumed w.l.o.g. if \(\mathcal{F}\) and \(\mathcal{G}\) are closed under addition and scalar multiplication, and contain the constant function.
an SCM by 3

\[
\begin{align*}
A & := \varepsilon_A \\
H & := \varepsilon_H \\
X & := (g(A) + h_2(H, \varepsilon_X)) \\
Y & := f(X) + h_1(H, \varepsilon_Y)
\end{align*}
\]

with \((\varepsilon_X, \varepsilon_Y, \varepsilon_A, \varepsilon_H) \sim Q\), where the assignments for \(A, H, X\) and \(Y\) consist of \(r, q, d\) and 1 coordinate(s), respectively. For each model \(M = (f, g, h_1, h_2, Q) \in \mathcal{M}\), we refer to \(f\) as the causal function (for the pair \((X, Y)\)), and denote by \(\mathbb{P}_M\) the joint distribution over the observed variables \((X, Y, A)\). We assume that this distribution has finite second moments. If no exogenous variables \(A\) exist, one can think of the function \(g\) as being constant. A model \(M\) that correctly models the training and test distributions will be referred to as the ‘true model’.

2.3 Interventions

Each SCM \(M \in \mathcal{M}\) can now be modified by the concept of interventions, e.g., [45], [47]. An intervention corresponds to replacing one or more of the structural assignments of the SCM (see Section 4.2 for details). For example, we intervene on some of the covariates \(X\) by replacing the corresponding assignments with, e.g., a Gaussian random vector that is independent of the other noise variables. Importantly, an intervention on some of the variables does not change the assignment of any other variable. In particular, an intervention on \(X\) does not change the conditional distribution of \(Y\), given \(X\) and \(H\) (this is an instance of the invariance property mentioned in Section 1) but it may change the conditional distribution of \(Y\), given \(X\).

The problems addressed in this work require us to simultaneously consider several different SCMs that are all subject to the same (set of) interventions. Formally, we therefore regard an intervention \(i\) as a mapping from the model class \(\mathcal{M}\) into a (possibly larger) set of SCMs, which takes as input a model \(M \in \mathcal{M}\) and outputs another model \(M(i)\) over variables \((X', A', Y', H')\), the intervened model. We do not need to assume that the intervened model \(M(i)\) belongs to the model class \(\mathcal{M}\) but we require that \(M(i)\) induces a joint distribution over \((X', Y', A', H')\) with finite second moments. We denote the corresponding distribution over the observed \((X', Y', A')\) by \(\mathbb{P}_{M(i)}\) and use \(I\) for a collection of interventions. In our work, the test distributions are modeled as distributions generated by these types of intervened models, and the set \(I\) therefore indexes the set of test distributions. We will be interested in the mean squared prediction error on each test distribution \(i\), formally written as \(\mathbb{E}_{M(i)}[(Y - f_i(X))^2]\). (In this work, we consider a univariate \(Y\), but writing \(\mathbb{E}[(Y - f_i(X))^2] = \sum_{j=1}^d \mathbb{E}[(Y_j - f_i,j(X))^2]\), most our results extend straightforwardly to a d-dimensional response.)

The support of random variables under interventions will play an important role for the analysis of distribution

generalization. Throughout this paper, \(\text{supp}^M(Z)\) denotes the support of the random variable \(Z \in \{A, X, H, Y\}\) under the distribution induced by the SCM \(M \in \mathcal{M}\). Moreover, \(\text{supp}^i(Z)\) denotes the union of \(\text{supp}^{M(i)}(Z)\) over all interventions \(i \in I\). We call a collection of interventions on \(Z\) support-reducing (w.r.t. \(M\)) if \(\text{supp}^i(Z) \subseteq \text{supp}^M(Z)\) and support-extending (w.r.t. \(M\)) if \(\text{supp}^i(Z) \not\subseteq \text{supp}^M(Z)\). Whenever it is clear from the context which model is considered, we may drop the indication of \(M\) altogether and simply write \(\text{supp}(Z)\).

2.4 Distribution Generalization

Let \(\mathcal{M}\) be a fixed model class, let \(M = (f, g, h_1, h_2, Q) \in \mathcal{M}\) and let \(I\) be a class of interventions. In this work, we aim to find a function \(f^* : \mathbb{R}^d \to \mathbb{R}\) such that the predictive model \(Y = f^*(X)\) has low worst-case risk over all test distributions induced by the interventions \(I\) in model \(M\). We therefore consider, for the true \(M\), the optimization problem

\[
\min_{f \in \mathcal{F}} \sup_{i \in I} \mathbb{E}_{M(i)}[(Y - f_i(X))^2],
\]

where \(\mathbb{E}_{M(i)}\) is the expectation in the intervened model \(M(i)\). In general, this optimization problem is neither guaranteed to have a solution, nor is the solution, if it exists, ensured to be unique. Whenever a solution \(f^*\) to (2) exists, we refer to it as a minimax solution (for model \(M\) w.r.t. \((I, \mathcal{F})\)).

Depending on the model class \(M\), there may be several models \(M \in \mathcal{M}\) that induce the observational distribution \(\mathbb{P}_M\), that is, the same distribution over the observed variables \(A, X\) and \(Y\), but do not agree with \(M\) on all intervention distributions induced by \(I\). Thus, each such model induces a potentially different minimax problem with different solutions. Given knowledge only of \(\mathbb{P}_M\), it is therefore generally not possible to identify a solution to (2). In this paper, we study conditions on \(\mathcal{M}, \mathbb{P}_M\) and \(I\), under which this becomes possible. More precisely, we aim to characterize under which conditions \((\mathbb{P}_M, \mathcal{M})\) admits distribution generalization to \(I\).

Definition 2.1 (distribution generalization). \((\mathbb{P}_M, \mathcal{M})\) is said to admit distribution generalization to \(I\), or simply to admit generalization to \(I\), if for every \(\varepsilon > 0\) there exists a function \(f^\varepsilon \in \mathcal{F}\) such that, for all models \(M \in \mathcal{M}\) with \(\mathbb{P}_M = \mathbb{P}_M\), it holds that

\[
\left| \sup_{i \in I} \mathbb{E}_{M(i)}[(Y - f^\varepsilon_i(X))^2] - \inf_{f_i \in \mathcal{F}} \sup_{i \in I} \mathbb{E}_{M(i)}[(Y - f_i(X))^2] \right| \leq \varepsilon.
\]

Distribution generalization does not require the existence of a minimax solution in \(\mathcal{F}\) (which would require further assumptions on the function class \(\mathcal{F}\)) and instead focuses on whether an approximate solution can be identified based only on the observational distribution \(\mathbb{P}_M\). If, however, there exists a function \(f^\varepsilon \in \mathcal{F}\) for every \(M \in \mathcal{M}\) with \(\mathbb{P}_M = \mathbb{P}_M\), then, in particular, \((\mathbb{P}_M, \mathcal{M})\) admits generalization to \(I\).

Our framework also includes several settings of multi-task learning (MTL) and domain adaptation [50], where one often assumes to observe different training tasks. In MTL, one is then interested in using the different tasks to improve
the predictive performance on either one or all training tasks – this is often referred to as asymmetric and symmetric MTL, respectively. In our framework, such a setup can be modeled using a categorical variable $X$. If, however, one is interested in predicting on an unseen task or if one does not know which of the observed tasks the new test data come from, one may instead use a categorical $A$ with support-extending or support-reducing interventions, respectively.

3 **Minimax Solutions and the Causal Function**

To address the question of distribution generalization, we first study properties of the minimax optimization problem (2). In the simplest case, where $I$ consists only of the trivial intervention, that is, $P_M = P_M(i)$, we are looking for the best predictor on the observational distribution. In that case, the minimax solution is attained at any conditional mean function, $f^*: x \mapsto \mathbb{E}[Y|X = x]$ (provided that $f^* \in \mathcal{F}$). For larger classes of interventions, however, the conditional mean may become sub-optimal in terms of prediction. To see this, it is instructive to decompose the risk under an intervention.

Let $\xi_Y$ denote the small-$\theta$ intervention, that is confounding-removing. Then, the minimax solution remains unchanged if an even stronger statement holds: The causal function on the test distributions.

$M$ is already a minimax solution if $M$ contains at least one confounding-removing intervention on $X$. This choice, however, comes at a cost: it relies on the fact that we know the class of interventions $I$. In general, being a minimax solution is not entirely robust with respect to misspecification of $I$. In particular, if the set $I_1$ of interventions describing the test distributions is misspecified by a set $I_1 \neq I_2$, then the considered minimax solution with respect to $I_1$ may perform worse than the causal function on the test distributions.

Proposition 3.3 (Properties of the minimax solution under mis-specified interventions). Let $I_1$ and $I_2$ be any two sets of interventions on $X$, and let $f^*_1 \in \mathcal{F}$ be a minimax solution w.r.t. $I_1$. Then, if $I_2 \subseteq I_1$, it holds that

$$\sup_{i \in I_2} \mathbb{E}_{M(i)}[(Y - f^*_i(X))^2] \leq \sup_{i \in I_1} \mathbb{E}_{M(i)}[(Y - f(X))^2].$$

If $I_2 \not\subseteq I_1$, however, it can happen (even if $\mathcal{F}$ is linear) that

$$\sup_{i \in I_2} \mathbb{E}_{M(i)}[(Y - f^*_i(X))^2] > \sup_{i \in I_2} \mathbb{E}_{M(i)}[(Y - f(X))^2].$$

The second part of the proposition should be understood as a non-robustness property of non-causal minimax solutions. Improvements on the causal function are possible in situations, where one has reasons to believe that the test distributions do not stem from a set of interventions that is much larger than the specified set.

4 **Distribution Generalization**

As described in Section 2.4, we consider a fixed model class $M$ containing the true (but unknown) model $M$, and let $I$ be a class of interventions. By definition, the optimizer of the minimax problem (2) depends on the true model $M$. The unbounded eigenvalue condition above is satisfied if $I$ is the set of all shift interventions on $X$. These interventions, formally defined in Section 4.2.2, appear in linear IV models and recently gained further attention in the causal community [54], [55]. The proposition above considers a linear function class $\mathcal{F}$; in this way, shift interventions are related to linear models.
Section 3 relates this optimizer to the causal function $f$, whose knowledge, too, requires knowing $M$. In practice, however, we do not have access to the true model $M$, but only to its observational distribution $P_M$. This motivates the notion of distribution generalization, see (3). In words, it states that approximate minimax solutions (which depend on the intervention distributions $P_{M(i)}$, $i \in I$) are identified from the observational distribution $P_M$. This holds true, in particular, if the intervention distributions themselves are identified from $P_M$.

**Proposition 4.1 (Sufficient conditions for distribution generalization).** Assume that for all $M \in \mathcal{M}$ it holds that

$$P_M = P_M \implies P_{M(i)}^{(X,Y)} = P_{M(i)}^{(X,Y)} \forall i \in I,$$

where $P_{M(i)}^{(X,Y)}$ is the joint distribution of $(X,Y)$ under $M(i)$. Then, $(P_M, \mathcal{M})$ admits generalization to $I$.

Proposition 4.1 provides verifiable conditions for distribution generalization, and can be used to prove possibility statements. It is, however, not a necessary condition. Indeed, we will see that, under certain types of interventions, distribution generalization becomes possible even in cases where the interventional marginal of $X$ is not identified.

In this section, we study conditions on $\mathcal{M}$, $P_M$ and $I$ which ensure generalization, and present corresponding impossibility results proving the necessity of some of these conditions. Two aspects will be of central importance. The first is related to causal identifiability, i.e., whether the causal function $f$ is sufficiently identified from the observational distribution $P_M$ (Section 4.1). The other aspect is related to the types of interventions (Section 4.2). We consider interventions on $X$ in Section 4.3 and interventions on $A$ in Section 4.4. Parts of our results are summarized in Table I.

### 4.1 Identifiability of the Causal Function

For specific types of interventions, the causal function $f$ is itself a minimax solution, see Propositions 3.1 and 3.2. If, in addition, these interventions are support-reducing, generalization is directly implied by the following assumption.

**Assumption 1 (Identifiability of $f$ on the support of $X$).**

For all $M = (\mathbf{f}, \ldots) \in \mathcal{M}$ with $P_M = P_M$, it holds that $\mathbf{f}(x) = f(x)$ for all $x \in \text{supp}(X)$.

Assumption 1 will play a central role in proving distribution generalization even in situations where the causal function is not a minimax solution. We use it as a starting point for most of our results. The assumption is violated, for example, in a linear Gaussian setting with a single covariate $X$ (without $A$). Here, in general, we cannot identify $f$ and distribution generalization does not hold. Assumption 1, however, is not necessary for generalization. In Section 4.4 we discuss a linear setting where distribution generalization is possible, even if Assumption 1 does not hold.

The question of causal identifiability has received a lot of attention in the literature. In linear instrumental variables settings, for example, one assumes that the functions $f$ and $g$ are linear and identifiability follows if the product moment between $A$ and $X$ has rank at least the dimension of $X$, e.g., [66]. In linear non-Gaussian models, one can identify the function $f$ even if there are no instruments [30]. For nonlinear models, restricted SCMs can be exploited, too. In that case, Assumption 1 holds under regularity conditions if $h_1(H, c_Y)$ is independent of $X$ [46, 47, 68] and first attempts have been made to extend such results to non-trivial confounding cases [33]. The nonlinear IV setting, e.g., [3], [42], [43] is discussed in more detail in Appendix B, available in the online supplemental material, where we give a brief overview of identifiability results for linear, parametric and non-parametric function classes. Assumption 1 states that $f$ is identifiable, even on $P_M$-null sets, which is usually achieved by placing further constraints on the function class, such as smoothness. Even though this issue seems technical, it becomes important when considering hard interventions that set $X$ to a fixed value, for example.

### 4.2 Types of Interventions

Whether distribution generalization is admitted depends on the intervention class $I$. In this work, we only consider interventions on the covariates $X$ and $A$. Each of these types of interventions can be characterized by a measurable function $\psi^i$, which determines the structural assignment of the intervened variable, and a (possibly degenerate) random vector $I^i$, which serves as an independent noise innovation. More formally, for an intervention on $X$, the pair $(\psi^i, I^i)$ defines the intervention which maps the input model $M = (f, g, h_1, h_2, Q) \in \mathcal{M}$ to the intervened model $M(i)$ given by the assignments

$$A^i := \varepsilon_A^i, \quad H^i := \varepsilon_H^i,$$
$$X^i := \psi(g, h_2, A^i, H^i, c^i_X, I^i),$$
$$Y^i := f(X^i) + h_1(H^i, c_Y^i).$$

Similarly, for an intervention on $A$, $(\psi^i, I^i)$ specifies the intervention which outputs

$$A^i := \psi(I^i, \varepsilon_A^i), \quad H^i := \varepsilon_H^i,$$
$$X^i := g(A^i) + h_2(H^i, c^i_X),$$
$$Y^i := f(X^i) + h_1(H^i, c_Y^i).$$

In both cases, $(c^i_X, \varepsilon_Y^i, \varepsilon_A^i, \varepsilon_H^i) \sim Q$ and $I^i \perp (c^i_X, \varepsilon_Y^i, \varepsilon_A^i, \varepsilon_H^i)$. We will see below that this class of interventions is rather flexible. It does, however, not allow for arbitrary manipulations of $M$. For example, it does not allow for changes in the structural assignments for $Y$ or $H$, or for the noise variable $\varepsilon_Y^i$ to enter the assignment of the intervened variable. As
the following section highlights, further constraints on the types of interventions are necessary to ensure distribution generalization.

4.2.1 Impossibility of Generalization Without Constraints on the Interventions

Let \( Q \) be a class of product distributions on \( \mathbb{R}^4 \), such that for all \( Q \in \mathcal{Q} \), the coordinates of \( Q \) are non-degenerate, zero-mean with finite second moment. Let \( \mathcal{M} \) be the class of all models of the form

\[
A := \varepsilon_A, H := \sigma_H, \quad X := \gamma_A + \varepsilon_X + \frac{1}{\sigma} H, \quad Y := \beta_X + \varepsilon_Y + \frac{1}{\sigma} H,
\]

with \( \gamma, \beta \in \mathbb{R}, \sigma > 0 \) and \( (\varepsilon_A, \varepsilon_X, \varepsilon_Y, \varepsilon_H) \sim Q \in \mathcal{Q} \). Assume that \( P_M \) is induced by some model \( M = M(\gamma, \beta, \sigma, Q) \) from the above model class (here, we slightly adapt the notation from Section 2). The following proposition shows that, without constraining the set of interventions \( \mathcal{I} \), distribution generalization is not always ensured.

Proposition 4.2 (Impossibility of generalization without constraining the class of interventions). Assume that \( \mathcal{M} \) is given as defined above, let \( \mathcal{I} \subseteq \mathbb{R}^2 \) be a compact, non-empty set and define the interventions on \( X \) by \( \psi'(g, h_2, A', H', \varepsilon_X', I') = \chi_i H, \) for \( i \in \mathcal{I} \). Then, \( P_M \), \( M \) does not admit generalization to \( \mathcal{I} \) (even if Assumption 1 is satisfied). In addition, any prediction model other than the causal model may perform arbitrarily bad under the interventions \( \mathcal{I} \). That is, for any \( b \neq \beta \) and any \( c > 0 \), there exists a model \( \tilde{M} \in \mathcal{M} \) with \( P_M = P_{\tilde{M}} \), such that

\[
\left| \sup_{i \in \mathcal{I}} E_{\tilde{M}(i)}[(Y - bX)^2] - \inf_{b \in \mathbb{R}} \sup_{i \in \mathcal{I}} E_{\tilde{M}(i)}[(Y - bX)^2] \right| \geq c.
\]

We now give some intuition about the above result. By definition, distribution generalization is ensured if there exist prediction functions that are (approximately) minimax optimal for all models which induce the same observational distribution as \( M \). Since, in the above example, the distribution of \((X, Y, A)\) does not depend on \( \sigma \), this includes all models of the form \( \tilde{M}_\sigma = M(\gamma, \beta, \tilde{\sigma}, Q) \) for some \( \tilde{\sigma} > 0 \). However, while agreeing on the observational distribution, each of these models induces fundamentally different intervention distributions (under \( M_i(i), (X, Y) \) is equal in distribution to \((i \varepsilon_H, (\beta i + 1/\sqrt{2}) \varepsilon_H))\) and results in different (approximate) minimax solutions. Below, we introduce two types of interventions which ensure distribution generalization in a wide range of settings by constraining the influence of \( H \) on \( X \).

4.2.2 Interventions Which Allow for Generalization

In Section 3, we already introduced confounding-removing interventions, which break the dependence between \( X \) and \( H \). For an intervention set \( \mathcal{I} \) which contains at least one confounding-removing intervention, the causal function \( f \) is always a minimax solution (see Proposition 3.1) and, in the case of support-reducing interventions, distribution generalization is therefore achieved by requiring Assumption 1 to hold. The intervention \( i \) with intervention map \( \psi' \) is called confounding-preserving if there exists a map \( \phi' \), such that

\[
\psi'(g, h_2, A', H', \varepsilon_X', I') = \phi'(A', g(A'), h_2(H'), \varepsilon_X'), I').
\]

Confounding-preserving interventions contain, e.g., shift interventions on \( X \), which linearly shift the original assignment by \( I \), that is, \( \psi'(g, h_2, A', H', \varepsilon_X', I') = g(A') + h_2(H', \varepsilon_X') + \varepsilon_X' \), which is the same as in the original model. The name ‘confounding-preserving’ stems from the fact that the confounding variables \( H \) only enter the intervened structural assignment of \( X \) via the term \( h_2(H', \varepsilon_X') \), which is the same as in the original model. This property fails to hold true for the interventions in Proposition 4.2. If \( I \) consists only of confounding-preserving interventions, the causal function is generally not a minimax solution. However, we will see that, under Assumption 1, these types of interventions lead to identifiability of the intervention distributions \( P_{M(i)} \), \( i \in \mathcal{I} \), and therefore ensure generalization via Proposition 4.1.

Some interventions are both confounding-removing and confounding-preserving, but not every confounding-removing intervention is confounding-preserving. For example, the intervention \( \psi'(g, h_2, A', H', \varepsilon_X') = \varepsilon_X' \) is confounding-removing but, in general, not confounding-preserving. Similarly, not all confounding-preserving interventions are confounding-removing. We call a set of interventions \( \mathcal{I} \) well-behaved either if it consists only of confounding-preserving interventions or if it contains at least one confounding-removing intervention.

4.3 Generalization to Interventions on \( X \)

We now formally prove in which sense the two types of interventions defined above allow for distribution generalization. We will see that this question is closely linked to the relation between the support of \( P_M \) and the support of the intervention distributions. Below, we therefore distinguish between support-reducing and support-extending interventions on \( X \).

4.3.1 Support-Reducing Interventions

For support-reducing interventions, Assumption 1 is sufficient for distribution generalization even in nonlinear settings, under a large class of interventions.

Proposition 4.3 (Generalization to support-reducing interventions on \( X \)). Let \( \mathcal{I} \) be a well-behaved set of interventions on \( X \), and assume that \( \text{supp}_H(X) \subseteq \text{supp}(X) \). Then, under Assumption 1, \( P_M \), \( M \) admits generalization to the interventions \( \mathcal{I} \). If one of the interventions is confounding-removing, then the causal function is a minimax solution.

In the case of support-extending interventions, further assumptions are required to ensure distribution generalization.

4.3.2 Support-Extending Interventions

If the interventions in \( \mathcal{I} \) extend the support of \( X \), i.e., \( \text{supp}_H(X) \subseteq \text{supp}(X) \), Assumption 1 is not sufficient for ensuring distribution generalization. This is because there may exist a model \( M \in \mathcal{M} \) which agrees with \( M \) on the observational distribution, but whose corresponding causal function \( f \) differs from \( f \) outside of the support of \( X \). In that case, a support-extending intervention on \( X \) may result in different dependencies between \( X \) and \( Y \) in the two models, and therefore potentially induce a different set of minimax solutions. The following assumption on the model class \( \mathcal{F} \) ensures that any \( f \in \mathcal{F} \) is uniquely determined by its values on \( \text{supp}(X) \).
Assumption 2 (Extrapolation of $\mathcal{F}$). For all $\hat{f}, \tilde{f} \in \mathcal{F}$ with $\hat{f}(x) = \tilde{f}(x)$ for all $x \in \text{supp}(X)$, it holds that $f = \tilde{f}$.

We will see that this assumption is sufficient (Proposition 4.4) for generalization to well-behaved interventions on $X$. Furthermore, it is also necessary (Proposition 4.7) if $\mathcal{F}$ is sufficiently flexible. The following proposition can be seen as an extension of Proposition 4.3.

Proposition 4.4 (Generalization to support-extending interventions on $X$). Let $\mathcal{I}$ be a well-behaved set of interventions on $X$. Then, under Assumptions 1 and 2, $(\mathbb{P}_{\tilde{M}}, \tilde{\mathcal{M}})$ admits generalization to $\mathcal{I}$. If one of the interventions is confounding-removing, then the causal function is a minimax solution.

Because the interventions may change the marginal distribution of $X$, the preceding proposition includes examples, in which distribution generalization is possible even if some of the considered joint (test) distributions are arbitrarily far from the training distribution, in terms of any reasonable divergence measure over distributions, such as Wasserstein distance or $f$-divergence.

Proposition 4.4 relies on Assumption 2. Even though this assumption is restrictive, it is satisfied by several reasonable function classes, which therefore allow for generalization to any set of well-behaved interventions. Below, we give two examples of such function classes.

4.3.2.1 Sufficient conditions for generalization: Assumption 2 states that every function in $\mathcal{F}$ is globally identified by its values on supp($X$). This is, for example, satisfied if $\mathcal{F}$ is a linear space of functions with domain $\mathcal{D} \subseteq \mathbb{R}^d$ which are linearly independent on supp($X$). More precisely, $\mathcal{F}$ is linearly closed, i.e.,

$$f_1, f_2 \in \mathcal{F}, c \in \mathbb{R} \Rightarrow f_1 + f_2 \in \mathcal{F}, cf_1 \in \mathcal{F},$$

and $\mathcal{F}$ is linearly independent on supp($X$), i.e.,

$$f_1(x) = 0 \ \forall x \in \text{supp}(X) \Rightarrow f_1(x) = 0 \ \forall x \in \mathcal{D}.$$ (5)

Examples of such classes include (i) globally linear parametric function classes, i.e., $\mathcal{F}$ is of the form

$$\mathcal{F}^\perp := \{f_0: \mathcal{D} \rightarrow \mathbb{R} | \exists \nu \in \mathbb{R}^k \text{ s.t. } \forall x \in \mathcal{D} : f_0(x) = \nu^T v(x)\},$$

where $v = (v_1, \ldots, v_k)$ consists of real-valued, linearly independent functions satisfying that $\mathbb{E}_M[v(x)]v(x)^T$ is strictly positive definite, and (ii) the class of differentiable functions that extend linearly outside of supp($X$), that is, $\mathcal{F}$ is of the form

$$\mathcal{F}^\perp := \{f_0: \mathcal{D} \rightarrow \mathbb{R} | f_0(x) = f_0(x_0) + \nabla f_0(x_0)(x-x_0) \}$$

where $x_0 := \arg\min_{x \in \text{supp}(X)} ||x-z||$ and supp($X$) is assumed to be closed with non-empty interior. Clearly, both of the above function classes are linearly closed. To see that $\mathcal{F}^\perp$ satisfies (5), let $x \in \mathbb{R}^k$ be s.t. $y^T v(x) = 0$ for all $x \in \text{supp}(X)$. Then, it follows that $0 = \mathbb{E}_M[(y^T v(x))^2] = y^T \mathbb{E}_M[v(x)v(x)^T] y$ and hence that $y = 0$. To see that $\mathcal{F}^\perp$ satisfies (5), let $f_0 \in \mathcal{F}^\perp$ and assume that $f_0(x) = 0$ for all $x \in \text{supp}(X)$. Then, $f_0(x) = 0$ for all $x \in \mathcal{D}$ and thus $\mathcal{F}^\perp$ uniquely defines the function on the entire domain $\mathcal{D}$.

By Proposition 4.4, generalization with respect to these model classes is possible for any well-behaved set of interventions. In practice, it may often be more realistic to impose bounds on the higher order derivatives of the functions in $\mathcal{F}$. We now prove that this still allows for what we will call approximate distribution generalization, see Propositions 4.5 and 4.6.

4.3.2.2 Sufficient conditions for approximate generalization: For differentiable functions, exact generalization cannot always be achieved. Bounding the first derivative, however, allows us to achieve approximate generalization.

We therefore consider the following function class

$$\mathcal{F}^2 := \{f_0: \mathcal{D} \rightarrow \mathbb{R} | f_0 \in C^1 \text{ with } \|\nabla f_0\|_\infty \leq K\}$$ (6)

for some fixed $K < \infty$, where $\nabla f_0$ denotes the gradient and $\mathcal{D} \subseteq \mathbb{R}^d$. We then have the following result.

Proposition 4.5 (Approx. generalization with bdd. derivatives (confounding-removing)). Let $\mathcal{F}$ be as defined in (6). Let $\mathcal{I}$ be a set of interventions on $X$ containing at least one confounding-removing intervention, and assume that Assumption 1 holds true. (In this case, the causal function $f$ is a minimax solution.) Then, for all $f^*$ with $f^* = f$ on supp($X$) and all $\tilde{M} \in \tilde{\mathcal{M}}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_{\tilde{M}}$, it holds that

$$\left| \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)} [(Y - f^*(X))^2] - \inf_{f \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)} [(Y - f(X))^2] \right| \leq 4\delta^2 K^2 + 4\delta K \sqrt{\text{Var}_M(Y)},$$

where $\delta := \sup_{x \in \text{supp}(X)} \inf_{z \in \text{supp}(X)} ||x-z||$. If $\mathcal{I}$ consists only of confounding-removing interventions, the same statement holds when replacing the bound by $4\delta K^2$.

Proposition 4.5 states that the deviation of the worst-case generalization error from the best possible value is bounded by a term that grows with the square of $\delta$. Intuitively, this means that under the function class defined in (6), approximate generalization is reasonable only for interventions that are close to the support of $X$. We now prove a similar result for cases in which the minimax solution is not necessarily the causal function. The following proposition bounds the worst-case generalization error for arbitrary confounding-preserving interventions. Here, the bound additionally accounts for the approximation to the minimax solution.

Proposition 4.6 (Approx. generalization with bdd. derivatives (confounding-preserving)). Let $\mathcal{F}$ be as defined in (6). Let $\mathcal{I}$ be a set of confounding-preserving interventions on $X$, and assume that Assumption 1 is satisfied. Let $\varepsilon > 0$ and let $f^*$ be $\mathcal{F}$ be such that

$$\left| \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)} [(Y - f^*(X))^2] - \inf_{f \in \mathcal{F}^\perp} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)} [(Y - f(X))^2] \right| \leq \varepsilon.$$ (7)

Then, for all $\tilde{M} \in \tilde{\mathcal{M}}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_{\tilde{M}}$, it holds that

$$\left| \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)} [(Y - f^*(X))^2] - \inf_{f \in \mathcal{F}^\perp} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)} [(Y - f(X))^2] \right| \leq \varepsilon + 12\delta^2 K^2 + 32\delta K \sqrt{\text{Var}_M(Y)} + 4/32K \sqrt{\varepsilon},$$

where $\delta := \sup_{x \in \text{supp}(X)} \inf_{z \in \text{supp}(X)} ||x-z||$. We can take $f^*$ to be the minimax solution if it exists. In that case, the terms involving $\varepsilon$ disappear from the bound, which then becomes more similar to the one in Proposition 4.5.
4.3.2.3 Impossibility of generalization without constraints on $\mathcal{F}$: If we do not constrain the function class $\mathcal{F}$, generalization is impossible. Even if we consider the set of all continuous functions $\mathcal{F}$, we cannot generalize to interventions outside the support of $X$. This statement holds even if Assumption 1 is satisfied.

Proposition 4.7 (Impossibility of extrapolation). Assume that $\mathcal{F} = \{f_x : \mathbb{R}^d \rightarrow \mathbb{R} \mid f_x \text{ is continuous}\}$. Let $I$ be a well-behaved set of support-extending interventions on $X$, such that $\text{supp}_I(X) \setminus \text{supp}(X)$ has non-empty interior. Then, $(\mathcal{P}_M, \mathcal{M})$ does not admit generalization to $I$, even if Assumption 1 is satisfied. In particular, for any function $f \in \mathcal{F}$ and any $c > 0$, there exists a model $M \in \mathcal{M}$, with $\pi_M = \pi_M$, such that

$$\left| \sup_{i \in I} \mathbb{E}_{M(i)} [(Y - \bar{f}(X))^2] - \inf_{f \in \mathcal{F}} \sup_{i \in I} \mathbb{E}_{M(i)} [(Y - f_i(X))^2] \right| \geq c.$$ 

The above impossibility result is visualized in Fig. 1.

4.4 Generalization to Interventions on $A$

We will see that, for interventions on $A$, parts of the analysis simplify. Since $A$ influences the system only via the covariates $X$, any such intervention may, in terms of its effect on $(X, Y)$, be equivalently expressed as an intervention on $X$ in which the structural assignment of $X$ is altered in a way that depends on the functional relationship $g$ between $X$ and $A$. We can therefore employ several of the results from Section 4.3 by imposing an additional assumption on the identifiability of $g$.

Assumption 3 (Identifiability of $g$). For all $M = (\bar{f}, \bar{g}, \bar{h}_1, \bar{h}_2, \bar{Q}) \in \mathcal{M}$ with $\pi_{\bar{M}} = \pi_{\bar{M}}$, it holds that $\bar{g}(a) = g(a)$ for all $a \in \text{supp}(A) \cup \text{supp}_I(A)$.

Since $g(A)$ is a conditional mean for $X$ given $A$, the values of $g$ are identified from $\mathcal{P}_M$ for $\mathcal{P}_M$-almost all $a$. If $\text{supp}_I(A) \subseteq \text{supp}(A)$, Assumption 3 therefore holds if, for example, $\mathcal{G}$ contains continuous functions only. The pointwise identifiability of $g$ is necessary, for example, if some of the test distributions are induced by hard interventions on $A$, which set $A$ to some fixed value $a \in \mathbb{R}^p$. In the case where the interventions $I$ extend the support of $A$, we additionally require the function class $\mathcal{G}$ to extrapolate from $\text{supp}(A)$ to $\text{supp}(A) \cup \text{supp}_I(A)$; this is similar to the conditions on $\mathcal{F}$ which we made in Section 4.3.2 and requires further restrictions on $\mathcal{G}$. Under Assumption 3, we obtain a result corresponding to Propositions 4.3 and 4.4.

Proposition 4.8 (Generalization to interventions on $A$). Let $\mathcal{I}$ be a set of interventions on $A$ and assume Assumption 3 is satisfied. Then, $(\mathcal{P}_M, \mathcal{M})$ admits generalization to $\mathcal{I}$ if either $\text{supp}_I(A) \subseteq \text{supp}(A)$ and Assumption 1 is satisfied or if both Assumptions 1 and 2 are satisfied.

As becomes clear from the proof of this proposition, in general, the causal function does not need to be a minimax solution. Further, Assumption 1 is not necessary for generalization. In the case where $\mathcal{F}, \mathcal{G}, \mathcal{H}_1, \mathcal{H}_2$ consist of linear functions, anchor regression [54] and K-class estimators [32] consider certain sets of interventions on $A$ which render minimax solutions identifiable (and estimate them consistently) even if Assumption 1 does not hold. Similarly, if for a categorical $A$, we have $\text{supp}_I(A) \subseteq \text{supp}(A)$, it is possible to drop Assumption 1.

4.4.1 Impossibility of Generalization Without Constraining $\mathcal{G}$

Without restrictions on the model class $\mathcal{G}$, generalization to interventions on $A$ is impossible. This holds true even under strong assumptions on the true causal function (such as $f$ is known to be linear). Below, we give a formal impossibility result for hard interventions on $A$, which set $A$ to some fixed value, and where $\mathcal{G}$ is the set of all continuous functions.

Proposition 4.9 (Impossibility of generalization to interventions on $A$). Assume that $\mathcal{F} = \{f_x : \mathbb{R}^d \rightarrow \mathbb{R} \mid f_x \text{ is linear}\}$ and $\mathcal{G} = \{g_x : \mathbb{R}^r \rightarrow \mathbb{R}^d \mid g_x \text{ is continuous}\}$.

Let $A \subseteq \mathbb{R}^s$ be bounded, and let $I$ denote the set of all hard interventions which set $A$ to some fixed value from $A$. Assume that $A \setminus \text{supp}(A)$ has nonempty interior. Assume further that $\mathbb{E}_M[\xi X \mathcal{G}] \neq 0$ (this excludes the case of no hidden confounding). Then, $(\mathcal{P}_M, \mathcal{M})$ does not admit generalization to $I$. In addition, any function other than $f$ may perform arbitrarily bad under the interventions in $I$. That is, for any $f \neq f$ and $c > 0$, there exists a model $M \in \mathcal{M}$ with $\pi_M = \pi_M$ such that

$$\left| \sup_{i \in I} \mathbb{E}_{M(i)} [(Y - \bar{f}(X))^2] - \inf_{f \in \mathcal{F}} \sup_{i \in I} \mathbb{E}_{M(i)} [(Y - f_i(X))^2] \right| \geq c.$$ 

This proposition is part of the argument showing that anchor regression [54] can be extended to nonlinear settings only under strong assumptions; the setting of a linear class $\mathcal{G}$ and a potentially nonlinear class $\mathcal{F}$ is covered in Section 4.3.2, by rewriting interventions on $A$ as interventions on $X$.

An impossibility result similar to the proposition above can be shown if $A$ is categorical. As long as not all categories have been observed during training it is possible that the intervention which sets $A$ to a previously unseen category can result in a support-extending distribution shift on $X$. Using Proposition 4.7, it therefore follows that generalization can become impossible. Since a categorical $A$ can encode
settings of multi-task learning and domain generalization (see Section 2.4), this result then complements well-known impossibility results for these problems, even under the covariate shift assumption, e.g., [18].

5 LEARNING GENERALIZING MODELS FROM DATA

So far, our focus has been on the possibility to generalize, that is, we have investigated under which conditions it is possible to identify generalizing models from the observational distribution. In practice, generalizing models need to be estimated from finitely many data. This task is challenging for several reasons. First, analytical solutions to the minimax problem (2) are only known in few cases. Even if generalization is possible, the inferential target thus often remains a complicated object, given as a well-defined but unknown function of the observational distribution. Second, we have seen that the ability to generalize depends strongly on whether the interventions extend the support of \( X \), see Propositions 4.4 and 4.7. In a setting with a finite amount of data, the empirical support of the data lies within some bounded region, and suitable constraints on the function class \( \mathcal{F} \) are necessary when aiming to achieve empirical generalization outside this region, even if \( X \) comes from a distribution with full support. As we show in our simulations in Section 5.2.4 (see figures), constraining the function class can also improve the prediction performance at the boundary of the support.

In Section 5.1, we survey existing methods for learning generalizing models. Often, these methods assume either a globally linear model class \( \mathcal{F} \) or are completely non-parametric and therefore do not generalize outside the empirical support of the data. Motivated by this observation, we introduce in Section 5.2 a novel estimator, which exploits an instrumental variable setup and a particular extrapolation assumption to learn a globally generalizing model.

5.1 Existing Methods

As discussed in Section 1, a wide range of methods have been proposed to guard against various types of distributional changes. Here, we review methods that fit into the causal framework in the sense that the distributions that in the minimax formulation the supremum is taken over are induced by interventions.

For well-behaved interventions on \( X \) which contain at least one confounding-removing intervention, estimating minimax solutions reduces to the well-studied problem of estimating causal relationships. One class of algorithms for this task is given by linear instrumental variable (IV) approaches. They assume that \( \mathcal{F} \) is linear and require identifiability of the causal function (Assumption 1) via a rank condition on the observational distribution, see Appendix B, available in the online supplemental material. Their target of inference is to estimate the causal function, which by Proposition 3.1 will coincide with the minimax solution if the set \( \mathcal{F} \) consists of well-behaved interventions with at least one of them being confounding-removing. A basic estimator for linear IV models is the two-stage least squares (TSLS) estimator, which minimizes the norm of the prediction residuals projected onto the subspace spanned by the observed instruments (TSLS objective). TSLS estimators are consistent but do not come with strong finite sample guarantees; e.g., they do not have finite moments in a just-identified setup, e.g., [38]. K-class estimators [64] have been proposed to overcome some of these issues. They minimize a linear combination of the residual sum of squares (OLS objective) and the TSLS objective. K-class estimators can be seen as utilizing a bias-variance trade-off. For fixed and non-trivial relative weights, they have, in a Gaussian setting, finite moments up to a certain order that depends on the sample-size and the number of predictors used. If the weights are such that the OLS objective is ignored asymptotically, they consistently estimate the causal parameter, e.g., [38]. More recently, PULSE has been proposed [32], a data-driven procedure for choosing the relative weights such that the prediction residuals ‘just’ pass a test for simultaneous uncorrelatedness with the instruments.

In cases where the minimax solution does not coincide with the causal function, only few algorithms exist. Anchor regression [54] is a procedure that can be used when \( \mathcal{F} \) and \( \mathcal{G} \) are linear and \( h_1 \) is additive in the noise component. It finds the minimax solution if the set \( \mathcal{I} \) consists of all interventions on \( A \) up to a fixed intervention strength, and is applicable even if Assumption 1 is not necessarily satisfied.

In a linear setting, where the regression coefficients differ between different environments, it is also possible to minimize the worst-case risk among the observed environments [40]. In its current formulation, this approach does not quite fit into the above framework, as it does not allow for changing distributions of the covariates. A summary of the mentioned methods and their assumptions is given in Table 2.

If \( \mathcal{F} \) is a nonlinear or non-parametric class of functions, the task of finding minimax solutions becomes more

### Table 2

| Model Class | Interventions | Support of \( X \) | Assumptions | Algorithm |
|-------------|---------------|--------------------|-------------|-----------|
| \( \mathcal{F} \) linear | on \( X \) or \( A \) of which at least one is confounding-removing | – | Assumption 1 | linear IV (e.g., two-stage least squares, K-class or PULSE [32, 64]) |
| \( \mathcal{F}, \mathcal{G} \) linear | on \( A \) | bounded strength | – | anchor regression [54], see also [64] |
| \( \mathcal{F} \) smooth | on \( X \) or \( A \) of which at least one is confounding-removing | support-reducing | Assumption 1 | nonlinear IV (e.g., NPREGIV [52], Deep IV [27], Sieve IV [14], [43], Kernel IV [60]) NILE Section 5.2 |
| \( \mathcal{F} \) smooth and linearly extrapolated | on \( X \) or \( A \) of which at least one is confounding-removing | – | Assumption 1 | }
difficult. In cases where the causal function is among such solutions, this problem has been studied in the econometrics community. For example, [42], [43] treat the identifiability and estimation of causal functions in non-parametric function classes. Several non-parametric IV procedures exist, e.g., NPREGIV [52] contains modified implementations of [29] and [16], which we will refer to as NPREGIV-1 and NPREGIV-2, respectively. Other procedures include Deep IV [27], Sieve IV [14], [43] and Kernel IV [60]. Identifiability and estimation of the causal function using nonlinear IV methods in parametric function classes is discussed in Appendix B, available in the online supplemental material. Unlike in the linear case, most of the methods do not aim to extrapolate and only recover the causal function inside the support of $X$, that is, they cannot be used to predict interventions outside of this domain. In the following section, we propose a procedure that is able to extrapolate when $F$ consists of functions which extend linearly outside of the support of $X$. In principle, any other extrapolation rule may be employed here, as long as all functions from $F$ are uniquely determined by their values on the support of $X$, that is, Assumption 2 is satisfied.

In our simulations, we see that our method can improve the prediction performance on the boundary of the support and outperforms other methods when comparing the estimation on the support.

### 5.2 NILE

We have seen in Proposition 4.7 that in order to generalize to interventions which extend the support of $X$, we require additional assumptions on the function class $F$. In this section, we start from such assumptions and verify both theoretically and practically that they allow us to perform distribution generalization in the considered setup. Along the way, several choices can be made and usually several options are possible. We will see that our choices yield a method with competitive performance, but we do not claim optimality of our procedure. Several of our choices were partially made to keep the theoretical exposition simple and the method computationally efficient. We first consider the univariate case (i.e., $X$ and $A$ are real-valued) and comment later on the possibility to extend the methodology to higher dimensions. Unless specific background knowledge is given, it might be reasonable to assume that the causal function extends linearly outside a fixed interval $[a, b]$. By additionally imposing differentiability on $F$, any function from $F$ is uniquely defined by its values within $[a, b]$, see also Section 4.3.2. Given an estimate $f$ on $[a, b]$, the linear extrapolation property then yields a global estimate on the whole of $\mathbb{R}$. In principle, any class of differentiable functions can be used. Here, we assume that, on the interval $[a, b]$, the causal function $f$ is contained in the linear span of a B-spline basis. More formally, let $B = (B_1, \ldots, B_k)$ be a fixed B-spline basis on $[a, b]$, and define $g := (a, b, B)$. Our procedure assumes that the true causal function $f$ belongs to the function class $F_g := \{ f_g(\cdot; \theta) : \theta \in \mathbb{R}^k \}$, where for every $x \in \mathbb{R}$ and $\theta \in \mathbb{R}^k$, $f_g(x; \theta)$ is given as

$$f_g(x; \theta) := \begin{cases} B(a)^\top \theta + B'(a)^\top (x-a) & \text{if } x < a \\ B(x)^\top \theta & \text{if } x \in [a, b] \\ B(b)^\top \theta + B'(b)^\top (x-b) & \text{if } x > b, \end{cases}$$

where $B' := (B'_1, \ldots, B'_k)$ denotes the component-wise derivative of $B$. In our algorithm, $\eta = (a, b, B)$ is a hyper-parameter, which can be set manually, or be chosen from data.

#### 5.2.1 Estimation Procedure

We now introduce our estimation procedure for fixed choices of all hyper-parameters. Section 5.2.2 describes how these can be chosen from data in practice. Let $(X, Y, A) \in \mathbb{R}^{n \times d}$ be $n$ i.i.d. realizations sampled from a distribution over $(X, Y, A)$, let $\eta = (a, b, B)$ be fixed and assume that $\text{supp}(X) \subseteq [a, b]$. Our algorithm aims to learn the causal function $f_g(\cdot; \theta^0) \in F_g$, which is determined by the linear causal parameter $\theta^0$ of a $k$-dimensional vector of covariates $(B_1(X), \ldots, B_k(X))$. From standard linear IV theory, it is known that at least $k$ instrumental variables are required to identify the $k$ causal parameters, see Appendix B, available in the online supplemental material. We therefore artificially generate such instruments by nonlinearly transforming $A$, by using another B-spline basis $C = (C_1, \ldots, C_k)$. The parameter $\theta^0$ can then be identified from the observational distribution under appropriate rank conditions, see Section 5.2.3. In that case, the hypothesis $H_0(\theta) : \theta = \theta^0$ is equivalent to the hypothesis $H_0(\theta^0) : \mathbb{E}[C(A)(Y - B(X)^\top \theta^0)] = 0$. Let $B \in \mathbb{R}^{n \times k}$ and $C \in \mathbb{R}^{n \times k}$ be the associated design matrices, for each $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, k\}$ given as $B_{ij} = B_j(X_i)$ and $C_{ij} = C_j(A_i)$. A straightforward choice would be to construct the standard TSLS estimator, i.e., $\hat{\theta} = \arg \min_{\theta \in \mathbb{R}^k} \| P(Y - B \theta) \|_2^2$, where $P$ is the projection matrix onto the columns of $C$, see also [26]. Even though this procedure may result in an asymptotically consistent estimator, there are several reasons why it may be suboptimal in a finite sample setting. First, the above estimator can have large finite sample bias, in particular if $k$ is large. Indeed, in the extreme case where $k = n$, and assuming that all columns in $C$ are linearly independent, $P$ is equal to the identity matrix, and $\hat{\theta}$ coincides with the OLS estimator. Second, since $\theta$ corresponds to the linear parameter of a spline basis, it seems reasonable to impose constraints on $\theta$ which enforce smoothness of the resulting spline function. Both of these points can be addressed by introducing additional penalties into the estimation procedure. Let therefore $K \in \mathbb{R}^{k \times k}$ and $M \in \mathbb{R}^{k \times k}$ be the matrices that are, for each $i, j \in \{1, \ldots, k\}$, defined as $K_{ij} := \int B_i'(x) B_j'(x) dx$ and $M_{ij} := \int C_i'(\gamma) C_j'(\gamma) d\gamma$, and let $\gamma, \delta > 0$ be the respective penalties associated with $K$ and $M$. For $\lambda \geq 0$ and with $\mu := (\gamma, \delta, C)$, we then define the estimator

$$\hat{\theta^0}_{\lambda, \mu} := \arg \min_{\theta \in \mathbb{R}^k} \| Y - B \theta \|_2^2 + \lambda \| P(Y - B \theta) \|_2^2 + \gamma \theta^\top K \theta,$$

where $P := C(C^\top C + \delta M)^{-1} C^\top$ is the ‘hat’-matrix for a penalized regression onto the columns of $C$. By choice of $K$, the term $\theta^\top K \theta$ is equal to the integrated squared curvature of the spline function parameterized by $\theta$. The regularization induced by the second summand in (8) is similar to the one from K-class estimators in linear settings [64]. The function class (7) enforces linear extrapolation. In principle, the above approach extends to situations where $X$ and $A$ are higher-dimensional, in which case $B$ and $C$ consist of multivariable functions. For example, [22] propose the use of tensor product splines, and introduce multivariate smoothness.
penalties based on pairwise first- or second order parameter differences of basis functions which are close-by with respect to some suitably chosen metric. Similarly to (8), such penalties result in a convex optimization problem. However, due to the large number of involved variables, the optimization procedure becomes computationally burdensome already in small dimensions.

Within the function class \( \mathcal{F}_q \), the above defines the global estimate \( \hat{f}_q(x; \hat{\theta}_{L,q,M}) \), for every \( x \in \mathbb{R} \), given by

\[
\begin{align*}
B(a) \hat{\theta}^q_{L,q,M} + B'(a) \hat{\theta}^q_{L,q,M}(x-a) & \quad \text{if } x < a \\
B(x) \hat{\theta}^q_{L,q,M} & \quad \text{if } x \in [a, b] \\
B(b) \hat{\theta}^q_{L,q,M} + B'(b) \hat{\theta}^q_{L,q,M}(x-b) & \quad \text{if } x > b.
\end{align*}
\]

We deliberately distinguish between three different groups of hyper-parameters \( \eta, \mu \) and \( \lambda \). The parameter \( \eta = (a, b, \mathcal{B}) \) defines the function class to which the causal function \( f \) is assumed to belong. To prove consistency of our estimator, we require this function class to be correctly specified. In turn, the parameters \( \lambda \) and \( \mu = (\gamma, \delta, C) \) are algorithmic parameters that do not describe the statistical model. Their values only affects the finite sample behavior of our algorithm, whereas consistency is ensured as long as \( C \) satisfies certain rank conditions, see Assumption (B2) in Section 5.2.3.

In practice, \( \gamma \) and \( \delta \) are chosen via a cross-validation procedure, see Section 5.2.2. The parameter \( \lambda \) determines the relative contribution of the OLS and TSLS losses to the objective function. To choose \( \lambda \) from data, we use an idea similar to the PULSE [32].

### 5.2.2 Algorithm

Let for now \( \eta, \mu \) be fixed. In the limit \( \lambda \rightarrow \infty \), our estimation procedure becomes equivalent to minimizing the TSLS loss \( \theta \rightarrow \| \mathbb{P}_h(Y - B \theta) \|_2^2 \), which may be interpreted as searching for the parameter \( \theta \) which complies ‘best’ with the hypothesis \( H_0(\theta) : \mathbb{E}[L(A)(Y - B \theta)^T] = 0 \). For finitely many data, following the idea introduced in [32], we propose to choose the value for \( \lambda \) such that \( H_0(\hat{\theta}_{L,q,M}^q) \) is just accepted (e.g., at a significance level \( \alpha = 0.05 \)). That is, among all \( \lambda \geq 0 \) which result in an estimator that is not rejected as a candidate for the causal parameter, we chose the one which yields maximal contribution of the OLS loss to the objective function. More formally, let for every \( \theta \in \mathbb{R}^k, T(\theta) = (T_n(\theta))_{n \in \mathbb{N}} \) be a statistical test at (asymptotic) level \( \alpha \) for \( H_0(\theta) \) with rejection threshold \( q(\alpha) \). That is, \( T_n(\theta) \) does not reject \( H_0(\theta) \) if and only if \( T_n(\theta) \leq q(\alpha) \). The penalty \( \lambda^*_n \) is then chosen in the following data-driven way

\[
\lambda^*_n := \inf \{ \lambda \geq 0 : T_n(\hat{\theta}^q_{L,q,M}) \leq q(\alpha) \}.
\]

In general, \( \lambda^*_n \) is not guaranteed to be finite for an arbitrary test statistic \( T_n \). Even for a reasonable test statistic it might happen that \( T_n(\hat{\theta}^q_{L,q,M}) > q(\alpha) \) for all \( \lambda \geq 0 \); see [32] for further details. We can remedy the problem by reverting to another well-defined and consistent estimator, such as the TSLS (which minimizes the TSLS loss above) if \( \lambda^*_n \) is not finite. Furthermore, if \( \lambda \rightarrow T_n(\hat{\theta}^q_{L,q,M}) \) is monotonic, \( \lambda^*_n \) can be computed efficiently by a binary search procedure. In our algorithm, the test statistic \( T \) and rejection threshold \( q \) can be supplied by the user. Conditions on \( T \) that are sufficient to yield a consistent estimator \( \hat{f}_q(\cdot, \hat{\theta}_{L,q,M}) \), given that \( \mathcal{F}_q \) is correctly specified, are presented in Section 5.2.3. Two packages of test statistics which are implemented in our code package can be found in Appendix C, available in the online supplemental material.

For every \( \gamma \geq 0 \), let \( Q_\gamma = B(\mathbb{B}^T + \gamma \mathbb{K})^{-1}\mathbb{B}^T \) be the ‘hat’-matrix for the penalized regression onto \( B \). Our method then proceeds as shown in Algorithm 1.

#### Algorithm 1. NILE (“Nonlinear Intervention-Robust Linear Extrapolator”)

1. input: data \((X, Y, A) \in \mathbb{R}^{n \times 3}\);
2. options: \(k, T, q, \alpha\);
3. Begin
4. \( a \leftarrow \min(X_i), b \leftarrow \max(X_i)\);
5. construct cubic B-spline bases \( B = (B_1, \ldots, B_k) \)
6. and \( C = (C_1, \ldots, C_k) \) at equidistant knots, with boundary knots at respective extreme values of \( X \) and \( A \);
7. define \( \hat{\eta} \leftarrow (a, b, B)\);
8. choose \( \delta_{CV}^q > 0 \) by 10-fold CV to minimize the out-of-sample mean squared error of \( Y = \mathbb{P}_h Y \);
9. choose \( \gamma_{CV}^q > 0 \) by 10-fold CV to minimize the out-of-sample mean squared error of \( Y = Q_\gamma Y \);
10. define \( \mu_{CV}^q \leftarrow (\delta_{CV}^q, \gamma_{CV}^q, C)\);
11. approx. \( \lambda^*_n \leftarrow \inf \{ \lambda \geq 0 : T_n(\hat{\theta}^q_{L,q,M}) \leq q(\alpha) \} \) by binary search;
12. update \( q_{CV}^q \leftarrow (1 + \lambda^*_n) q_{CV}^q \);
13. compute \( \theta_{L,q,M}^q \) using (8);
14. end
15. output: \( \hat{f}_q : \lambda^*_n \cdot \theta_{L,q,M}^q \) defined by (9);

The penalty parameter \( \gamma_{CV}^q \) is chosen to minimize the out-of-sample mean squared error of the prediction model \( \hat{Y} = Q_\gamma Y \), which corresponds to the solution of (8) for \( \lambda = 0 \). After choosing \( \lambda^*_n \), the objective function in (8) increases by the term \( \lambda^*_n \| \mathbb{P}_h (Y - B \theta) \|_2^2 \). In order for the penalty term \( \gamma \theta^T K \theta \) to impose the same degree of smoothness in the altered optimization problem, the penalty parameter \( \gamma \) needs to be adjusted accordingly. The heuristic update in our algorithm is motivated by the simple observation that for all \( \delta, \gamma \geq 0 \),

\[
\| Y - B \theta \|_2^2 + \lambda \| \mathbb{P}_h (Y - B \theta) \|_2^2 \leq (1 + \lambda) \| Y - B \theta \|_2^2.
\]

#### 5.2.3 Asymptotic Generalization (Consistency)

We now prove consistency of our estimator in the case where the hyper-parameters \( (\eta, \mu) \) are fixed (rather than data-driven), and the function class \( \mathcal{F}_q \) is correctly specified. Fix any \( a < b \) and a basis \( B = (B_1, \ldots, B_k) \). Let \( \mathcal{M} = \mathbb{R}^k \times \mathbb{R}^k \times \mathbb{R}^2 \times \mathbb{R}^2 \), where \( \mathcal{F}_0 \) is as described in Section 5.2. Assume that the data-generating model \( M = (f_0(\cdot; \theta^0), g, b_1, b_2, Q) \in \mathcal{M} \) induces an observational distribution \( \mathbb{P}_0 \) such that \( \mathbb{P}^0(X) \subseteq (a, b) \). Let further \( T \) be a set of interventions on \( X \) or \( A \), and let \( \alpha \in (0, 1) \) be a fixed significance level.

We prove asymptotic generalization (consistency) for an idealized version of the NILE estimator which utilizes \( \eta_{\alpha} \), rather than the data-driven values. Choose any \( \delta, \gamma \geq 0 \) and basis \( C = (C_1, \ldots, C_k) \) and let \( \mu = (\delta, \gamma, C) \). We will make use of the following assumptions.
For all $M \in \mathcal{M}$ with $P_M = P_{\tilde{M}}$ it holds that $\sup_{i \in I} \mathbb{E}_{X(i)}[X^2], \sup_{i \in I} \lambda_{\max}(\mathbb{E}_{X(i)}[B(X)B(X)^\top]) < \infty$.

The matrices $\mathbb{E}_M[B(X)B(X)^\top], \mathbb{E}_M[C(A)C(A)^\top]$ and $\mathbb{E}_M[C(A)B(X)^\top]$ are of full rank.

$T(\theta)$ has uniform asymptotic power on any compact set of alternatives.

$\lambda_n := \inf \{\lambda \geq 0 : T_n(\tilde{\theta}_{R_0,\mu}) \leq \tilde{q}(\alpha)\}$ is almost surely finite.

$\lambda \mapsto T_n(\tilde{\theta}_{R_0,\mu})$ is weakly decreasing and $\theta \mapsto T_n(\theta)$ is continuous.

Assumptions (B1)–(B2) ensure consistency of the estimator as long as $\lambda_n$ tends to infinity. Intuitively, in this case, we can apply arguments similar to those that prove consistency of the TSLS estimator. Assumptions (C1)–(C3) ensure that consistency is achieved when choosing $\lambda_n$ in the data-driven fashion described in Section 5.2.2. Assumption (B1), $\lambda_{\max}$ denotes the largest eigenvalue. In words, the assumption states that, under each model $M \in \mathcal{M}$ with $P_M = P_{\tilde{M}}$, there exists a finite upper bound on the variance of any linear combination of the basis functions $B(X)$, uniformly over all distributions induced by $I$. The first two rank conditions of (B2) enable certain limiting arguments to be valid and they guarantee that estimators are asymptotically well-defined. The last rank condition of (B2) is the so-called rank condition for identification. It guarantees that $\theta_0$ is identified from the observational distribution in the sense that the hypothesis $H_0(\theta) : \theta = \theta_0$ becomes equivalent with $H_0(\theta) : \mathbb{E}_{x,X}(C(A)(Y - B(X)^\top)\theta) = 0$. (C1) means that for any compact set $K \subseteq \mathbb{R}^d$ with $\theta \not\in K$ it holds that $\lim_{n \to \infty} P(\inf_{\theta \in K} T_n(\theta) \leq q(\alpha)) = 0$. If the considered test has, in addition, a level guarantee, such as pointwise asymptotic level, the interpretation of the finite sample estimator discussed in Section 5.2.2 remains valid (such level guarantee may potentially yield improved finite sample performance, too). (C2) is made to simplify the consistency proof. As previously discussed in Section 5.2.2, if (C2) is not satisfied, we can output another well-defined and consistent estimator on the event ($\lambda_n = \infty$), ensuring that consistency still holds.

Under these conditions, we have the following asymptotic generalization guarantee.

**Proposition 5.1 (Asymptotic generalization).** Let $I$ be a set of interventions on $X$ or $A$ of which at least one is confounding-removing. If assumptions (B1)–(B2) and (C1)–(C3) hold true, then, for any $M \in \mathcal{M}$ with $P_M = P_{\tilde{M}}$, and any $\epsilon > 0$, it holds that

$$P_M \left[ \sup_{i \in I} \mathbb{E}_{X(i)} \left( Y - f_0(X; \tilde{\theta}_{R_0,\mu})^2 \right) - \inf_{i \in I} \sup_{\theta \in \tilde{M}_{i}(i)} \mathbb{E}_{X(i)} \left( Y - f_0(X)^2 \right) \leq \epsilon \right] \to 1,$$

as $n \to \infty$. In the above event, only $\tilde{\theta}_{R_0,\mu}$ is stochastic.

**5.2.4 Experiments**

We now investigate the empirical performance of our proposed estimator, the NILE, with $k = 50$ spline basis functions. To choose $\lambda_n$, we use the test statistic $T_n^2$, which tests the slightly stronger hypothesis $H_0$, see Appendix C, available in the online supplemental material. In all experiments use the significance level $\alpha = 0.05$. We include two other approaches as baseline: (i) the method NPRGIV-1 (using its default options) introduced in Section 5.1, and (ii) a linearly extrapolating estimator of the ordinary regression of $Y$ on $X$ (which corresponds to the NILE with $\lambda \equiv 0$). In all experiments, we generate data sets of size $n = 200$ as independent replications from

$$A := \varepsilon_A, \quad H := \varepsilon_H, \quad X := \alpha_A A + \alpha_H H + \alpha_c \varepsilon_X, \quad Y := f(X) + 0.3H + 0.2\varepsilon_Y,$$

where $(\varepsilon_A, \varepsilon_H, \varepsilon_X, \varepsilon_Y)$ are jointly independent with Uniform $(-1, 1)$ marginals. To make results comparable across different parameter settings, we impose the constraint $\alpha_A^2 + \alpha_H^2 + \alpha_c^2 = 1$, which ensures that in all models, $X$ has variance 1/3. The function $f$ is drawn from the linear span of a basis of four natural cubic splines with knots placed equidistantly within the 90 percent inner quantile range of $X$. By well-known properties of natural splines, any such function extends linearly outside the boundary knots. Fig. 2 (left) shows an example data set from (10), where the causal function is indicated in green. We additionally display estimates obtained by each of the considered methods, based on 20 i.i.d. datasets. Due to the confounding variable $H$, the OLS estimator is clearly biased. NPRGIV-1 exploits $A$ as an instrumental variable and obtains good results within the support of the observed data. Due to its non-parametric nature, however, it cannot extrapolate outside this domain. The NILE estimator exploits the linear extrapolation assumption on $f$ to produce global estimates.

We further investigate the empirical worst-case risk across several different models of the form (10). That is, for a fixed set of parameters $(\alpha_A, \alpha_H, \alpha_c)$, we construct several models $M_1, \ldots, M_N$ of the form (10) by randomly sampling causal functions $f_1, \ldots, f_N$ (see Appendix D, available in the online supplemental material, for further details on the sampling procedure). For every $x \in [0, 2]$, let $I_x$ denote the set of hard interventions which set $X$ to some fixed value in $[-x, x]$. We then characterize the performance of each method using the average (across different models) worst-case risk (across the interventions in $I_x$), i.e., for each estimator $\hat{f}$, we consider

![Fig. 2. A sample dataset from the model (10) with $\alpha_A = \sqrt{1/3}, \alpha_H = \sqrt{2/3}, \alpha_c = 0$. The true causal function is indicated by a green dashed line. For each method, we show 20 estimates of this function, each based on an independent sample from (10). For values within the support of the training data (vertical dashed lines mark the inner 90 percent quantile range), NPRGIV-1 correctly estimates the causal function well. As expected, when moving outside the support of $X$, the estimates become unreliable, and we gain an increasing advantage by exploiting the linear extrapolation assumed by the NILE.](image-url)
in this experiment, the empirical average of \(X = 0\), the NILE 
\[ a_x = \text{a noise term for} \, N \] 
\( X \rightarrow Y \) for different confounding- and intervention strengths (see Appendix D for details on the sampling procedure), available in the online supplemental material. The performances under individual models are shown by thin lines; the average performance (11) across all models is indicated by thick lines. In all considered models, the optimal prediction error (green dashed line) is equal to \( E(\xi^2) \) (by consistency, for any fixed function \( f \), NILE’s worst-case risk converges point-wise to this value for increasing sample size). The grey area indicates the inner 90 percent quantile range of \( X \) in the training distribution; the white area can be seen as an area of generalization.

\[
\frac{1}{N} \sum_{i=1}^{N} \sup_{x \in I_x} E_{M[j]}[(Y - \hat{f}(X))^2] = E(\xi^2) + \frac{1}{N} \sum_{i=1}^{N} \sup_{x \in [-x,x]} (f(x) - \hat{f}(\bar{x}))^2, \tag{11}
\]

where \( \xi_x := 0.3H + 0.2\gamma \) is the noise term for \( Y \) (which is fixed across all experiments). In practice, we evaluate the functions \( \tilde{f}, f_1, \ldots, f_N \) on a fine grid on \([-x, x]\) to approximate the above supremum. Fig. 3 plots the average worst-case risk versus intervention strength for varying degree of confounding (\( \alpha_x \)). The optimal worst-case risk \( E(\xi^2) \) is indicated by a green dashed line. The results show that the linear extrapolation property of the NILE estimator is beneficial in particular for strong interventions. In the case of no confounding (\( \alpha_H = 0 \)), the minimax solution coincides with the regression of \( Y \) on \( X \), hence even the OLS estimator yields good predictive performance. In this case, the hypothesis \( H_0(\lambda^{(0)}_{\gamma} \mid \{\theta^0_{x} \}_{x \in [-x,x]} \}) \) is accepted already for small values of \( \lambda \) (in this experiment, the empirical average of \( \lambda^{(0)}_n \) equals 0.015), and the NILE estimator becomes indistinguishable from the OLS. As the confounding strength increases, the OLS becomes increasingly biased, and the NILE objective function differs more notably from the OLS (average \( \lambda^{(0)}_n \) of 2.412 and 5.136, respectively). The method NPREGIV-1 slightly outperforms the NILE inside the support of the observed data, but drops in performance for stronger interventions. We believe that the increase in extrapolation performance of the NILE for stronger confounding (increasing \( \alpha_H \)) might stem from the fact that, as the \( \lambda_n \) increases, also the smoothness penalty \( \gamma \) increases, see Algorithm 1. While this results in slightly worse in-sample prediction, it seems beneficial for extrapolation (at least for the particular function class that we consider). We do not claim that our algorithm has theoretical guarantees which explain this increase in performance.

Fig. 4 shows the worst-case risk for varying instrument strength (\( \alpha_A \)). In the case where all exogenous noise comes from the unobserved variable \( \varepsilon_X \) (i.e., \( \alpha_A = 0 \)), the NILE coincides with the OLS estimator. In such settings, standard IV methods are known to perform poorly, although also the NPREGIV-1 method seems robust to such scenarios. As the instrument strength increases, the NILE clearly outperforms OLS and NPREGIV-1 for interventions on \( X \) which include values outside the training data.

We further compare NILE’s ability to estimate the causal function on the support of the covariate \( X \) in a nonlinear IV setting and compare it with the results from other state-of-the-art procedures for nonlinear IV estimation, following the experimental setup by [60]. Here, the authors consider a predictor variable \( X \sim \text{Uniform}(0,1) \) which causally influences the target variable \( Y \) via the structural assignment \( Y := f(X) + \xi_Y \), where \( f \) is the nonlinear causal function \( f(x) = \log [(16x - 8) + 1 \cdot \text{sgn}(x - 1/2)] \), and \( \xi_Y \) is an additive error term which is correlated with \( X \). They compare their proposed procedure Kernel IV to the methods NPREGIV-2 ([60] refer to this method as ‘Smooth IV’), Sieve IV and Deep IV (see Section 5.1). As a baseline, they also include a method for standard kernel ridge regression (‘Kernel Reg’) [56], which ignores the existence of hidden confounders. Each procedure yields a different estimator \( \hat{f} \). Based on 40 independent simulations, the estimators are then compared in terms of the average squared distance between \( f \) and \( \hat{f} \) across 1,000 equidistant points in the interval [0,1]. We refer to [60, Appendix A.11] for a precise description of the experimental setup. Fig. 5 shows the results of the above experiment (corresponding to Fig. 2 in [60]), where we have also included the NILE. Our method outperforms all other procedures, in particular for large sample sizes. There is slight difference in the way the different algorithms use the available data. In order to reduce finite sample bias, [60] use sample splitting, where the first and second step of the two-stage-least-squares procedure are performed on disjoint data sets. The NILE, in contrast, uses all of the data at once. However, even when running our procedure on only half of the data, we still outperform the other procedures by a distinct margin, see Fig. 3. We believe that the superior MSE performance of NILE could be due to the different approaches of regularization. For example, NILE uses causal regularization similar to that of PULSE, i.e., a data-driven K-class regularization; in linear IV settings, this type of regularization often yields a smaller MSE than standard IV methods such as TSLS [32].
6 Discussion and Future Work

In many real world problems, the test distribution may differ from the training distribution. This requires statistical methods that come with a provable guarantee in such a setting. It is possible to characterize robustness by considering predictive performance for distributions that are close to the training distribution in terms of standard divergences or metrics, such as KL divergences or Wasserstein distance. As an alternative view point, we have introduced a novel framework that formalizes the task of distribution generalization when considering distributions that are induced by a set of interventions. Based on the concept of modularity, interventions modify parts of the joint distribution and leave other parts invariant. Thereby, they impose constraints on the changes of the distributions that are qualitatively different from considering balls in the above metrics. As such, we see them as a useful language to describe realistic changes between training and test distributions.

Our framework is general in that it allows us to model a wide range of causal models and interventions, which do not need to be known beforehand. We have proved several generalization guarantees, some of which show robustness for distributions that are not close to the training distribution by considering almost any of the standard metrics. Here, generalization can be obtained by causal functions, but also by non-causal functions; in general, however, the minimizer changes when the intervention class is altered (or misspecified). We have further proved impossibility results that indicate the limits of what is possible to learn from the training distribution. In particular, in nonlinear models, strong assumptions are required for distribution generalization to a different support of the covariates. As such, methods such as anchor regression cannot be expected to work in nonlinear models, unless strong restrictions are placed on the function class $\mathcal{G}$.

Our work can be extended into several directions. It may, for example, be worthwhile to investigate the sharpness of the bounds we provide in Section 4.3.2 and other extrapolation assumptions on $\mathcal{F}$. Our results make use of the form of the squared loss and it remains an open question to which extent they hold for general convex loss functions. While our results can be applied to situations where causal background knowledge is available, via a transformation of SCMs, our analysis is deliberately agnostic about such information. It would be interesting to see whether stronger theoretical results can be obtained by including causal background information. We showed that the type of the interventions play a crucial role in determining whether the causal function is a minimax optimal solution. Building on this, it would be interesting to find empirical procedures which test whether an intervention is confounding-removing, confounding-preserving or neither. Finally, it could be worthwhile to investigate whether NILE, which outperforms existing approaches with respect to extrapolation, can be combined with non-parametric methods to further improve in-sample performance. While our current framework already contains certain settings of multi-task learning and domain generalization, it could be instructive to additionally include the possibility to model unlabeled data in the test task. Finally, our results concern the infinite sample case, but we believe that they can form the basis for a corresponding analysis involving rates or even finite sample results.

We view our work as a step towards understanding the problem of distribution generalization. We hope that considering the concepts of interventions may help to shed further light into the question of generalizing knowledge that was acquired during training to a different test distribution.

Acknowledgments

The authors thank Thomas Kneib for helpful discussions and two anonymous reviewers for valuable comments. RC and JP were supported by a research grant (18968) from VILLUM FONDEN; MEJ and JP were supported by the Carlsberg Foundation.

References

[1] S. S. Abadeh, P. M. M. Esfahani, and D. Kuhn, “Distributionally robust logistic regression,” in Proc. 28th Int. Conf. Neural Inf. Process. Syst., 2015, pp. 1576–1584.
[2] J. Aldrich, “Autonomy,” Oxford Econ. Papers, vol. 41, pp. 15–34, 1989.
[3] T. Amemiya, “The nonlinear two-stage least-squares estimator,” J. Econometrics, vol. 2, pp. 105–110, 1974.
[4] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, “Invariant risk minimization,” 2019, arXiv:1907.02839.
[5] P. Auer, N. Cesa-Bianchi, and P. Fischer, “Finite-time analysis of the multiarmed bandit problem,” Mach. Learn., vol. 47, no. 2/3, pp. 235–256, 2002.
[6] J. A. Bagnell, “Robust supervised learning,” in Proc. 20th Natl. Conf. Artif. Intell., 2005, pp. 714–719.
[7] P. L. Bartlett, V. Dani, T. Hayes, S. Kakade, A. Rakhlin, and A. Tewari, “High-probability regret bounds for bandit online linear optimization,” in Proc. 21st Ann. Conf. Learn. Theory, 2008, pp. 335–342.
[8] J. Baxter, “A model of inductive bias learning,” J. Artif. Intell. Res., vol. 12, pp. 149–198, 2000.
[9] S. Bickel, M. Brückner, and T. Scheffer, “Discriminative learning under covariate shift,” J. Mach. Learn. Res., vol. 10, no. 75, pp. 2137–2155, 2009.
[10] J. Blanchet, Y. Kang, K. Murthy, and F. Zhang, “Data-driven optimal transport cost selection for distributionally robust optimization,” in Proc. Winter Simul. Conf., 2019, pp. 3740–3751.
[11] K. A. Bollen, Structural Equations With Latent Variables. New York, NY, USA: Wiley, 1989.
[12] R. J. Bowden and D. A. Turkington, Instrumental Variables. Econometric Society Monographs. Cambridge, U.K.: Cambridge Univ. Press, 1985.
[13] R. Caruana, “Multitask learning,” Mach. Learn., vol. 28, pp. 41–75, 1997.
[14] X. Chen and T. M. Christensen, “Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression,” Quantitative Econ., vol. 9, no. 1, pp. 39–84, 2018.
[15] G. Csurka, “Domain adaptation for visual applications: A comprehensive survey,” 2017, arXiv:1907.02893.
[16] S. Darolles, Y. Fan, J.-P. Florens, and E. Renault, “Nonparametric instrumental regression,” Econometrica, vol. 79, no. 5, pp. 1541–1565, 2011.
[17] S. Dhamak, T. A. S. Kneib, and J. P. Scheffer, “Discriminative learning under covariate shift,” J. Mach. Learn. Res., vol. 10, no. 75, pp. 2137–2155, 2009.
[18] J. Blanchet, Y. Kang, K. Murthy, and F. Zhang, “Data-driven optimal transport cost selection for distributionally robust optimization,” in Proc. Winter Simul. Conf., 2019, pp. 3740–3751.
[19] K. A. Bollen, Structural Equations With Latent Variables. New York, NY, USA: Wiley, 1989.
[20] R. J. Bowden and D. A. Turkington, Instrumental Variables. Econometric Society Monographs. Cambridge, U.K.: Cambridge Univ. Press, 1985.
[21] R. Curran, “Multitask learning,” Mach. Learn., vol. 28, pp. 41–75, 1997.
[22] X. Chen and T. M. Christensen, “Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression,” Quantitative Econ., vol. 9, no. 1, pp. 39–84, 2018.
[23] G. Csurka, “Domain adaptation for visual applications: A comprehensive survey,” 2017, arXiv:1907.02893.
[24] S. Darolles, Y. Fan, J.-P. Florens, and E. Renault, “Nonparametric instrumental regression,” Econometrica, vol. 79, no. 5, pp. 1541–1565, 2011.
Rune Christiansen received the PhD degree from the University of Copenhagen, Denmark. He is currently a postdoc at the Department of Mathematical Sciences, University of Copenhagen, Denmark.

Niklas Pfister received the PhD degree from the Seminar for Statistics, ETH Zurich, Switzerland. He is currently an assistant professor at the Department of Mathematical Sciences, University of Copenhagen, Denmark.

Nicola Gnecco is currently working toward the PhD degree in the Research Center for Statistics, University of Geneva, Switzerland.

Jonas Peters is currently a professor at the Department of Mathematical Sciences, University of Copenhagen, Denmark. Previously, he has been at the MPI for Intelligent Systems in Tübingen and the Seminar for Statistics, ETH Zurich, Switzerland.

Martin Emil Jakobsen is currently working toward the PhD degree with the Department of Mathematical Sciences, University of Copenhagen, Denmark.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/csdl.