A causal framework for distribution generalization

Rune Christiansen[♭] Niklas Pfister[♭] Martin Emil Jakobsen[♭]
Nicola Gnecco[♯] Jonas Peters[♭]

[♭] *University of Copenhagen, Denmark*
{krunechristiansen, np, m.jakobsen, jonas.peters}@math.ku.dk

[♯] *University of Geneva, Switzerland*
nicola.gnecco@unige.ch

September 8, 2020

Abstract

We consider the problem of predicting a response from a set of covariates when the test distribution may differ from the training distribution. Motivated by the idea that such differences may have causal explanations, we consider a class of test distributions that emerge from interventions in a structural causal model, and focus on minimizing the worst-case risk over this class. Causal regression models, which regress the response variable on all of its direct parents, remain valid under arbitrary interventions on any subset of covariates, but they are not always optimal in the above sense. For example, in linear models, for a set of interventions with bounded strength, 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 and indirect interventions on the covariates. It takes into account that, in practice, minimax solutions need to be identified from observational data. Our framework allows us to characterize under which class of interventions the causal function is minimax optimal. We prove several sufficient conditions for distribution generalization and present corresponding impossibility results. We propose a practical method, called NILE, that achieves distribution generalization in a nonlinear instrumental variables setting with linear extrapolation. We prove consistency and present empirical results.

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 [55, 48, 17, 5]. 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 $\sup_{P\in\mathcal{N}} \mathbb{E}_P[(Y - f(X))^2]$ 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, 63]. A typical approach is to define a ρ-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 [7, 33]. 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, 63, 22, 11].

In our analysis, we do not start from a divergence measure, but instead construct a neighborhood of distributional changes based on the concept of interventions [49, 51]. 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, 28, 49]. Such a causal regression model yields the minimal worst-case risk when considering all interventions on variables other than Y [e.g., 57, Theorem 1, Appendix]. It has therefore been suggested to use causal models in problems of distributional shifts [60, 57, 30, 40, 43, 5, 5, 53]. One may argue, however, 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 [58, 52], see also Section 5.1. So-called anchor regression [58] 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 observational distribution.

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 [26] or physical [23] 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 [9, 54, 15, 41] (see Section 2.4 for more details), with a focus on minimizing the worst-case risk. In settings of covariate shift [e.g., 61, 64, 65], 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 [19, 10, 20, 45]. Sometimes, it is additionally assumed that the support of the training distribution covers that of the test distribution [61]. 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 a successful technique, for example in image classification, to adapt prediction procedures to such types of distribution shifts. These methods increase the diversity of the training dataset by changing the geometry and the color of the images (e.g., by rotation, cropping or changing saturation) [69, 62]. This allows the user to create models that generalize better to unseen environments [e.g., 67]. We view these approaches as a way 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 [21, 38], too. It can also be formulated in terms of minimizing the regret in a multi-armed bandit problem [39, 6, 8]. 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.

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, which learns a minimax solution from i.i.d. observational data (Section 5).

Our modeling framework describes how structural causal models can be used as technical devices for generating 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., 57], 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 model class 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 nonlinear 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 [58] – can be extended to nonlinear settings. Further, our framework implies impossibility statements for multi-task learning that relate to existing results [20].

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. Appendix A shows how to represent several causal models in our framework. Appendix B summarizes how existing results on identifiability in IV models can be exploited when considering Assumption 1. Appendix C provides details on the test statistic that we use for NILE. Appendix D contains an additional experiment. All proofs are provided in Appendix E.

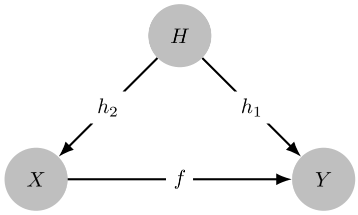
2 Framework

For a real-valued response variable $Y \in \mathbb{R}$ 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) [12, 49]. More precisely, denoting by $H \in \mathbb{R}^q$ some additional (unobserved) variables, we consider the SCM

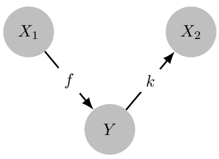
$$\begin{aligned} H &:= \varepsilon_H & &q \text{ assignments} \\ X &:= h_2(H, \varepsilon_X) & &d \text{ assignments} \\ Y &:= f(X) + h_1(H, \varepsilon_Y) & &1 \text{ assignment} \end{aligned}$$



Here, f, h_1 and h_2 are measurable functions, and the innovation terms ε_X, ε_Y and ε_H are independent vectors with possibly dependent coordinates. Two comments are in order. The joint distribution of (X, Y) is constrained only by requiring that X and $h_1(H, \varepsilon_Y)$ enter the equation of 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. Furthermore, we neither assume that the above SCM represents the true causal relationships between the random variables, nor do we assume any causal background knowledge of the system. Instead, the SCM is used only to construct the test distributions (by considering interventions on X) for which we are analyzing the predictive performance of different methods – similar to how one could have considered a ball around the training distribution. If causal background knowledge exists, however, e.g., in the form of an SCM over variables X and Y, it can be fit into the above framework. As such, our framework covers a large variety of models, including SCMs in which some of the variables in X are not ancestors but descendants of Y (this requires adapting the set of interventions appropriately), see Appendix A for details. The following remark shows such an example, and may be interesting to readers with a special interest in causality. It can be skipped at first reading.

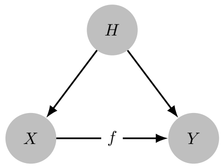
Remark 1 (Rewriting causal background knowledge)**.** *If a priori causal background knowledege is available, e.g., in form of an SCM, our framework is still applicable after an appropriate transformation. The following example shows a reformulation of an SCM over variables X_1, X_2 and Y.*

$$\begin{aligned} X_1 &:= \varepsilon_1 \\ X_2 &:= k(Y) + \varepsilon_2 \\ Y &:= f(X_1) + \varepsilon_3, \end{aligned}$$
with $(\varepsilon_1, \varepsilon_2, \varepsilon_3) \sim Q$.



$\xrightarrow{\textit{rewrite}}$

$$\begin{aligned} H &:= \varepsilon_3 \\ X &:= h_2(H, (\varepsilon_1, \varepsilon_2)) \\ Y &:= f(X_1) + H, \end{aligned}$$
with $(\varepsilon_1, \varepsilon_2, \varepsilon_3) \sim Q$.



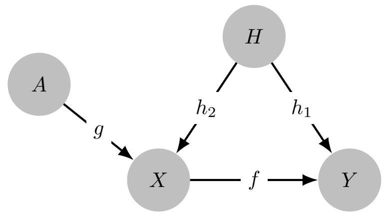
Here, $h_2(H, (\varepsilon_1, \varepsilon_2)) := (\varepsilon_1, k(f(\varepsilon_1) + H) + \varepsilon_2)$. Both SCMs induce the same observational distribution over (X_1, X_2, Y) and any intervention on the covariates in the SCM on the left-hand side can be rewritten as an intervention on the covariates in the SCM on the right-hand side. Details and a more general treatment are provided in Appendix A.

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 causal function 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., 14, 27], a well-established tool for ensuring that f can be uniquely recovered from the observational distribution of (X, Y). (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 article, we will therefore consider a slightly larger class of SCMs that also includes exogenous variables A. It contains the SCM presented at the beginning of Section 2.1 as a special case.[1] We derive results for settings with and without exogeneous 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 $H \in \mathbb{R}^q$. Let further $\mathcal{F} \subseteq \{f : \mathbb{R}^d \to \mathbb{R}\}$, $\mathcal{G} \subseteq \{g : \mathbb{R}^r \to \mathbb{R}^d\}$, $\mathcal{H}_1 \subseteq \{h_1 : \mathbb{R}^{q+1} \to \mathbb{R}\}$ and $\mathcal{H}_2 \subseteq \{h_2 : \mathbb{R}^{q+d} \to \mathbb{R}^d\}$ be fixed sets of measurable functions. Moreover, let $\mathcal{Q}$ be a collection of probability distributions on $\mathbb{R}^{d+1+r+q}$, such that for all $Q \in \mathcal{Q}$ it holds that if $(\varepsilon_X, \varepsilon_Y, \varepsilon_A, \varepsilon_H) \sim Q$, then $\varepsilon_X, \varepsilon_Y, \varepsilon_A$ and ε_H are jointly independent, and for all $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$ it holds that $\xi_Y := h_1(\varepsilon_H, \varepsilon_Y)$ and $\xi_X := h_2(\varepsilon_H, \varepsilon_X)$ 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 an SCM by[3]

$$\begin{aligned} A &:= \varepsilon_A && r \text{ assignments} \\ H &:= \varepsilon_H && q \text{ assignments} \\ X &:= g(A) + h_2(H, \varepsilon_X) && d \text{ assignments} \\ Y &:= f(X) + h_1(H, \varepsilon_Y) && 1 \text{ assignment} \end{aligned}$$



with $(\varepsilon_X, \varepsilon_Y, \varepsilon_A, \varepsilon_H) \sim Q$. 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 a constant function.

2.3 Interventions

Each SCM $M \in \mathcal{M}$ can now be modified by the concept of interventions [e.g., 49, 51]. 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).

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^i, A^i, Y^i, H^i), 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^i, Y^i, A^i, H^i)[4], which has finite second moments. We denote the corresponding distribution over the observed (X^i, Y^i, A^i) by $\mathbb{P}_{M(i)}$, and use $\mathcal{I}$ to denote a collection of interventions. In our work, the test distribution is modeled as the distribution generated by these types of intervened models and the set $\mathcal{I}$ therefore indexes all possible test distributions. We will be

[1] This follows from choosing A as an independent noise variable and a constant g.

[2] This can be assumed without loss of generality if $\mathcal{F}$ and $\mathcal{G}$ are closed under addition and scalar multiplication, and contain the constant function.

[3] For an appropriate choice of h_2, the model includes settings in which (parts of) A directly influence Y.

[4] If the context does not allow for any ambiguity, we omit the superscript i.

interested in the mean squared prediction error on each test distribution i, formally written as $\mathbb{E}_{M(i)}[(Y - f(X))^2]$.

The support of random variables under interventions will play an important role for the analysis of distribution generalization. Throughout this paper, $\mathrm{supp}^M(Z)$ 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, $\mathrm{supp}^M_{\mathcal{I}}(Z)$ denotes the union of $\mathrm{supp}^{M(i)}(Z)$ over all interventions $i \in \mathcal{I}$. We call a collection of interventions on Z *support-reducing* (w.r.t. M) if $\mathrm{supp}^M_{\mathcal{I}}(Z) \subseteq \mathrm{supp}^M(Z)$ and *support-extending* (w.r.t. M) if $\mathrm{supp}^M_{\mathcal{I}}(Z) \not\subseteq \mathrm{supp}^M(Z)$. Whenever it is clear from the context which model is considered, we may drop the indication of M altogether and simply write $\mathrm{supp}(Z)$.

2.4 Distribution generalization

Let $\mathcal{M}$ be a fixed model class, let $M = (f, g, h_1, h_2, Q) \in \mathcal{M}$ be the true (but unknown) data generating model, and let $\mathcal{I}$ 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 $\hat{Y} = f^*(X)$ has low worst-case risk over all distributions induced by the interventions in $\mathcal{I}$. We therefore consider the optimization problem

$$\operatorname*{argmin}_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f_\diamond(X))^2\big], \tag{1}$$

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^* exists, we refer to it as a *minimax solution* (for model M w.r.t. $(\mathcal{F}, \mathcal{I})$).

Depending on the model class $\mathcal{M}$, there may be several models $\tilde{M} \in \mathcal{M}$ 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 $\mathcal{I}$. 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 (1). In this paper, we study conditions on $\mathcal{M}$, $\mathbb{P}_M$ and $\mathcal{I}$, under which this becomes possible. More precisely, we aim to characterize under which conditions $(\mathbb{P}_M, \mathcal{M})$ admits distribution generalization to $\mathcal{I}$.

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

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

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 $f^* \in \mathcal{F}$ which, for every $\tilde{M} \in \mathcal{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, is a minimax solution for $\tilde{M}$ w.r.t. $(\mathcal{F}, \mathcal{I})$, then, in particular, $(\mathbb{P}_M, \mathcal{M})$ admits generalization to $\mathcal{I}$.

Our framework also includes several settings of multitask learning (MTL) and domain adaptation [54], 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 (1). In the simplest case, where $\mathcal{I}$ consists only of the trivial intervention, that is, $\mathbb{P}_M = \mathbb{P}_{M(i)}$, 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. Since the structural assignment for Y remains unchanged for all interventions that we consider in this work, it holds for all $f_\diamond \in \mathcal{F}$ and all interventions i on either A or X that

$$\mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] = \mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] + \mathbb{E}_M[\xi_Y^2] + 2\mathbb{E}_{M(i)}[\xi_Y(f(X) - f_\diamond(X))].$$

Here, the middle term does not depend on i since $\xi_Y = h_1(H, \varepsilon_Y)$ remains fixed. If i is an intervention such that $\xi_Y \perp\!\!\!\perp X$ under $\mathbb{P}_{M(i)}$ (which is the case for interventions we will call confounding-removing), then, because of $\mathbb{E}_M[\xi_Y] = 0$, the last term in the above equation vanishes. Therefore, if $\mathcal{I}$ consists only of confounding-removing interventions, the causal function is a solution to the minimax problem (1). The following proposition shows that an even stronger statement holds: The causal function is already a minimax solution if $\mathcal{I}$ contains at least one confounding-removing intervention on X.

Proposition 3.1 (confounding-removing interventions on X)**.** *Let $\mathcal{I}$ be a set of interventions on X or A such that there exists at least one $i \in \mathcal{I}$ that we call confounding-removing, which means that for all $M \in \mathcal{M}$, the variables X and H are independent under $M(i)$. Then, the minimal worst-case risk is attained at a confounding-removing intervention, and the causal function f is a minimax solution.*

We now prove that, in a linear setting, the causal function is also a minimax solution if the interventions create unbounded variability in all directions of the covariance matrix of X.

Proposition 3.2 (unbounded interventions on X with linear $\mathcal{F}$)**.** *Let $\mathcal{F}$ be the class of all linear functions, and let $\mathcal{I}$ be a set of interventions on X or A s.t. $\sup_{i\in\mathcal{I}} \lambda_{\min}\big(\mathbb{E}_{M(i)}\big[XX^\top\big]\big) = \infty$, where $\lambda_{\min}$ denotes the smallest eigenvalue. Then, the causal function f is the unique minimax solution.*

The unbounded eigenvalue condition above is satisfied if $\mathcal{I}$ 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 [58, 59]. The proposition above considers a linear function class $\mathcal{F}$; in this way, shift interventions are related to linear models.

Even if the causal function f does not solve the minimax problem (1), the difference between the minimax solution and the causal function cannot be arbitrarily large. The following proposition shows that the worst-case L_2-distance between f and any function $f_\diamond$ that performs better than f (in terms of worst-case risk) can be bounded by a term which is related to the strength of the confounding.

Proposition 3.3 (difference between causal function and minimax solution)**.** *Let $\mathcal{I}$ be a set of interventions on X or A. Then, for any function $f_\diamond \in \mathcal{F}$ which satisfies that*

$$\sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] \leq \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - f(X))^2],$$

it holds that

$$\sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] \leq 4\operatorname{Var}_M[\xi_Y].$$

Even though the difference can be bounded, it may be non-zero, and one may benefit from choosing a function that differs from the causal function f. This choice, however, comes at a cost: it relies on the fact that we know the class of interventions $\mathcal{I}$. In general, being a minimax solution is not entirely robust with respect to misspecification of $\mathcal{I}$. In particular, if the set $\mathcal{I}_2$ of interventions describing the test distributions is misspecified by a set $\mathcal{I}_1 \neq \mathcal{I}_2$, then the considered minimax solution with respect to $\mathcal{I}_1$ may perform worse than the causal function on the test distributions.

Proposition 3.4 (properties of the minimax solution under mis-specified interventions)**.** *Let $\mathcal{I}_1$ and $\mathcal{I}_2$ be any two sets of interventions on X, and let $f_1^* \in \mathcal{F}$ be a minimax solution w.r.t. $\mathcal{I}_1$. Then, if $\mathcal{I}_2 \subseteq \mathcal{I}_1$, it holds that*

$$\sup_{i \in \mathcal{I}_2} \mathbb{E}_{M(i)}\big[(Y - f_1^*(X))^2\big] \leq \sup_{i \in \mathcal{I}_2} \mathbb{E}_{M(i)}\big[(Y - f(X))^2\big].$$

If $\mathcal{I}_2 \not\subseteq \mathcal{I}_1$, however, it can happen (even if $\mathcal{F}$ is linear) that

$$\sup_{i \in \mathcal{I}_2} \mathbb{E}_{M(i)}\big[(Y - f_1^*(X))^2\big] > \sup_{i \in \mathcal{I}_2} \mathbb{E}_{M(i)}\big[(Y - f(X))^2\big].$$

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 $\mathcal{M}$ containing the unknown data generating model M, and let $\mathcal{I}$ be a class of interventions. By definition, the optimizer of the minimax problem (1) depends on the true model M. 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 $\mathbb{P}_M$. This motivates the notion of distribution generalization, see (2). In words, it states that approximate minimax solutions (which depend on the intervention distributions $\mathbb{P}_{M(i)}$, $i \in \mathcal{I}$) are identified from the observational distribution $\mathbb{P}_M$. This holds true, in particular, if the intervention distributions themselves are identified from $\mathbb{P}_M$.

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

$$\mathbb{P}_{\tilde{M}} = \mathbb{P}_M \quad \Rightarrow \quad \mathbb{P}_{\tilde{M}(i)}^{(X,Y)} = \mathbb{P}_{M(i)}^{(X,Y)} \quad \forall i \in \mathcal{I},$$

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

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 $\mathcal{M}$, $\mathbb{P}_M$ and $\mathcal{I}$ which ensure distribution 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 $\mathbb{P}_M$ (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 1.

intervention	$\text{supp}_{\mathcal{I}}(X)$	assumptions	result
on X (well-behaved)	$\subseteq \text{supp}(X)$	Assumption 1	Proposition 4.3
on X (well-behaved)	$\not\subseteq \text{supp}(X)$	Assumptions 1 and 2	Proposition 4.4
on A	$\subseteq \text{supp}(X)$	Assumptions 1 and 3	Proposition 4.8
on A	$\not\subseteq \text{supp}(X)$	Assumptions 1, 2 and 3	Proposition 4.8

Table 1: Summary of conditions under which generalization is possible. Corresponding impossibility results are shown in Propositions 4.2, 4.7 and 4.9.

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, distribution generalization is directly implied by the following assumption.

Assumption 1 (Identifiability of f on the support of X). *For all $\tilde{M} = (\tilde{f}, \dots) \in \mathcal{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, it holds that $\tilde{f}(x) = f(x)$ 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., 68]. In linear non-Gaussian models, one can identify the function f even if there are no instruments [32]. For nonlinear models, restricted structural causal models can be exploited, too. In that case, Assumption 1 holds under regularity conditions if $h_1(H, \varepsilon_Y)$ is independent of X [70, 50, 51] and first attempts have been made to extend such results to non-trivial confounding cases [35]. The nonlinear IV setting [e.g., 3, 46, 47] is discussed in more detail in Appendix B, 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 $\mathbb{P}_M$-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 $\mathcal{I}$. 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 ψ^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 (ψ^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

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

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

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

In both cases, $(\varepsilon_X^i, \varepsilon_Y^i, \varepsilon_A^i, \varepsilon_H^i) \sim Q$ and $I^i \perp\!\!\!\perp (\varepsilon_X^i, \varepsilon_Y^i, \varepsilon_A^i, \varepsilon_H^i)$. 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 ε_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 contraints on the interventions

Let $\mathcal{Q}$ 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

$$A := \varepsilon_A, \quad H := \sigma\varepsilon_H, \quad X := \gamma A + \varepsilon_X + \tfrac{1}{\sigma}H, \quad Y := \beta X + \varepsilon_Y + \tfrac{1}{\sigma}H, \tag{3}$$

with $\gamma, \beta \in \mathbb{R}$, $\sigma > 0$ and $(\varepsilon_A, \varepsilon_X, \varepsilon_Y, \varepsilon_H) \sim Q \in \mathcal{Q}$. Assume that $\mathbb{P}_M$ 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 $\mathcal{I} \subseteq \mathbb{R}_{>0}$ be a compact, non-empty set and define the interventions on X by $\psi^i(g, h_2, A^i, H^i, \varepsilon_X^i, I^i) = iH$, for $i \in \mathcal{I}$. Then, $(\mathbb{P}_M, \mathcal{M})$ 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 $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, such that*

$$\left| \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - bX)^2\big] - \inf_{b_\diamond\in\mathbb{R}} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - b_\diamond X)^2\big] \right| \geq c.$$

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 σ, this includes all models of the form $M_{\tilde{\sigma}} = M(\gamma, \beta, \tilde{\sigma}, Q)$ for some $\tilde{\sigma} > 0$. However, while agreeing on the observational distribution, each of these models induces fundamentally different intervention distributions (under $M_{\tilde{\sigma}}(i)$, (X, Y) is equal in distribution to $(i\varepsilon_H, (\beta i + \frac{1}{\tilde{\sigma}})\varepsilon_H))$ 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

Let i be an intervention on X with intervention map ψ^i. As already mentioned in Section 3, the intervention is then called

confounding-removing	if for all models $M \in \mathcal{M}$, $\psi^i(g, h_2, A^i, H^i, \varepsilon_X^i, I^i) \perp\!\!\!\perp H^i$ under $M(i)$.

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 is called

confounding-preserving	if there exists a map φ^i, such that $\psi^i(g, h_2, A^i, H^i, \varepsilon_X^i, I^i) = \varphi^i(A^i, g(A^i), h_2(H^i, \varepsilon_X^i), I^i)$.

Confounding-preserving interventions contain, e.g., *shift interventions* on X, which linearly shift the original assignment by I^i, that is, $\psi^i(g, h_2, A^i, H^i, \varepsilon_X^i, I^i) = g(A^i) + h_2(H^i, \varepsilon_X^i) + I^i$. The name 'confounding-preserving' stems from the fact that the unobserved (confounding) variables H only enter the intervened structural assignment of X via the term $h_2(H^i, \varepsilon_X^i)$, which is the same as in the original model. (This property fails to hold true for the interventions in Proposition 4.2.) If $\mathcal{I}$ 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 $\mathbb{P}_{M(i)}$, $i \in \mathcal{I}$, and therefore ensure distribution 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 $\psi^i(g, h_2, A^i, H^i, \varepsilon_X^i, I^i) = \varepsilon_X^i$ 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 $\mathbb{P}_M$ 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 $\operatorname{supp}_{\mathcal{I}}(X) \subseteq \operatorname{supp}(X)$. Then, under Assumption 1, $(\mathbb{P}_M, \mathcal{M})$ 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., $\operatorname{supp}_{\mathcal{I}}(X) \not\subseteq \operatorname{supp}(X)$, Assumption 1 is not sufficient for ensuring distribution generalization. This is because there may exist a model $\tilde{M} \in \mathcal{M}$ which agrees with M on the observational distribution, but whose corresponding causal function $\tilde{f}$ 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 $\operatorname{supp}(X)$.

Assumption 2 (Extrapolation of $\mathcal{F}$)**.** *For all $\tilde{f}, \bar{f} \in \mathcal{F}$ with $\tilde{f}(x) = \bar{f}(x)$ for all $x \in \operatorname{supp}(X)$, it holds that $\tilde{f} \equiv \bar{f}$.*

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, $(\mathbb{P}_M, \mathcal{M})$ 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 satisifed 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.

Sufficient conditions for generalization Assumption 2 states that every function in $\mathcal{F}$ is globally identified by its values on $\operatorname{supp}(X)$. 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 $\operatorname{supp}(X)$. More precisely,

$$\mathcal{F} \text{ is linearly closed}: \quad f_1, f_2 \in \mathcal{F}, c \in \mathbb{R}, \implies f_1 + f_2 \in \mathcal{F}, cf_1 \in \mathcal{F}, \text{ and} \tag{4}$$

$$\mathcal{F} \text{ is lin. ind. on } \operatorname{supp}(X): \quad f_1(x) = 0 \quad \forall x \in \operatorname{supp}(X) \implies f_1(x) = 0 \quad \forall x \in \mathcal{D}. \tag{5}$$

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

$$\mathcal{F}^1 := \{f_\diamond : \mathcal{D} \to \mathbb{R} \mid \text{ there exists } \gamma \in \mathbb{R}^k \text{ s.t. } \forall x \in \mathcal{D} : f_\diamond(x) = \gamma^\top \nu(x)\},$$

where $\nu = (\nu_1, \ldots, \nu_k)$ consists of real-valued, linearly independent functions satisfying that $\mathbb{E}_M[\nu(X)\nu(X)^\top]$ is strictly positive definite, and (ii) the class of differentiable functions that extend linearly outside of $\operatorname{supp}(X)$, that is, $\mathcal{F}$ is of the form

$$\mathcal{F}^2 := \{f_\diamond : \mathcal{D} \to \mathbb{R} \mid f_\diamond \in C^1 \text{ and } \forall x \in \mathcal{D} \setminus \operatorname{supp}(X) : f_\diamond(x) = f_\diamond(x_b) + \nabla f_\diamond(x_b)(x - x_b)\},$$

where $x_b := \operatorname{argmin}_{z \in \operatorname{supp}(X)} \|x - z\|$ and $\operatorname{supp}(X)$ is assumed to be closed with non-empty interior. Clearly, both of the above function classes are linearly closed. To see that $\mathcal{F}^1$ satisfies (5), let $\gamma \in \mathbb{R}^k$ be s.t. $\gamma^\top \nu(x) = 0$ for all $x \in \operatorname{supp}(X)$. Then, it follows that $0 = \mathbb{E}_M[(\gamma^\top \nu(X))^2] = \gamma^\top \mathbb{E}_M[\nu(X)\nu(X)^\top]\gamma$ and hence that $\gamma = 0$. To see that $\mathcal{F}^2$ satisfies (5), let $f_\diamond \in \mathcal{F}^2$ and assume that $f_\diamond(x) = 0$ for all $x \in \operatorname{supp}(X)$. Then, $f_\diamond(x) = 0$ for all $x \in \mathcal{D}$ and thus $\mathcal{F}^2$ 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.

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

$$\mathcal{F}^3 := \{f_\diamond : \mathcal{D} \to \mathbb{R} \mid f_\diamond \text{ is continuously differentiable with } \|\nabla f_\diamond\|_\infty \leq K\}, \tag{6}$$

for some fixed $K < \infty$, where $\nabla f_\diamond$ 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 $\operatorname{supp}(X)$ and all $\tilde{M} \in \mathcal{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, it holds that*

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

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

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 δ. 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 $f^* \in \mathcal{F}$ be such that,*

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

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

$$\begin{aligned} &\left| \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f^*(X))^2\right] - \inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f_\diamond(X))^2\right] \right| \\ &\quad \le \varepsilon + 12\delta^2 K^2 + 32\delta K \sqrt{\mathrm{Var}_M(\xi_Y)} + 4\sqrt{2}\delta K \sqrt{\varepsilon} \end{aligned}$$

where $\delta := \sup_{x\in\mathrm{supp}_{\mathcal{I}}^M(X)} \inf_{z\in\mathrm{supp}^M(X)} \|x - z\|$.

We can take f^* to be the minimax solution if it exists. In that case, the terms involving ε disappear from the bound, which then becomes more similar to the one in Proposition 4.5.

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

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

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 $\tilde{M} = (\tilde{f}, \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q}) \in \mathcal{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, it holds that $\tilde{g}(a) = g(a)$ for all $a \in \mathrm{supp}(A) \cup \mathrm{supp}_{\mathcal{I}}(A)$.*

Since $g(A)$ is a conditional mean for X given A, the values of g are identified from $\mathbb{P}_M$ for $\mathbb{P}_M$-almost all a. If $\mathrm{supp}_{\mathcal{I}}(A) \subseteq \mathrm{supp}(A)$, 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 $a \in \mathbb{R}^r$. In the case where the interventions $\mathcal{I}$ extend the support of A, we additionally require the function class $\mathcal{G}$ to extrapolate from $\text{supp}(A)$ to $\text{supp}(A) \cup \text{supp}_{\mathcal{I}}(A)$; 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, $(\mathbb{P}_M, \mathcal{M})$ admits generalization to $\mathcal{I}$ if either $\text{supp}_{\mathcal{I}}(X) \subseteq \text{supp}(X)$ 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 $\mathcal{F}$, $\mathcal{G}$, $\mathcal{H}_1$ and $\mathcal{H}_2$ consist of linear functions, Anchor regression [58] and K-class estimators [34] 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 $\text{supp}_{\mathcal{I}}(A) \subseteq \text{supp}(A)$, it is possible to drop Assumption 1.

4.4.1 Impossibility of generalization without constraints on $\mathcal{G}$

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

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

This proposition is part of the argument showing that anchor regression [58] 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., 20].

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 (1) 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

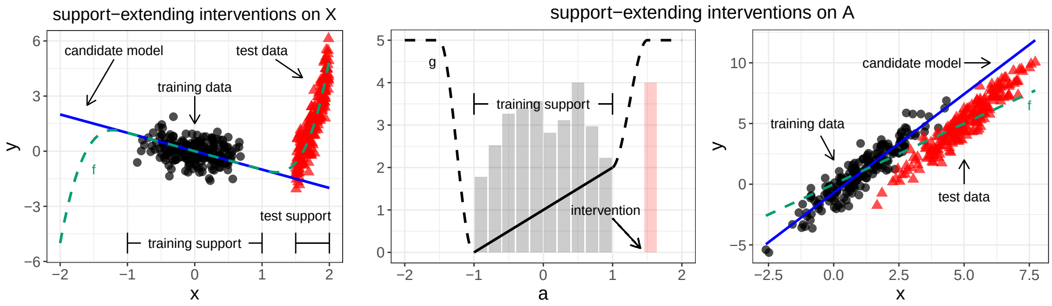


Figure 1: The left plot illustrates the straight-forward idea behind the impossibility result in Proposition 4.7. The plots in the middle and on the right-hand side illustrate the impossibility result in Proposition 4.9. All plots visualize the case of univariate variables. Under well-behaved interventions on X (left; here using confounding-removing interventions) which extend the support of X, generalization is impossible without further restrictions on the function class $\mathcal{F}$. This holds true even if Assumption 1 is satisfied. Indeed, although the candidate model (blue line) coincides with the causal model (green dashed curve) on the support of X, it may perform arbitrarily bad on test data generated under support-extending interventions. Under interventions on A (right and middle), generalization is impossible even under strong assumptions on the function class $\mathcal{F}$ (here, $\mathcal{F}$ is the class of all linear functions). Any support-extending intervention on A shifts the marginal distribution of X by an amount which depends on the (unknown) function g, resulting in a distribution of (X, Y) which cannot be identified from the observational distribution. Without further restrictions on the function class $\mathcal{G}$, any candidate model apart from the causal model may result in arbitrarily large worst-case risk.

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, 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. Their target of inference is to estimate the causal function, which by Proposition 3.1 will coincide with the minimax solution if the set $\mathcal{I}$ 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

model class	interventions	$\text{supp}_{\mathcal{I}}(X)$	assumptions	algorithm
$\mathcal{F}$ linear	on X or A of which at least one is confounding-removing	–	Ass. 1	linear IV (e.g., two-stage least squares, K-class or PULSE [66, 34])
$\mathcal{F}, \mathcal{G}$ linear	on A	bounded strength	–	anchor regression [58], see also [66]
$\mathcal{F}$ smooth	on X or A of which at least one is confounding-removing	support-reducing	Ass. 1	nonlinear IV (e.g., NPREGIV [56])
$\mathcal{F}$ smooth and linearly extrapolates	on X or A of which at least one is confounding-removing	–	Ass. 1	**NILE** Section 5.2

Table 2: List of algorithms to learn the generalizing function from data, the considered model class, types of interventions, support under interventions, and additional model assumptions. Sufficient conditions for Assumption 1 are given, for example, in the IV literature by generalized rank conditions, see Appendix B.

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., 42]. K-class estimators [66] 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., 42]. More recently, PULSE has been proposed [34], 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 [58] 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 [44]. 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 difficult. In cases where the causal function is among such solutions, this problem has been studied in the econometrics community. For example, [46, 47] treat the identifiability and estimation of causal functions in non-parametric function classes. Several non-parametric IV procedures exists, e.g., NPREGIV [56] contains modified implementations of [18] and [31]. Identifiability and estimation of the causal function using nonlinear IV methods in parametric function classes is discussed in Appendix B. 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 $\mathcal{F}$ 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 $\mathcal{F}$ are uniquely determined by their values on the support of X, that is, Assumption 2 is satisfied. In our simulations, we show that our method can improve the prediction performance on the boundary of 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 $\mathcal{F}$. In this section, we start from such assumptions and verify both theoretically and practically that they allows 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 $\mathcal{F}$, any function from $\mathcal{F}$ 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 $B = (B_1, ..., B_k)$ be a fixed B-spline basis on $[a, b]$, and define $\eta := (a, b, B)$. Our procedure assumes that the true causal function f belongs to the function class $\mathcal{F}_\eta := \{f_\eta(\cdot; \theta) : \theta \in \mathbb{R}^k\}$, where for every $x \in \mathbb{R}$ and $\theta \in \mathbb{R}^k$, $f_\eta(x; \theta)$ is given as

$$f_\eta(x;\theta) := \begin{cases} B(a)^\top\theta + B'(a)^\top\theta(x-a) & \text{if } x < a \\ B(x)^\top\theta & \text{if } x \in [a,b] \\ B(b)^\top\theta + B'(b)^\top\theta(x-b) & \text{if } x > b, \end{cases} \tag{7}$$

where $B' := (B'_1, \dots, B'_k)$ 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 $(\mathbf{X}, \mathbf{Y}, \mathbf{A}) \in \mathbb{R}^{n\times 3}$ be n i.i.d. realizations sampled from a distribution over (X, Y, A), let $\eta = (a, b, B)$ be fixed and assume that $\mathrm{supp}(X) \subseteq [a, b]$. Our algorithm aims to learn the causal function $f_\eta(\cdot; \theta^0) \in \mathcal{F}_\eta$, which is determined by the linear causal parameter θ^0 of a k-dimensional vector of covariates $(B_1(X), \dots, B_k(X))$. 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. We therefore artificially generate such instruments by nonlinearly transforming A, by using another B-spline basis $C = (C_1, \dots, C_k)$. The parameter θ^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 $\tilde{H}_0(\theta) : \mathbb{E}[C(A)(Y - B(X)^\top\theta)] = 0$. Let $\mathbf{B} \in \mathbb{R}^{n\times k}$ and $\mathbf{C} \in \mathbb{R}^{n\times k}$ be the associated design matrices, for each $i \in \{1, \dots, n\}$, $j \in \{1, \dots, k\}$ given as $\mathbf{B}_{ij} = B_j(X_i)$ and $\mathbf{C}_{ij} = C_j(A_i)$. A straightforward choice would be to construct the standard TSLS estimator, i.e., $\hat{\theta}$ as the minimizer of $\theta \mapsto \|\mathbf{P}(\mathbf{Y} - \mathbf{B}\theta)\|_2^2$, where $\mathbf{P}$ is the projection matrix onto the columns of $\mathbf{C}$, see also [29]. 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 $\mathbf{C}$ are linearly independent, $\mathbf{P}$ is equal to the identity matrix, and $\hat{\theta}$ coincides with the OLS estimator. Second, since θ corresponds to the linear parameter of a spline basis, it seems reasonable to impose constraints on θ 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 $\mathbf{K} \in \mathbb{R}^{k \times k}$ and $\mathbf{M} \in \mathbb{R}^{k \times k}$ be the matrices that are, for each $i, j \in \{1, \dots, k\}$, defined as $\mathbf{K}_{ij} = \int B_i''(x) B_j''(x) dx$ and $\mathbf{M}_{ij} = \int C_i''(a) C_j''(a) da$, and let $\gamma, \delta > 0$ be the respective penalties associated with $\mathbf{K}$ and $\mathbf{M}$. For $\lambda \geq 0$ and with $\mu := (\gamma, \delta, C)$, we then define the estimator

$$\hat{\theta}^n_{\lambda,\eta,\mu} := \underset{\theta \in \mathbb{R}^k}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{B}\theta\|_2^2 + \lambda \|\mathbf{P}_\delta (\mathbf{Y} - \mathbf{B}\theta)\|_2^2 + \gamma \theta^\top \mathbf{K} \theta, \tag{8}$$

where $\mathbf{P}_\delta := \mathbf{C}(\mathbf{C}^\top \mathbf{C} + \delta \mathbf{M})^{-1} \mathbf{C}^\top$ is the 'hat'-matrix for a penalized regression onto the columns of $\mathbf{C}$. By choice of $\mathbf{K}$, the term $\theta^\top \mathbf{K} \theta$ is equal to the integrated squared curvature of the spline function parametrized by θ. The above may thus be seen as a nonlinear extension of K-class estimators [66], with an additional penalty term which 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 multivariate functions. For example, [24] 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 $\mathcal{F}_\eta$, the above defines the global estimate $f_\eta(\cdot; \hat{\theta}^n_{\lambda,\eta,\mu})$, for every $x \in \mathbb{R}$ given by

$$f_\eta(x; \hat{\theta}^n_{\lambda,\eta,\mu}) := \begin{cases} B(a)^\top \hat{\theta}^n_{\lambda,\eta,\mu} + B'(a)^\top \theta^n_{\lambda,\eta,\mu}(x - a) & \text{if } x < a \\ B(x)^\top \hat{\theta}^n_{\lambda,\eta,\mu} & \text{if } x \in [a, b] \\ B(b)^\top \theta^n_{\lambda,\eta,\mu} + B'(b)^\top \theta^n_{\lambda,\eta,\mu}(x - b) & \text{if } x > b. \end{cases} \tag{9}$$

We deliberately distinguish between three different groups of hyper-parameters η, μ and λ. The parameter $\eta = (a, b, B)$ 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 λ 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, γ and δ are chosen via a cross-validation procedure, see Section 5.2.2. The parameter λ determines the relative contribution of the OLS and TSLS losses to the objective function. To choose λ from data, we use an idea similar to the PULSE [34].

5.2.2 Algorithm

Let for now η, μ be fixed. In the limit $\lambda \to \infty$, our estimation procedure becomes equivalent to minimizing the TSLS loss $\theta \mapsto \|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2$, which may be interpreted as searching for the parameter θ which complies 'best' with the hypothesis $\tilde{H}_0(\theta) : \mathbb{E}[C(A)(Y - B(X)^\top \theta)] = 0$. For finitely many data, following the idea introduced in [34], we propose to choose the value for λ such that $\tilde{H}_0(\hat{\theta}^n_{\lambda,\eta,\mu})$ 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 α for $\tilde{H}_0(\theta)$ with rejection threshold $q(\alpha)$. That is, $T_n(\theta)$ does not reject $\tilde{H}_0(\theta)$ if and only if $T_n(\theta) \leq q(\alpha)$. The penalty $\lambda_n^\star$ is then chosen in the following data-driven way

$$\lambda_n^\star := \inf\{\lambda \geq 0 : T_n(\hat{\theta}^n_{\lambda,\eta,\mu}) \leq q(\alpha)\}.$$

In general, $\lambda_n^\star$ is not guaranteed to be finite for an arbitrary test statistic T_n. Even for a reasonable test statistic it might happen that $T_n(\hat{\theta}^n_{\lambda,\eta,\mu}) > q(\alpha)$ for all $\lambda \geq 0$; see [34] 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 $\lambda_n^\star$ is not finite. Furthermore, if $\lambda \mapsto T_n(\hat{\theta}^n_{\lambda,\eta,\mu})$ is monotonic, $\lambda_n^\star$ 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 $f_\eta(\cdot, \hat{\theta}_{\lambda_n^\star,\mu,\eta})$, given that $\mathcal{F}_\eta$ is correctly specified, are presented in Section 5.2.3. Two choices of test statistics which are implemented in our code package can be found in Appendix C.

For every $\gamma \geq 0$, let $\mathbf{Q}_\gamma = \mathbf{B}(\mathbf{B}^\top \mathbf{B} + \gamma \mathbf{K})^{-1}\mathbf{B}^\top$ be the 'hat'-matrix for the penalized regression onto $\mathbf{B}$. Our algorithm then proceeds as follows.

Algorithm 1: NILE ("Nonlinear Intervention-robust Linear Extrapolator")

input: data $(\mathbf{X}, \mathbf{Y}, \mathbf{A}) \in \mathbb{R}^{n\times 3}$;
options: k, T, q, α;
begin
  $a \leftarrow \min_i X_i$, $b \leftarrow \max_i X_i$;
  construct cubic B-spline bases $B = (B_1, \ldots, B_k)$ and $C = (C_1, \ldots, C_k)$ at equidistant knots, with boundary knots at respective extreme values of $\mathbf{X}$ and $\mathbf{A}$;
  define $\hat{\eta} \leftarrow (a, b, B)$;
  choose $\delta^n_{\mathrm{CV}} > 0$ by 10-fold CV to minimize the out-of-sample mean squared error of $\hat{\mathbf{Y}} = \mathbf{P}_\delta \mathbf{Y}$;
  choose $\gamma^n_{\mathrm{CV}} > 0$ by 10-fold CV to minimize the out-of-sample mean squared error of $\hat{\mathbf{Y}} = \mathbf{Q}_\gamma \mathbf{Y}$;
  define $\mu^n_{\mathrm{CV}} \leftarrow (\delta^n_{\mathrm{CV}}, \gamma^n_{\mathrm{CV}}, C)$;
  approximate $\lambda_n^\star = \inf\{\lambda \geq 0 : T_n(\hat{\theta}^n_{\lambda,\mu^n_{\mathrm{CV}},\hat{\eta}}) \leq q(\alpha)\}$ by binary search;
  update $\gamma^n_{\mathrm{CV}} \leftarrow (1 + \lambda_n^\star) \cdot \gamma^n_{\mathrm{CV}}$;
  compute $\hat{\theta}^n_{\lambda_n^\star,\mu^n_{\mathrm{CV}},\hat{\eta}}$ using (8);
end
output: $\hat{f}^n_{\mathrm{NILE}} := f_{\hat{\eta}}(\,\cdot\,; \hat{\theta}^n_{\lambda_n^\star,\mu^n_{\mathrm{CV}},\hat{\eta}})$ defined by (9);

The penalty parameter γ^n_{CV} is chosen to minimize the out-of-sample mean squared error of the prediction model $\hat{\mathbf{Y}} = \mathbf{Q}_\gamma \mathbf{Y}$, which corresponds to the solution of (8) for $\lambda = 0$. After choosing $\lambda_n^\star$, the objective function in (8) increases by the term $\lambda_n^\star \|\mathbf{P}_{\delta^n_{\mathrm{CV}}}(\mathbf{Y} - \mathbf{B}\theta)\|_2^2$. In order for the penalty term $\gamma\theta^\top \mathbf{K}\theta$ to impose the same degree of smoothness in the altered optimization problem, the penalty parameter γ needs to be adjusted accordingly. The heuristic update in our algorithm is motivated by the simple observation that for all $\delta, \lambda \geq 0$, $\|\mathbf{Y} - \mathbf{B}\theta\|_2^2 + \lambda \|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2 \leq (1+\lambda)\|\mathbf{Y} - \mathbf{B}\theta\|_2^2$.

5.2.3 Asymptotic generalization (consistency)

We now prove consistency of our estimator in the case where the hyper-parameters (η, μ) are fixed (rather than data-driven), and the function class $\mathcal{F}_\eta$ is correctly specified. Fix any $a < b$ and a basis $B = (B_1, \ldots, B_k)$. Let $\eta_0 = (a, b, B)$ and let the model class be given by $\mathcal{M} = \mathcal{F}_{\eta_0} \times \mathcal{G} \times \mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{Q}$, where $\mathcal{F}_{\eta_0}$ is as described in Section 5.2. Assume that the data-generating model $M = (f_{\eta_0}(\,\cdot\,; \theta^0), g, h_1, h_2, Q) \in \mathcal{M}$ induces an observational distribution $\mathbb{P}_M$ such that $\mathrm{supp}^M(X) \subseteq (a, b)$. Let further $\mathcal{I}$ 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 η_0, rather than the data-driven values. Choose any $\delta, \gamma \geq 0$ and basis $C = (C_1, ..., C_k)$ and let $\mu = (\delta, \gamma, C)$. We will make use of the following assumptions.

(B1) $\forall \tilde{M} \in \mathcal{M}$ s.t. $\mathbb{P}_M = \mathbb{P}_{\tilde{M}}$: $\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[X^2]$, $\sup_{i\in\mathcal{I}} \lambda_{\max}(\mathbb{E}_{\tilde{M}(i)}[B(X)B(X)^\top]) < \infty$.

(B2) $\mathbb{E}_M[B(X)B(X)^\top]$, $\mathbb{E}_M[C(A)C(A)^\top]$ and $\mathbb{E}_M[C(A)B(X)^\top]$ are of full rank.

(C1) $T(\theta)$ has uniform asymptotic power on any compact set of alternatives.

(C2) $\lambda_n^\star := \inf\{\lambda \geq 0 : T_n(\hat{\theta}^n_{\lambda,\eta_0,\mu}) \leq q(\alpha)\}$ is almost surely finite.

(C3) $\lambda \mapsto T_n(\hat{\theta}^n_{\lambda,\eta_0,\mu})$ is weakly decreasing and $\theta \mapsto T_n(\theta)$ is continuous.

Assumptions (B1)–(B2) ensure consistency of the estimator as long as $\lambda_n^\star$ 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 $\lambda_n^\star$ in the data-driven fashion described in Section 5.2.2. In Assumption (B1), $\lambda_{\max}$ denotes the largest eigenvalue. In words, the assumption states that, under each model $\tilde{M} \in \mathcal{M}$ with $\mathbb{P}_M = \mathbb{P}_{\tilde{M}}$, 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 $\mathcal{I}$. 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 θ^0 is identified from the observational distribution in the sense that the hypothesis $H_0(\theta) : \theta = \theta^0$ becomes equivalent with $\tilde{H}_0(\theta) : \mathbb{E}_M[C(A)(Y - B(X)^\top\theta)] = 0$. (C1) means that for any compact set $K \subseteq \mathbb{R}^k$ with $\theta^0 \notin K$ 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 asympotic 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 $(\lambda_n^\star = \infty)$, ensuring that consistency still holds.

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

Proposition 5.1 (Asymptotic generalization). *Let $\mathcal{I}$ 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 $\tilde{M} \in \mathcal{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, and any $\varepsilon > 0$, it holds that*

$$\mathbb{P}_M\left(\Big| \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f_{\eta_0}(X; \hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu}))^2\big] - \inf_{f_\diamond\in\mathcal{F}_{\eta_0}} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f_\diamond(X))^2\big] \Big| \leq \varepsilon \right) \to 1,$$

as $n \to \infty$. In the above event, only $\hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu}$ 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 $\lambda_n^\star$, we use the test statistic T_n^2, which tests the slightly stronger hypothesis $\bar{H}_0$, see Appendix C. In all experiments use the significance level $\alpha = 0.05$. We include two other approaches as baseline: (i) the method NPREGIV (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 $\lambda^\star \equiv 0$). In all experiments, we generate data sets of size $n = 200$ as independent replications from

$$A := \varepsilon_A, \quad H := \varepsilon_H, \quad X := \alpha_A A + \alpha_H H + \alpha_\varepsilon \varepsilon_X, \quad Y := f(X) + 0.3H + 0.2\varepsilon_Y, \qquad (10)$$

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 $\alpha_A^2 + \alpha_H^2 + \alpha_\varepsilon^2 = 1$, 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% inner quantile range of X. By well-known properties of natural splines, any such function extends linearly outside the boundary knots. Figure 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. NPREGIV exploits A as an

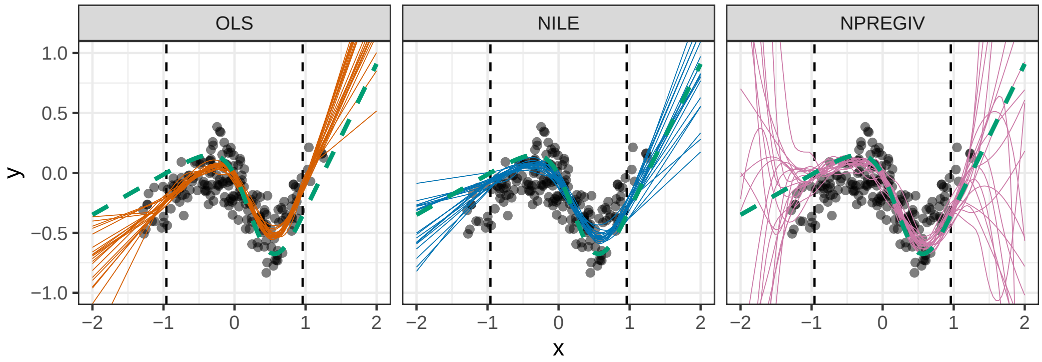


Figure 2: A sample dataset from the model (10) with $\alpha_A = \sqrt{1/3}$, $\alpha_H = \sqrt{2/3}$, $\alpha_\varepsilon = 0$. The true causal function is indicated by a green dashed line. For each method, we show 20 estimates of this function, each based on an independent sample from (10). For values within the support of the training data (vertical dashed lines mark the inner 90% quantile range), NPREGIV 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.

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 $(\alpha_A, \alpha_H, \alpha_\varepsilon)$, we construct several models $M_1, \dots, M_N$ of the form (10) by randomly sampling causal functions $f_1, \dots, f_N$ (see Appendix D for further details on the sampling procedure). For every $x \in [0, 2]$, let $\mathcal{I}_x$ 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 $\mathcal{I}_x$), i.e., for each estimator $\hat{f}$, we consider

$$\frac{1}{N}\sum_{j=1}^{N} \sup_{i \in \mathcal{I}_x} \mathbb{E}_{M_j(i)}\big[(Y - \hat{f}(X))^2\big] = \mathbb{E}[\xi_Y^2] + \frac{1}{N}\sum_{j=1}^{N} \sup_{\tilde{x} \in [-x,x]} (f_j(\tilde{x}) - \hat{f}(\tilde{x}))^2, \tag{11}$$

where $\xi_Y := 0.3H + 0.2\varepsilon_Y$ is the noise term for Y (which is fixed across all experiments). In practice, we evaluate the functions $\hat{f}$, $f_1, \dots, f_N$ on a fine grid on $[-x, x]$ to approximate the above supremum. Figure 3 plots the average worst-case risk versus intervention strength for different parameter settings. The optimal worst-case risk $\mathbb{E}[\xi_Y^2]$ 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 $\bar{H}_0(\hat{\theta}^n_{\lambda,\delta^n_{\mathrm{CV}},\gamma^n_{\mathrm{CV}}})$ is accepted already for small values of λ (in this experiment, the empirical average of $\lambda^\star_n$ 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 $\lambda^\star_n$ of 2.412 and 5.136, respectively). The method NPREGIV 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 α_H) might stem from the fact that, as the $\lambda^\star_n$ increases, also the smoothness penalty γ 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.

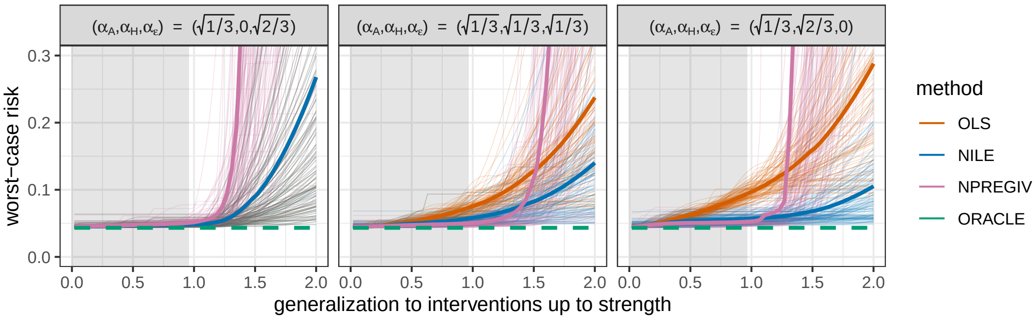


Figure 3: Predictive performance under confounding-removing interventions on X for different confounding- and intervention strengths (see alpha values in the grey panel on top). The right panel corresponds to the same parameter setting as in Figure 2. The plots in each panel are based on data sets of size $n = 200$, generated from $N = 100$ different models of the form (10). For each model, we draw a different function f, resulting in a different minimax solution (see Appendix D for details on the sampling procedure). 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 is equal to $\mathbb{E}[\xi_Y^2]$ (green dashed line). The grey area indicates the inner 90 % quantile range of X in the training distribution; the white area can be seen as an area of generalization.

In the case, where all exogenous noise comes from the unobserved variable ε_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 method seems robust to such scenarios. As the instrument strength increases, the NILE clearly outperforms OLS and NPREGIV for interventions on X which include values outside the training data.

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 minmizer 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

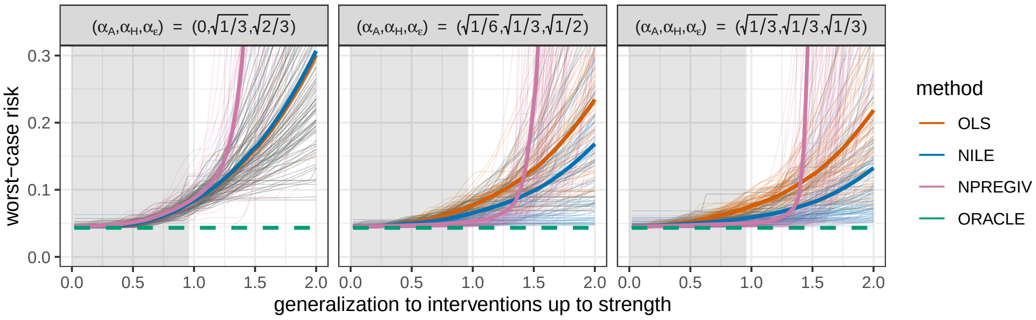


Figure 4: Predictive performance for varying instrument strength. If the instruments have no influence on X ($\alpha_A = 0$), the second term in the objective function (8) is effectively constant in θ, and the NILE therefore coincides with the OLS estimator (which uses $\lambda = 0$). This guards the NILE against the large variance which most IV estimators suffer from in a weak instrument setting. For increasing influence of A, it clearly outperforms both alternative methods for large intervention strengths.

investigate the sharpness of the bounds we provide in Section 4.3.2 and other extrapolation assumptions on $\mathcal{F}$. 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. This could yield an even better performance on estimating the causal function within the support of the covariates. 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 unlabelled 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 under which assumptions it is possible to generalize knowledge that was acquired during training to a different test distribution.

A Transforming causal models

As illustrated in Remark 1, our framework is able to model cases where causal relations between the observed variables are given explicitly, e.g., by an SCM. The key insight is that most of these causal relations can be absorbed by the hidden confounding H on which we make few restrictions. To show how this can be done in a general setting, let us consider the following SCM

$$\begin{aligned} A &:= \varepsilon_A & X &:= w(X,Y) + g(A) + h_2(H, \varepsilon_X) \\ H &:= \varepsilon_H & Y &:= f(X) + h_1(H, \varepsilon_Y). \end{aligned} \tag{12}$$

Assume that this SCM is uniquely solvable in the sense that there exists a unique function F such that $(A, H, X, Y) = F(\varepsilon_A, \varepsilon_H, \varepsilon_X, \varepsilon_Y)$ almost surely, see [13] for more details. Denote by F_X the coordinates of F that correspond to the X variable (i.e., the coordinates from

$r+q+1$ to $r+q+d$). Assume further that there exist functions $\tilde{g}$ and $\tilde{h}_2$ such that

$$F_X(\varepsilon_A, \varepsilon_H, \varepsilon_X, \varepsilon_Y) = \tilde{g}(\varepsilon_A) + \tilde{h}_2((\varepsilon_H, \varepsilon_Y), \varepsilon_X). \tag{13}$$

This decomposition is not always possible, but it exists in the following settings, for example: (i) *There are no A variables.* As discussed in Section 2.1 our framework also works if no A variables exist. In these cases, the additive decomposition (13) becomes trivial. (ii) *There are further constraints on the full SCM.* The additive decomposition (13) holds if, for example, w is a linear function or A only enters the structural assignments of covariates X which have at most Y as a descendant.

Using the decomposition in (13), we can define the following SCM

$$\begin{aligned} A &:= \varepsilon_A & \qquad X &:= \tilde{g}(A) + \tilde{h}_2(\tilde{H}, \varepsilon_X) \\ \tilde{H} &:= \varepsilon_{\tilde{H}} & \qquad Y &:= f(X) + h_1(\tilde{H}), \end{aligned} \tag{14}$$

where $\varepsilon_{\tilde{H}}$ has the same distribution as $(\varepsilon_H, \varepsilon_Y)$ in the previous model. This model fits the framework described in Section 2.1, where the noise term in Y is now taken to be constantly zero. Both SCMs (12) and (14) induce the same observational distribution and the same function f appears in the assignments of Y.

It is further possible to express the set of interventions on the covariates X in the original SCM (12) as a set of interventions on the covariates in the reduced SCM (14). The description of a class of interventions in the full SCM (12) may, however, become more complex if we consider them in the reduced SCM (14). In particular, to apply the developed methodology, one needs to check whether the interventions in the reduced SCM is a well-behaved set of interventions (this is not necessarily the case) and how the support of all X variables behaves under that specific intervention. We now discuss the case that the causal graph induced by the full SCM is a directed acyclic graph (DAG).

Intervention type. First, we consider which types of interventions in (12) translate to well-behaved interventions in (14). Importantly, interventions on A in the full SCM reduce to regular interventions A also in the reduced SCM. Similarly, performing hard interventions on all components of X in the full SCM leads to the same intervention in the reduced SCM, which is in particular both confounding-removing and confounding-preserving. For interventions on subsets of the X, this is not always the case. To see that, consider the following example.

$$\begin{aligned} A &:= \varepsilon_A \\ X_1 &:= \varepsilon_1, \quad X_2 := Y + \varepsilon_2 \\ Y &:= X_1 + \varepsilon_Y \end{aligned} \tag{15}$$

$$\begin{aligned} A &:= \varepsilon_A, \quad H := \varepsilon_Y \\ X &:= (\varepsilon_1, H + \varepsilon_1 + \varepsilon_2) \\ Y &:= X_1 + H \end{aligned} \tag{16}$$

with $\varepsilon_A, \varepsilon_1, \varepsilon_2, \varepsilon_Y \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, where (15) represents the full SCM and (16) corresponds to the reduced SCM using our framework. Consider now, in the full SCM, the intervention $X_1 := i$, for some $i \in \mathbb{R}$. In the reduced SCM, this intervention corresponds to the intervention $X = (X_1, X_2) := (i, H + i + \varepsilon_2)$, which is neither confounding-preserving nor confounding-removing.[5] On the other hand, any intervention on X_2 or A in the full SCM model corresponds to the same intervention in the reduced SCM. We can generalize these observations to the following statements:

- *Interventions on A:* If we intervene on A in the full SCM (12) (i.e., by replacing the structural assignment of A with $\psi^i(I^i, \varepsilon^i_A)$), then this translates to an equivalent intervention in the reduced SCM (14).
- *Hard interventions on all X:* If we intervene on all X in the full SCM (12) by replacing the structural assignment of X with an independent random variable $I \in \mathbb{R}^d$, then this

[5] This may not come as a surprise since without the help of an instrument, it is impossible to distinguish whether a covariate is an ancestor or a descendant of Y.

translates to the same intervention in the reduced SCM (14) which is confounding-removing.

- *No X is a descendant of Y and there is no unobserved confounding H:* If we intervene on X in the full SCM (12) (i.e., by replacing the structural assignment of X with $\psi^i(g, A^i, \varepsilon_X^i, I^i)$), then this translates to a potentially different but confounding-removing intervention in the reduced SCM (14). This is because the reduced SCM (14) does not include unobserved variables H in this case.

- *Hard interventions on a variable X_j which has at most Y as a descendant:* If we intervene on X_j in the full SCM (12) by replacing the structural assignment of X_j with an independent random variable I, then this intervention translates to a potentially different but confounding-preserving intervention.

Other settings may yield well-behaved interventions, too, but may require more assumptions on the full SCM model (12) or further restrictions on the intervention classes.

Intervention support. A support-reducing intervention in the full SCM can translate to a support-extending intervention in the reduced SCM. Consider the following example.

$$\begin{aligned} X_1 &\coloneqq \varepsilon_1 \\ X_2 &\coloneqq X_1 + \mathbf{1}\{X_1 = 0.5\} \\ Y &\coloneqq X_2 + \varepsilon_Y \end{aligned} \tag{17}$$

$$\begin{aligned} X &\coloneqq (\varepsilon_1, \varepsilon_1 + \mathbf{1}\{\varepsilon_1 = 0.5\}) \\ Y &\coloneqq X_2 + \varepsilon_Y, \end{aligned} \tag{18}$$

with $\varepsilon_1, \varepsilon_Y \overset{i.i.d.}{\sim} \mathcal{U}(0,1)$. As before, (17) represents the full SCM, whereas (18) corresponds to the reduced SCM converted to fit our framework. Under the observational distribution, the support of X_1 and X_2 is equal to the open interval $(0,1)$. Consider now the support-reducing intervention $X_1 := 0.5$ in (17). Within our framework, such an intervention would correspond to the intervention $X = (X_1, X_2) := (0.5, 1.5)$, which is support-extending. This example is rather special in that the SCM consists of a function that changes on a null set of the observational distribution. With appropriate assumptions to exclude similar degenerate cases, it is possible to show that support-reducing interventions in (12) correspond to support-reducing interventions within our framework (14).

B Sufficient conditions for Assumption 1 in IV settings

Assumption 1 states that f is identified on the support of X from the observational distribution of (Y, X, A). Whether this assumption is satisfied depends on the structure of $\mathcal{F}$ but also on the other function classes $\mathcal{G}, \mathcal{H}_1, \mathcal{H}_2$ and $\mathcal{Q}$ that make up the model class $\mathcal{M}$ from which we assume that the distribution of (Y, X, A) is generated.

Identifiability of the causal function in the presence of instrumental variables is a well-studied problem in econometrics literature. Most prominent is the literature on identification in linear SCMs [e.g., 25, 27]. However, identification has also been studied for various other parametric function classes. We say that $\mathcal{F}$ is a parametric function class if it can be parametrized by some finite dimensional parameter set $\Theta \subseteq \mathbb{R}^p$. We here consider classes of the form

$$\mathcal{F} := \{f(\cdot, \theta) : \mathbb{R}^d \to \mathbb{R} \,|\, \theta : \Theta \to \mathbb{R}, \theta \mapsto f(x, \theta) \text{ is } C^2 \text{ for all } x \in \mathbb{R}^d\}.$$

Consistent estimation of the parameter θ_0 using instrumental variables in such function classes has been studied extensively in the econometric literature [e.g., 3, 36, 37]. These works also contain rigorous results on how instrumental variable estimators of θ_0 are constructed and under which conditions consistency (and thus identifiability) holds. Here, we give an argument on why the presence of the exogenous variables A yields identifiability under certain regularity conditions. Assume that $\mathbb{E}[h_1(H, \varepsilon_Y)|A] = 0$, which implies that the true causal function $f(\cdot, \theta_0)$ satisfies the population orthogonality condition

$$\mathbb{E}[l(A)^\top (Y - f(X, \theta_0))] = \mathbb{E}\big[l(A)^\top \mathbb{E}[h_1(H, \varepsilon_Y)|A]\big] = 0, \tag{19}$$

for some measurable mapping $l : \mathbb{R}^q \to \mathbb{R}^g$, for some $g \in \mathbb{N}_{>0}$. Clearly, θ_0 is identified from the observational distribution if the map $\theta \mapsto \mathbb{E}[l(A)^\top (Y - f(X, \theta))]$ is zero if and only if $\theta = \theta_0$. Furthermore, since $\theta \mapsto f(x, \theta)$ is differentiable for all $x \in \mathbb{R}^d$, the mean value theorem yields that, for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$, there exists an intermediate point $\tilde{\theta}(x, \theta, \theta_0)$ on the line segment between θ and θ_0 such that

$$f(x, \theta) - f(x, \theta_0) = D_\theta f(x, \tilde{\theta}(x, \theta, \theta_0))(\theta - \theta_0),$$

where, for each $x \in \mathbb{R}^d$, $D_\theta f(x, \theta) \in \mathbb{R}^{1 \times p}$ is the derivative of $\theta \mapsto f(x, \theta)$ evaluated in θ. Composing the above expression with the random vector X, multiplying with $l(A)$ and taking expectations yields that

$$\mathbb{E}[l(A)(Y - f(X, \theta_0))] - \mathbb{E}[l(A)(Y - f(X, \theta))] = \mathbb{E}[l(A) D_\theta f(X, \tilde{\theta}(X, \theta, \theta_0))](\theta_0 - \theta).$$

Hence, if $\mathbb{E}[l(A) D_\theta f(X, \tilde{\theta}(X, \theta, \theta_0))] \in \mathbb{R}^{g \times p}$ is of rank p for all $\theta \in \Theta$ (which implies $g \geq p$), then θ_0 is identifiable as it is the only parameter that satisfies the population orthogonality condition of (19). As θ_0 uniquely determines the entire function, we get identifiability of $f \equiv f(\cdot, \theta_0)$, not only on the support of X but the entire domain $\mathbb{R}^d$, i.e., both Assumptions 1 and 2 are satisfied. In the case that $\theta \mapsto f(x, \theta)$ is linear, i.e. $f(x, \theta) = f(x)^T \theta$ for all $x \in \mathbb{R}^d$, the above rank condition reduces to $\mathbb{E}[l(A) f(X)^T] \in \mathbb{R}^{g \times p}$ having rank p (again, implying that $g \geq p$). Furthermore, when $(x, \theta) \mapsto f(x, \theta)$ is bilinear, a reparametrization of the parameter space ensures that $f(x, \theta) = x^T \theta$ for $\theta \in \Theta \subseteq \mathbb{R}^d$. In this case, the rank condition can be reduced to the well-known rank condition for identification in a linear SCM, namely that $\mathbb{E}[AX^T] \in \mathbb{R}^{q \times p}$ is of rank p.

Finally, identifiability and methods of consistent estimation of the causal function have also been studied for non-parametric function classes. The conditions for identification are rather technical, however, and we refer the reader to [46, 47] for further details.

C Choice of test statistic

By considering the variables $B(X) = (B_1(X), \ldots, B_k(X))$ and $C(A) = (C_1(A), \ldots, C_k(A))$ as vectors of covariates and instruments, respectively, our setting in Section 5.2 reduces to the classical (just-identified) linear IV setting. We could therefore use a test statistics similar to the one propsed by the PULSE [34]. With a notation that is slightly adapted to our setting, this estimator tests $\tilde{H}_0(\theta)$ using the test statistic

$$T_n^1(\theta) = c(n) \frac{\|\mathbf{P}(\mathbf{Y} - \mathbf{B}\theta)\|_2^2}{\|\mathbf{Y} - \mathbf{B}\theta\|_2^2},$$

where $\mathbf{P}$ is the projection onto the columns of $\mathbf{C}$, and $c(n)$ is some function with $c(n) \sim n$ as $n \to \infty$. Under the null hypothesis, T_n^1 converges in distribution to the χ_k^2 distribution, and diverges to infinity in probability under the general alternative. Using this test statistic, $\tilde{H}_0(\theta)$ is rejected if and only if $T_n^1(\theta) > q(\alpha)$, where $q(\alpha)$ is the $(1 - \alpha)$-quantile of the χ_k^2 distribution. The acceptance region of this test statistic is asymptotically equivalent with the confidence region of the Anderson-Rubin test [4] for the causal parameter θ^0. Using the above test results in a consistent estimator for θ^0 [34, Theorem 3.12]; the proof exploits the particular form of T_n^1 without explicitly imposing that assumptions (C1) and (C2) hold.

If the number k of basis functions is large, however, numerical experiments suggest that the above test has low power in finite sample settings. As default, our algorithm therefore uses a different test based on a penalized regression approach. This test has been proposed in [16] for inference in nonparametric regression models. We now introduce this procedure with a notation that is adapted to our setting. For every $\theta \in \mathbb{R}^k$, let $R_\theta = Y - B(X)^\top \theta$ be the residual associated with θ. We then test the slightly stronger hypothesis

$$\bar{H}_0(\theta): \text{ there exists } \sigma_\theta^2 > 0 \text{ such that } \mathbb{E}[R_\theta \mid A] \overset{\text{a.s.}}{=} 0 \text{ and } \mathrm{Var}[R_\theta \mid A] = \sigma_\theta^2$$

against the alternative that $\mathbb{E}[R_\theta \mid A] = m(A)$ for some smooth function m. To see that the above hypothesis implies $\tilde{H}_0(\theta)$ (and therefore $H_0(\theta)$, see Section 5.2.1), let $\theta \in \mathbb{R}^k$ be such that $\bar{H}_0(\theta)$ holds true. Then,

$$\mathbb{E}[C(A)(Y - B(X)^\top \theta)] = \mathbb{E}[C(A)R_\theta] = \mathbb{E}[\mathbb{E}[C(A)R_\theta \mid A]] = \mathbb{E}[C(A)\mathbb{E}[R_\theta \mid A]] = 0,$$

showing that also $\tilde{H}_0(\theta)$ holds true. Thus, if $\tilde{H}_0(\theta)$ is false, then also $\bar{H}_0(\theta)$ is false. As a test statistic $T_n^2(\theta)$ for $\bar{H}_0(\theta)$, we use (up to a normalization) the squared norm of a penalized regression estimate of m, evaluated at the data $\mathbf{A}$, i.e., the TSLS loss $\|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2$. In the fixed design case, where $\mathbf{A}$ is non-random, it has been shown that, under $\bar{H}_0(\theta)$ and certain additional regularity conditions, it holds that

$$\frac{\|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2 - \sigma_\theta^2 c_n}{\sigma_\theta^2 d_n} \xrightarrow{\text{d}} \mathcal{N}(0,1),$$

where c_n and d_n are known functions of $\mathbf{C}$, $\mathbf{M}$ and δ [16, Theorem 1]. The authors further state that the above convergence is unaffected by exchanging σ_θ^2 with a consistent estimator $\hat{\sigma}_\theta^2$, which motivates our use of the test statistic

$$T_n^2(\theta) := \frac{\|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2 - \hat{\sigma}_{\theta,n}^2 c_n}{\hat{\sigma}_{\theta,n}^2 d_n},$$

where $\hat{\sigma}_{\theta,n}^2 := \frac{1}{n-1}\sum_{i=1}^n \|(\mathbf{I}_n - \mathbf{P}_\delta)(\mathbf{Y} - \mathbf{B}\theta)\|_2^2$. As a rejection threshold $q(\alpha)$ we use the $1-\alpha$ quantile of a standard normal distribution. For results on the asymptotic power of the test defined by T^2, we refer to Section 2.3 in [16].

In our software package, both of the above tests are available options.

D Addition to experiments

D.1 Sampling of the causal function

To ensure linear extrapolation of the causal function, we have chosen a function class consisting of natural cubic splines, which, by construction, extrapolate linearly outside the boundary knots. We now describe in detail how we sample functions from this class for the experiments in Section 5.2.4. Let $q_{\min}$ and $q_{\max}$ be the respective 5%- and 95% quantiles of X, and let $B_1, \dots, B_4$ be a basis of natural cubic splines corresponding to 5 knots placed equidistantly between $q_{\min}$ and $q_{\max}$. We then sample coefficients $\beta_i \overset{\text{iid}}{\sim} \text{Uniform}(-1,1)$, $i = 1, \dots, 4$, and construct f as $f = \sum_{i=1}^4 \beta_i B_i$. For illustration, we have included 18 realizations in Figure 5.

D.2 Violations of the linear extrapolation assumption

We have assumed that the true causal function extrapolates linearly outside the 90% quantile range of X. We now investigate the performance of our method for violations of this assumption. To do so, we again sample from the model (10), with $\alpha_A = \alpha_H = \alpha_\varepsilon = 1/\sqrt{3}$. For each data set, the causal function is sampled as follows. Let $q_{\min}$ and $q_{\max}$ be the 5%- and 95% quantiles of X. We first generate a function $\tilde{f}$ that linearly extrapolates outside $[q_{\min}, q_{\max}]$ as described in Section D.1. For a given threshold κ, we then draw $k_1, k_2 \overset{\text{iid}}{\sim} \text{Uniform}(-\kappa, \kappa)$ and construct f for every $x \in \mathbb{R}$ by

$$f(x) = \tilde{f}(x) + \frac{1}{2}k_1((x - q_{\min})_-)^2 + \frac{1}{2}k_2((x - q_{\max})_+)^2,$$

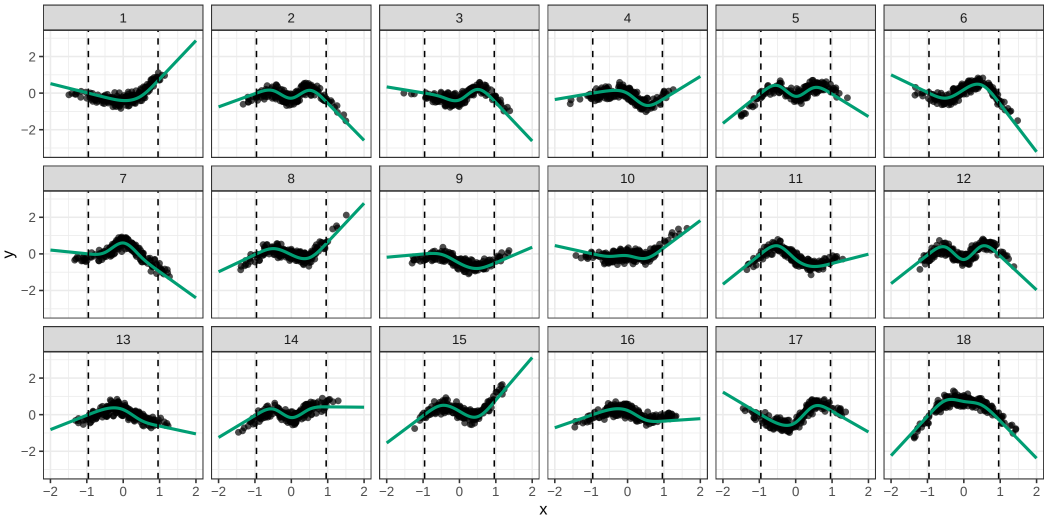


Figure 5: The plots show independent realizations of the causal function that is used in all our experiments. These are sampled from a linear space of natural cubic splines, as described in Appendix D.1. To ensure a fair comparison with the alternative method, NPREGIV, the true causal function is chosen from a model class different from the one assumed by the NILE.

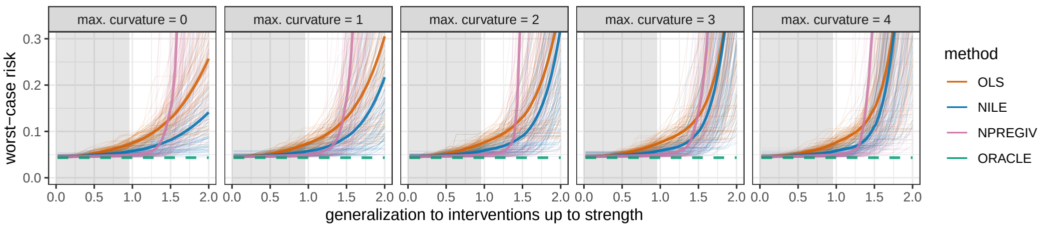


Figure 6: Worst-case risk for increasingly strong violations of the linear extrapolation assumption. The grey area marks the inner 90 % quantile range of X in the training distribution. As the curvature of f outside the domain of the observed data increases, it becomes difficult to predict the interventional behavior of Y for strong interventions. However, even in situations where the linear extrapolation assumption is strongly violated, it remains beneficial to extrapolate linearly.

such that the curvature of f on $(-\infty, q_{\text{min}}]$ and $[q_{\text{max}}, \infty)$ is k_1 and k_2, respectively. Figure 6 shows results for $\kappa = 0, 1, 2, 3, 4$. As the curvature increases, the ability to generalize decreases.

E Proofs

E.1 Proof of Proposition 3.1

Proof. Assume that $\mathcal{I}$ is a set of interventions on X with at least one confounding-removing intervention. Let $i \in \mathcal{I}$ and $f_\diamond \in \mathcal{F}$, then we have the following expansion

$$\mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] = \mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] + \mathbb{E}_{M(i)}[\xi_Y^2] + 2\mathbb{E}_{M(i)}[\xi_Y(f(X) - f_\diamond(X))], \quad (20)$$

where $\xi_Y = h_1(H, \varepsilon_Y)$. For any intervention $i \in \mathcal{I}$ the causal function f always yields an identical loss. In particular, it holds that

$$\sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - f(X))^2] = \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[\xi_Y^2] = \mathbb{E}_M[\xi_Y^2], \tag{21}$$

where we used that the distribution of ξ_Y is not affected by an intervention on X. The loss of the causal function can never be better than the minimax loss, that is,

$$\inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] \le \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - f(X))^2] = \mathbb{E}_M[\xi_Y^2]. \tag{22}$$

In other words, the minimax solution (if it exists) is always better than or equal to the causal function. We will now show that when $\mathcal{I}$ contains at least one confounding-removing intervention, then the minimax loss is dominated by any such intervention.

Fix $i_0 \in \mathcal{I}$ to be a confounding-removing intervention and let (X, Y, H, A) be generated by the SCM $M(i_0)$. Recall that there exists a map ψ^{i_0} such that $X := \psi^{i_0}(g, h_2, A, H, \varepsilon_X, I^{i_0})$ and that $X \perp\!\!\!\perp H$ as i_0 is a confounding-removing intervention. Furthermore, since the vectors A, H, ε_X, ε_Y and I^{i_0} are mutually independent, we have that $(X, H) \perp\!\!\!\perp \varepsilon_Y$ which together with $X \perp\!\!\!\perp H$ implies X, H and ε_Y are mutually independent, and hence $X \perp\!\!\!\perp h_1(H, \varepsilon_Y)$. Using this independence we get that $\mathbb{E}_{M(i_0)}[\xi_Y(f(X) - f_\diamond(X))] = \mathbb{E}_M[\xi_Y]\mathbb{E}_{M(i_0)}[(f(X) - f_\diamond(X))]$. Hence, (20) for the intervention i_0 together with the modeling assumption $\mathbb{E}_M[\xi_Y] = 0$ implies that for all $f_\diamond \in \mathcal{F}$,

$$\mathbb{E}_{M(i_0)}[(Y - f_\diamond(X))^2] = \mathbb{E}_{M(i_0)}[(f(X) - f_\diamond(X))^2] + \mathbb{E}_M[\xi_Y^2] \ge \mathbb{E}_M[\xi_Y^2].$$

This proves that the smallest loss at a confounding-removing intervention is achieved by the causal function. Denoting the non-empty subset of confounding-removing interventions by $\mathcal{I}_{\mathrm{cr}} \subseteq \mathcal{I}$, this implies

$$\begin{aligned}\inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] &\ge \inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}_{\mathrm{cr}}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] \\ &\ge \inf_{f_\diamond\in\mathcal{F}} \mathbb{E}_{M(i_0)}[(Y - f_\diamond(X))^2] \\ &= \mathbb{E}_M[\xi_Y^2].\end{aligned} \tag{23}$$

Combining (22) and (23) it immediately follows that

$$\inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] = \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - f(X))^2],$$

and hence

$$f \in \operatorname*{argmin}_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2],$$

which completes the proof of Proposition 3.1. □

E.2 Proof of Proposition 3.2

Proof. Let $\mathcal{F}$ be the class of all linear functions and let $\mathcal{I}$ denote the set of interventions on X that satisfy

$$\sup_{i\in\mathcal{I}} \lambda_{\min}\big(\mathbb{E}_{M(i)}\big[XX^\top\big]\big) = \infty.$$

We claim that the causal function $f(x) = b^\top x$ is the unique minimax solution of (1). We prove the result by contradiction. Let $\bar{f} \in \mathcal{F}$ (with $\bar{f}(x) = \bar{b}^\top x$) be such that

$$\sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - \bar{b}^\top X)^2] \le \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}[(Y - b^\top X)^2],$$

and assume that $\|\bar{b} - b\|_2 > 0$. For a fixed $i \in \mathcal{I}$, we get the following bound

$$\mathbb{E}_{M(i)}[(b^\top X - \bar{b}^\top X)^2] = (b - \bar{b})^\top \mathbb{E}_{M(i)}[XX^\top](b - \bar{b}) \ge \lambda_{\min}(\mathbb{E}_{M(i)}[XX^\top])\|b - \bar{b}\|_2^2.$$

Since we assumed that the minimal eigenvalue is unbounded, this means that we can choose $i \in \mathcal{I}$ such that $\mathbb{E}_{M(i)}[(b^\top X - \bar{b}^\top X)^2]$ can be arbitrarily large. However, applying Proposition 3.3, this leads to a contradiction since $\sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}[(b^\top X - \bar{b}^\top X)^2] \leq 4 \operatorname{Var}_M(\xi_Y)$ cannot be satisfied. Therefore, it must holds that $\bar{b} = b$, which moreover implies that f is indeed a solution to the minimax problem $\operatorname{argmin}_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2]$, as it achieves the lowest possible objective value. This completes the proof of Proposition 3.2. $\square$

E.3 Proof of Proposition 3.3

Proof. Let $\mathcal{I}$ be a set of interventions on X or A and let $f_\diamond \in \mathcal{F}$ with

$$\sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] \leq \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}[(Y - f(X))^2]. \tag{24}$$

For any $i \in \mathcal{I}$, the Cauchy-Schwartz inequality implies that

$$\begin{aligned}
&\mathbb{E}_{M(i)}[(Y - f_\diamond(X))^2] = \mathbb{E}_{M(i)}[(f(X) + \xi_Y - f_\diamond(X))^2] \\
&\quad = \mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] + \mathbb{E}_{M(i)}[\xi_Y^2] + 2\mathbb{E}_{M(i)}[\xi_Y(f(X) - f_\diamond(X))] \\
&\quad \geq \mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] + \mathbb{E}_{M}[\xi_Y^2] - 2\left(\mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2]\mathbb{E}_M[\xi_Y^2]\right)^{\frac{1}{2}}.
\end{aligned}$$

A similar computation shows that the causal function f satisfies

$$\mathbb{E}_{M(i)}[(Y - f(X))^2] = \mathbb{E}_M[\xi_Y^2].$$

So by condition (24) this implies for any $i \in \mathcal{I}$ that

$$\mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] + \mathbb{E}_M[\xi_Y^2] - 2\left(\mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2]\mathbb{E}_M[\xi_Y^2]\right)^{\frac{1}{2}} \leq \mathbb{E}_M[\xi_Y^2],$$

which is equivalent to

$$\begin{aligned}
&\mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] \leq 2\sqrt{\mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2]\mathbb{E}_M[\xi_Y^2]} \\
\iff \quad &\mathbb{E}_{M(i)}[(f(X) - f_\diamond(X))^2] \leq 4\mathbb{E}_M[\xi_Y^2].
\end{aligned}$$

As this inequality holds for all $i \in \mathcal{I}$, we can take the supremum over all $i \in \mathcal{I}$, which completes the proof of Proposition 3.3. $\square$

E.4 Proof of Proposition 3.4

Proof. As argued before, we have that for all $i \in \mathcal{I}_1$,

$$\mathbb{E}_{M(i)}\left[(Y - f(X))^2\right] = \mathbb{E}_{M(i)}\left[\xi_Y^2\right] = \mathbb{E}_M\left[\xi_Y^2\right].$$

Let now $f_1^* \in \mathcal{F}$ be a minimax solution w.r.t. $\mathcal{I}_1$. Then, using that the causal function f lies in $\mathcal{F}$, it holds that

$$\sup_{i \in \mathcal{I}_1} \mathbb{E}_{M(i)}\left[(Y - f_1^*(X))^2\right] \leq \sup_{i \in \mathcal{I}_1} \mathbb{E}_{M(i)}\left[(Y - f(X))^2\right] = \mathbb{E}_M\left[\xi_Y^2\right].$$

Moreover, if $\mathcal{I}_2 \subseteq \mathcal{I}_1$, then it must also hold that

$$\sup_{i \in \mathcal{I}_2} \mathbb{E}_{M(i)}\left[(Y - f_1^*(X))^2\right] \leq \mathbb{E}_M\left[\xi_Y^2\right] = \sup_{i \in \mathcal{I}_2} \mathbb{E}_{M(i)}\left[(Y - f(X))^2\right].$$

To prove the second part, we give a one-dimensional example. Let $\mathcal{F}$ be linear (i.e., $f(x) = bx$) and let $\mathcal{I}_1$ consist of shift interventions on X of the form

$$X^i := g(A^i) + h_2(H^i, \varepsilon_X^i) + c,$$

with $c \in [0, K]$. Then, the minimax solution f_1^* (where $f_1^*(x) = b_1^* x$) with respect to $\mathcal{I}_1$ is not equal to the causal function f as long as $\mathrm{Cov}(X, \xi_Y)$ is strictly positive. This can be seen by explicitly computing the OLS estimator for a fixed shift c and observing that the worst-case risk is attained at $c = K$. Now let $\mathcal{I}_2$ be a set of interventions of the same form as $\mathcal{I}_1$ but including shifts with $c > K$ such that $\mathcal{I}_2 \not\subseteq \mathcal{I}_1$. Since $\mathcal{F}$ consists of linear functions, we know that the loss $\mathbb{E}_{M(i)}\big[(Y - f_1^*(X))^2\big]$ can become arbitrarily large, since

$$
\begin{aligned}
&\mathbb{E}_{M(i)}\big[(Y - f_1^*(X))^2\big] \\
&\quad = (b - b_1^*)^2 \mathbb{E}_{M(i)}[X^2] + \mathbb{E}_M[\xi_Y^2] + 2(b - b_1^*)\mathbb{E}_{M(i)}[\xi_Y X] \\
&\quad = (b - b_1^*)^2 (c^2 + \mathbb{E}_M[X^2] + 2c\mathbb{E}_M[X]) + \mathbb{E}_M[\xi_Y^2] + 2(b - b_1^*)(\mathbb{E}_M[\xi_Y X] + \mathbb{E}_M[\xi_Y] c),
\end{aligned}
$$

and $(b - b^*)^2 > 0$. In contrast, the loss for the causal function is always $\mathbb{E}_M[\xi_Y^2]$, so the worst-case risk of f_1^* becomes arbitrarily worse than that of f. This completes the proof of Proposition 3.4. $\square$

E.5 Proof of Proposition 4.1

Proof. Let $\varepsilon > 0$. By definition of the infimum, we can find $f^* \in \mathcal{F}$ such that

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

Let now $\tilde{M} \in \mathcal{M}$ be s.t. $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$. By assumption, the left-hand side of the above inequality is unaffected by substituting M for $\tilde{M}$, and the result thus follows. $\square$

E.6 Proof of Proposition 4.2

Proof. We first show that the causal parameter β is not a minimax solution. Let $u := \sup \mathcal{I} < \infty$, since $\mathcal{I}$ is bounded, and take $b = \beta + 1/(\sigma u)$. By an explicit computation we get that

$$
\begin{aligned}
\inf_{b_\diamond \in \mathbb{R}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - b_\diamond X)^2\big] &\leq \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - bX)^2\big] = \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\big[(\varepsilon_Y + \tfrac{1}{\sigma} H - \tfrac{1}{\sigma u} i H)^2\big] \\
&= \sup_{i \in \mathcal{I}} \Big[1 + \big(1 - \tfrac{i}{u}\big)^2\Big] < 2 = \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - \beta X)^2\big],
\end{aligned}
$$

where the last inequality holds because $0 < 1 + (1 - i/u)^2 < 2$ for all $i \in \mathcal{I}$, and since $\mathcal{I} \subseteq \mathbb{R}_{>0}$ is compact with upper bound u. Hence,

$$
\sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - \beta X)^2\big] - \inf_{b_\diamond \in \mathbb{R}} \sup_{i \in \mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - b_\diamond X)^2\big] > 0,
$$

proving that the causal parameter is not a minimax solution for model M w.r.t. $(\mathcal{F}, \mathcal{I})$. Recall that in order to prove that $(\mathbb{P}_M, \mathcal{M})$ does not generalize with respect to $\mathcal{I}$ we have to show that there exists an $\varepsilon > 0$ such that for all $b \in \mathbb{R}$ it holds that

$$
\sup_{\tilde{M}: \mathbb{P}_{\tilde{M}} = \mathbb{P}_M} \Big| \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - bX)^2\big] - \inf_{b_\diamond \in \mathbb{R}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - b_\diamond X)^2\big] \Big| \geq \varepsilon.
$$

Thus, it remains to show that for all $b \neq \beta$ there exists a model $\tilde{M} \in \mathcal{M}$ with $\mathbb{P}_M = \mathbb{P}_{\tilde{M}}$ such that the generalization loss is bounded below uniformly by a positive constant. We will show the stronger statement that for any $b \neq \beta$, there exists a model $\tilde{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, such that under $\tilde{M}$, b results in arbitrarily large generalization error. Let $c > 0$ and $i_0 \in \mathcal{I}$. Define

$$
\tilde{\sigma} := \frac{\operatorname{sign}\left((\beta - b) i_0\right) \sqrt{1 + c} - 1}{(\beta - b) i_0} > 0,
$$

and let $\tilde{M} := M(\gamma, \beta, \tilde{\sigma}, Q)$. By construction of the model class $\mathcal{M}$, it holds that $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$. Furthermore, by an explicit computation we get that

$$\begin{aligned}
\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - bX)^2\big] &\geq \mathbb{E}_{\tilde{M}(i_0)}\big[(Y - bX)^2\big] = \mathbb{E}_{\tilde{M}(i_0)}\big[((\beta - b)i_0 H + \varepsilon_Y + \tfrac{1}{\tilde{\sigma}}H)^2\big] \\
&= \mathbb{E}_{\tilde{M}(i_0)}\big[([(\beta - b)i_0\tilde{\sigma} + 1]\varepsilon_H + \varepsilon_Y)^2\big] = [(\beta - b)i_0\tilde{\sigma} + 1]^2 + 1 \\
&= ((\beta - b)i_0\tilde{\sigma})^2 + 2(\beta - b)i_0\tilde{\sigma} + 2 \\
&= (\operatorname{sign}((\beta - b)i_0)\sqrt{1+c} - 1)^2 + 2\operatorname{sign}((\beta - b)i_0)\sqrt{1+c} \\
&= c + 2.
\end{aligned} \tag{25}$$

Finally, by definition of the infimum, it holds that

$$\inf_{b_\diamond\in\mathbb{R}} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - b_\diamond X)^2\big] \leq \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - \beta X)^2\big] = 2. \tag{26}$$

Combining (25) and (26) yields that the generalization error is bounded below by c. That is,

$$\big|\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - bX)^2\big] - \inf_{b_\diamond\in\mathbb{R}} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - b_\diamond X)^2\big]\big| \geq c.$$

The above results make no assumptions on γ, and hold true, in particular, if $\gamma \neq 0$ (in which case Assumption 1 is satisfied, see Appendix B). This completes the proof of Proposition 4.2. □

E.7 Proof of Proposition 4.3

Proof. Let $\mathcal{I}$ be a well-behaved set of interventions on X. We consider two cases; (A) all interventions in $\mathcal{I}$ are confounding-preserving and (B) there is at least one intervention in $\mathcal{I}$ that is confounding-removing.

Case (A): In this case, we prove the result in two steps: (i) We show that (A, ξ_X, ξ_Y) is identified from the observational distribution $\mathbb{P}_M$. (ii) We show that this implies that the intervention distributions (X^i, Y^i), $i \in \mathcal{I}$, are also identified from the observational distribution, and conclude by using Proposition 4.1. Some of the details will be slightly technical because we allow for a large class of distributions (e.g., there is no assumption on the existence of densities).

We begin with step (i). In this case, $\mathcal{I}$ is a set of confounding-preserving interventions on X, and we have that $\operatorname{supp}_{\mathcal{I}}(X) \subseteq \operatorname{supp}(X)$. Fix $\tilde{M} = (\tilde{f}, \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q}) \in \mathcal{M}$ such that $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$ and let $(\tilde{X}, \tilde{Y}, \tilde{H}, \tilde{A})$ be generated by the SCM of $\tilde{M}$. We have that $(X, Y, A) \stackrel{\text{d}}{=} (\tilde{X}, \tilde{Y}, \tilde{A})$ and by Assumption 1, we have that $f \equiv \tilde{f}$ on $\operatorname{supp}(X)$, hence $f(X) \stackrel{\text{a.s.}}{=} \tilde{f}(X)$. Further, fix any $B \in \mathcal{B}(\mathbb{R}^p)$ (i.e., in the Borel sigma-algebra on $\mathbb{R}^p$) and note that

$$\begin{aligned}
\mathbb{E}_M[\mathbb{1}_B(A)X|A] &= \mathbb{E}_M[\mathbb{1}_B(A)g(A) + \mathbb{1}_B(A)h_2(H, \varepsilon_X)|A] \\
&= \mathbb{E}_M[\mathbb{1}_B(A)g(A)|A] + \mathbb{1}_B(A)\mathbb{E}[h_2(H, \varepsilon_X)] = \mathbb{1}_B(A)g(A),
\end{aligned}$$

almost surely. Here, we have used our modeling assumption $\mathbb{E}[h_2(H, \varepsilon_X)] = 0$. Hence, by similar arguments for $\mathbb{E}_{\tilde{M}}(\mathbb{1}_B(\tilde{A})\tilde{X}|\tilde{A})$ and the fact that $(X, Y, A) \stackrel{\text{d}}{=} (\tilde{X}, \tilde{Y}, \tilde{A})$ we have that

$$\mathbb{1}_B(A)g(A) \stackrel{\text{a.s.}}{=} \mathbb{E}_M(\mathbb{1}_B(A)X|A) \stackrel{\text{d}}{=} \mathbb{E}_{\tilde{M}}(\mathbb{1}_B(\tilde{A})\tilde{X}|\tilde{A}) \stackrel{\text{a.s.}}{=} \mathbb{1}_B(\tilde{A})\tilde{g}(\tilde{A}).$$

We conclude that $\mathbb{1}_B(A)g(A) \stackrel{\text{d}}{=} \mathbb{1}_B(\tilde{A})\tilde{g}(\tilde{A})$ for any $B \in \mathcal{B}(\mathbb{R}^p)$. Let $\mathbb{P}$ and $\tilde{\mathbb{P}}$ denote the respective background probability measures on which the random elements (X, Y, H, A) and $(\tilde{X}, \tilde{Y}, \tilde{H}, \tilde{A})$ are defined. Fix any $F \in \sigma(A)$ (i.e., in the sigma-algebra generated by A) and note that there exists a $B \in \mathcal{B}(\mathbb{R}^p)$ such that $F = \{A \in B\}$. Since $A \stackrel{\text{d}}{=} \tilde{A}$, we have that,

$$\int_F g(A)\,\mathrm{d}\mathbb{P} = \int \mathbb{1}_B(A)g(A)\,\mathrm{d}\mathbb{P} = \int \mathbb{1}_B(\tilde{A})\tilde{g}(\tilde{A})\,\mathrm{d}\tilde{\mathbb{P}} = \int \mathbb{1}_B(A)\tilde{g}(A)\,\mathrm{d}\mathbb{P} = \int_F \tilde{g}(A)\,\mathrm{d}\mathbb{P}.$$

Both $g(A)$ and $\tilde{g}(A)$ are $\sigma(A)$-measurable and they agree integral-wise over every set $F \in \sigma(A)$, so we must have that $g(A) \overset{\text{a.s.}}{=} \tilde{g}(A)$. With $\eta(a,b,c) = (a, c - \tilde{f}(b), b - \tilde{g}(a))$ we have that

$$(A, \xi_Y, \xi_X) \overset{\text{a.s.}}{=} (A, Y - \tilde{f}(X), X - \tilde{g}(A)) = \eta(A, X, Y) \overset{\text{d}}{=} \eta(\tilde{A}, \tilde{X}, \tilde{Y}) = (\tilde{A}, \tilde{\xi}_Y, \tilde{\xi}_X),$$

so $(A, \xi_Y, \xi_X) \overset{\text{d}}{=} (\tilde{A}, \tilde{\xi}_Y, \tilde{\xi}_X)$. This completes step (i).

Next, we proceed with step (ii). Take an arbitrary intervention $i \in \mathcal{I}$ and let $\varphi^i, I^i, \tilde{I}^i$ with $I^i \overset{\text{d}}{=} \tilde{I}^i$, $I^i \perp\!\!\!\perp (\varepsilon_X^i, \varepsilon_Y^i, \varepsilon_H^i, \varepsilon_A^i) \sim Q$ and $\tilde{I}^i \perp\!\!\!\perp (\tilde{\varepsilon}_X^i, \tilde{\varepsilon}_Y^i, \tilde{\varepsilon}_H^i, \tilde{\varepsilon}_A^i) \sim \tilde{Q}$ be such that the structural assignments for X^i and $\tilde{X}^i$ in $M(i)$ and $\tilde{M}(i)$, respectively, are given as

$$X^i := \varphi^i(A^i, g(A^i), h_2(H^i, \varepsilon_X^i), I^i) \quad \text{and} \quad \tilde{X}^i := \varphi^i(\tilde{A}^i, \tilde{g}(\tilde{A}^i), \tilde{h}_2(\tilde{H}^i, \tilde{\varepsilon}_X^i), \tilde{I}^i).$$

Define $\xi_X^i := h_2(H^i, \varepsilon_X^i)$, $\xi_Y^i := h_1(H^i, \varepsilon_Y^i)$, $\tilde{\xi}_X^i := \tilde{h}_2(\tilde{H}^i, \tilde{\varepsilon}_X^i)$ and $\tilde{\xi}_Y^i := \tilde{h}_1(\tilde{H}^i, \tilde{\varepsilon}_Y^i)$. Then, it holds that

$$(A^i, \xi_X^i, \xi_Y^i) \overset{\text{d}}{=} (A, \xi_X, \xi_Y) \overset{\text{d}}{=} (\tilde{A}, \tilde{\xi}_X, \tilde{\xi}_Y) \overset{\text{d}}{=} (\tilde{A}^i, \tilde{\xi}_X^i, \tilde{\xi}_Y^i),$$

where we used step (i), that (A^i, ξ_X^i, ξ_Y^i) and (A, ξ_X, ξ_Y) are generated by identical functions of the noise innovations and that $(\varepsilon_X, \varepsilon_Y, \varepsilon_H, \varepsilon_A)$ and $(\varepsilon_X^i, \varepsilon_Y^i, \varepsilon_H^i, \varepsilon_A^i)$ have identical distributions. Adding a random variable with the same distribution, that is mutually independent with all other variables, on both sides does not change the distribution of the bundle, hence

$$(A^i, \xi_X^i, \xi_Y^i, I^i) \overset{\text{d}}{=} (\tilde{A}^i, \tilde{\xi}_X^i, \tilde{\xi}_Y^i, \tilde{I}^i).$$

Define $\kappa(a,b,c,d) := (\varphi^i(a, \tilde{g}(a), b, d), \tilde{f}(\varphi^i(a, \tilde{g}(a), b, d)) + c)$. As shown in step (i) above, we have that $g(A^i) \overset{\text{a.s.}}{=} \tilde{g}(A^i)$. Furthermore, since $\operatorname{supp}(X^i) \subseteq \operatorname{supp}(X)$ we have that $f(X^i) \overset{\text{a.s.}}{=} \tilde{f}(X^i)$, and hence

$$\begin{aligned}
(X^i, Y^i) &\overset{\text{a.s.}}{=} (X^i, \tilde{f}(X^i) + \xi_Y^i) \\
&= (\varphi^i(A^i, g(A^i), \xi_X^i, I^i), \tilde{f}(\varphi^i(A^i, g(A^i), \xi_X^i, I^i)) + \xi_Y^i) \\
&\overset{\text{a.s.}}{=} (\varphi^i(A^i, \tilde{g}(A^i), \xi_X^i, I^i), \tilde{f}(\varphi^i(A^i, \tilde{g}(A^i), \xi_X^i, I^i)) + \xi_Y^i) \\
&= \kappa(A^i, \xi_X^i, \xi_Y^i, I^i) \overset{\text{d}}{=} \kappa(\tilde{A}^i, \tilde{\xi}_X^i, \tilde{\xi}_Y^i, \tilde{I}^i) = (\tilde{X}^i, \tilde{Y}^i).
\end{aligned}$$

Thus, $\mathbb{P}_{M(i)}^{(X,Y)} = \mathbb{P}_{\tilde{M}(i)}^{(X,Y)}$, which completes step (ii). Since $i \in \mathcal{I}$ was arbitrary, the result now follows from Proposition 4.1.

Case (B): Assume that the set of interventions $\mathcal{I}$ contains at least one confounding-removing intervention. Let $\tilde{M} = (\tilde{f}, \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q}) \in \mathcal{M}$ be such that $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$. Then, by Proposition 3.1, it follows that the causal function $\tilde{f}$ is a minimax solution w.r.t. $(\tilde{M}, \mathcal{I})$. By Assumption 1, we further have that $\tilde{f}$ and f coincide on $\operatorname{supp}(X) \supseteq \operatorname{supp}_{\mathcal{I}}(X)$. Hence, it follows that

$$\inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[(Y - f_\diamond(X))^2] = \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[(Y - \tilde{f}(X))^2] = \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[(Y - f(X))^2],$$

showing that also f is a minimax solution w.r.t. $(\tilde{M}, \mathcal{I})$. This completes the proof of Proposition 4.3. $\square$

E.8 Proof of Proposition 4.4

Proof. Let $\tilde{M} \in \mathcal{M}$ be such that $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$. By Assumptions 1 and 2, it holds that $f \equiv \tilde{f}$. The proof now proceeds analogously to that of Proposition 4.3. $\square$

E.9 Proof of Proposition 4.5

Proof. By Assumption 1, f is identified on $\text{supp}^M(X)$ by the observational distribution $\mathbb{P}_M$. Let $\mathcal{I}$ be a set of interventions containing at least one confounding-removing intervention. For any $\tilde{M} = (\tilde{f}, \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q}) \in \mathcal{M}$, Proposition 3.1 yields that the causal function is a minimax solution. That is,

$$\begin{aligned}\inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f_\diamond(X))^2\big] &= \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - \tilde{f}(X))^2\big] = \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[\xi_Y^2] \\ &= \mathbb{E}_{\tilde{M}}[\xi_Y^2], \end{aligned} \tag{27}$$

where we used that any intervention $i \in \mathcal{I}$ does not affect the distribution of $\xi_Y = \tilde{h}_2(H, \varepsilon_Y)$. Now, assume that $\tilde{M} = (\tilde{f}, \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q}) \in \mathcal{M}$ satisfies $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$. Since $(\mathbb{P}_M, \mathcal{M})$ satisfies Assumption 1, we have that $f \equiv \tilde{f}$ on $\text{supp}^M(X) = \text{supp}^{\tilde{M}}(X)$. Let f^* be any function in $\mathcal{F}$ such that $f^* = f$ on $\text{supp}^M(X)$. We first show that $\|\tilde{f} - f^*\|_{\mathcal{I},\infty} \leq 2\delta K$, where $\|f\|_{\mathcal{I},\infty} := \sup_{x \in \text{supp}_{\mathcal{I}}^M(X)} \|f(x)\|$. By the mean value theorem, for all $f_\diamond \in \mathcal{F}$ it holds that $|f_\diamond(x) - f_\diamond(y)| \leq K\|x - y\|$, for all $x, y \in \mathcal{D}$. For any $x \in \text{supp}_{\mathcal{I}}^M(X)$ and $y \in \text{supp}^M(X)$ we have

$$\begin{aligned}\big|\tilde{f}(x) - f^*(x)\big| &= \big|\tilde{f}(x) - \tilde{f}(y) + f^*(y) - f^*(x)\big| \\ &\leq \big|\tilde{f}(x) - \tilde{f}(y)\big| + \big|f^*(y) - f^*(x)\big| \\ &\leq 2K\|x - y\|,\end{aligned}$$

where we used the fact that $\tilde{f}(y) = f(y) = f^*(y)$, for all $y \in \text{supp}^M(X)$. In particular, it holds that

$$\begin{aligned}\|\tilde{f} - f^*\|_{\mathcal{I},\infty} &= \sup_{x \in \text{supp}_{\mathcal{I}}^M(X)} \big|\tilde{f}(x) - f^*(x)\big| \\ &\leq 2K \sup_{x \in \text{supp}_{\mathcal{I}}^M(X)} \inf_{y \in \text{supp}^M(X)} \|x - y\| \\ &= 2\delta K.\end{aligned} \tag{28}$$

For any $i \in \mathcal{I}$ we have that

$$\begin{aligned}\mathbb{E}_{\tilde{M}(i)}\big[(Y - f^*(X))^2\big] &= \mathbb{E}_{\tilde{M}(i)}\big[(\tilde{f}(X) + \xi_Y - f^*(X))^2\big] \\ &= \mathbb{E}_{\tilde{M}}\big[\xi_Y^2\big] + \mathbb{E}_{\tilde{M}(i)}\big[(\tilde{f}(X) - f^*(X))^2\big] \\ &\quad + 2\mathbb{E}_{\tilde{M}(i)}\big[\xi_Y(\tilde{f}(X) - f^*(X))\big].\end{aligned} \tag{29}$$

Next, we can use Cauchy-Schwarz, (27) and (28) in (29) to get that

$$\begin{aligned}&\left|\sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f^*(X))^2\big] - \inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f_\diamond(X))^2\big]\right| \\ &\quad = \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f^*(X))^2\big] - \mathbb{E}_{\tilde{M}}[\xi_Y^2] \\ &\quad = \sup_{i \in \mathcal{I}} \Big(\mathbb{E}_{\tilde{M}(i)}\big[(\tilde{f}(X) - f^*(X))^2\big] + 2\mathbb{E}_{\tilde{M}(i)}\big[\xi_Y(\tilde{f}(X) - f^*(X))\big]\Big) \\ &\quad \leq 4\delta^2 K^2 + 4\delta K \sqrt{\text{Var}_M(\xi_Y)},\end{aligned} \tag{30}$$

proving the first statement. Finally, if $\mathcal{I}$ consists only of confounding-removing interventions, then the bound in (30) can be improved by using that $\mathbb{E}[\xi_Y] = 0$ together with $H \perp\!\!\!\perp X$. In that case, we get that $\mathbb{E}_{\tilde{M}(i)}\big[\xi_Y(\tilde{f}(X) - f(X))\big] = 0$ and hence the bound becomes $4\delta^2 K^2$. This completes the proof of Proposition 4.5. $\square$

E.10 Proof of Proposition 4.6

Proof. By Assumption 1, f is identified on $\text{supp}^M(X)$ by the observational distribution $\mathbb{P}_M$. Let $\mathcal{I}$ be a set of confounding-preserving interventions. For a fixed $\varepsilon > 0$, let $f^* \in \mathcal{F}$ be a function satisfying

$$|\sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f^*(X))^2\big] - \inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f_\diamond(X))^2\big]| \leq \varepsilon. \tag{31}$$

Fix any secondary model $\tilde{M} = (\tilde{f}, \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q}) \in \mathcal{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$. The general idea is to derive an upper bound for $\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[(Y - f^*(X))^2]$ and a lower bound for $\inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[(Y - f_\diamond(X))^2]$ which will allow us to bound the absolute difference of interest.

Since $(\mathbb{P}_M, \mathcal{M})$ satisfies Assumption 1, we have that $f \equiv \tilde{f}$ on $\text{supp}^M(X) = \text{supp}^{\tilde{M}}(X)$. We first show that $\|\tilde{f} - f\|_{\mathcal{I},\infty} \leq 2\delta K$, where $\|f\|_{\mathcal{I},\infty} := \sup_{x\in\text{supp}^M_{\mathcal{I}}(X)} \|f(x)\|$. By the mean value theorem, for all $f_\diamond \in \mathcal{F}$ it holds that $|f_\diamond(x) - f_\diamond(y)| \leq K\|x - y\|$, for all $x, y \in \mathcal{D}$. For any $x \in \text{supp}^M_{\mathcal{I}}(X)$ and $y \in \text{supp}^M(X)$ we have

$$\begin{aligned}
\big|\tilde{f}(x) - f(x)\big| &= \big|\tilde{f}(x) - \tilde{f}(y) + f(y) - f(x)\big| \\
&\leq \big|\tilde{f}(x) - \tilde{f}(y)\big| + \big|f(y) - f(x)\big| \\
&\leq 2K\|x - y\|,
\end{aligned}$$

where we used the fact that $\tilde{f}(y) = f(y)$, for all $y \in \text{supp}_M(X)$. In particular, it holds that

$$\begin{aligned}
\|\tilde{f} - f\|_{\mathcal{I},\infty} &= \sup_{x\in\text{supp}^M_{\mathcal{I}}(X)} \big|\tilde{f}(x) - f(x)\big| \\
&\leq 2K \sup_{x\in\text{supp}^M_{\mathcal{I}}(X)} \inf_{y\in\text{supp}^M(X)} \|x - y\| \qquad (32)\\
&= 2\delta K.
\end{aligned}$$

Let now $i \in \mathcal{I}$ be fixed. The term $\xi_Y = h_1(H, \varepsilon_Y)$ is not affected by the intervention i. Furthermore, $\mathbb{P}^{(X,\xi_Y)}_{M(i)} = \mathbb{P}^{(X,\xi_Y)}_{\tilde{M}(i)}$ since i is confounding-preserving (this can be seen by a slight modification to the arguments from case (A) in the proof of Proposition 4.3). Thus, for any $f_\diamond \in \mathcal{F}$ we have that

$$\begin{aligned}
&\mathbb{E}_{\tilde{M}(i)}\big[(Y - f_\diamond(X))^2\big] \\
&\quad= \mathbb{E}_{\tilde{M}(i)}\big[(\tilde{f}(X) + \xi_Y - f_\diamond(X) + f(X) - f(X))^2\big] \\
&\quad= \mathbb{E}_{\tilde{M}(i)}\big[\xi_Y^2\big] + \mathbb{E}_{\tilde{M}(i)}\big[(f(X) - f_\diamond(X))^2\big] + \mathbb{E}_{\tilde{M}(i)}\big[(\tilde{f}(X) - f(X))^2\big] \\
&\qquad\quad + 2\mathbb{E}_{\tilde{M}(i)}\big[\xi_Y(f(X) - f_\diamond(X))\big] \\
&\qquad\quad + 2\mathbb{E}_{\tilde{M}(i)}\big[(\tilde{f}(X) - f(X))(f(X) - f_\diamond(X))\big] \\
&\qquad\quad + 2\mathbb{E}_{\tilde{M}(i)}\big[\xi_Y(\tilde{f}(X) - f(X))\big] \\
&\quad= \mathbb{E}_{M(i)}\big[\xi_Y^2\big] + \mathbb{E}_{M(i)}\big[(f(X) - f_\diamond(X))^2\big] + \mathbb{E}_{M(i)}\big[(\tilde{f}(X) - f(X))^2\big] \\
&\qquad\quad + 2\mathbb{E}_{M(i)}\big[\xi_Y(f(X) - f_\diamond(X))\big] \\
&\qquad\quad + 2\mathbb{E}_{M(i)}\big[(\tilde{f}(X) - f(X))(f(X) - f_\diamond(X))\big] \\
&\qquad\quad + 2\mathbb{E}_{M(i)}\big[\xi_Y(\tilde{f}(X) - f(X))\big] \\
&\quad= \mathbb{E}_{M(i)}\big[(Y - f_\diamond(X))^2\big] + L_1^i(\tilde{f}) + L_2^i(\tilde{f}, f_\diamond) + L_3^i(\tilde{f}), \qquad (33)
\end{aligned}$$

where, we have made the following definitions

$$\begin{aligned}
L_1^i(\tilde{f}) &:= \mathbb{E}_{M(i)}\big[(\tilde{f}(X) - f(X))^2\big], \\
L_2^i(\tilde{f}, f_\diamond) &:= 2\mathbb{E}_{M(i)}\big[(\tilde{f}(X) - f(X))(f(X) - f_\diamond(X))\big], \\
L_3^i(\tilde{f}) &:= 2\mathbb{E}_{M(i)}\big[\xi_Y(\tilde{f}(X) - f(X))\big].
\end{aligned}$$

Using (32) it follows that

$$0 \leq L_1^i(\tilde{f}) \leq 4\delta^2 K^2, \tag{34}$$

and by the Cauchy-Schwarz inequality it follows that

$$\left|L_3^i(\tilde{f})\right| \leq 2\sqrt{\mathrm{Var}_M(\xi_Y)4\delta^2K^2} = 4\delta K\sqrt{\mathrm{Var}_M(\xi_Y)}. \tag{35}$$

Let now $f_\diamond \in \mathcal{F}$ be any function such that

$$\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f_\diamond(X))^2)\right] \leq \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - \tilde{f}(X))^2)\right], \tag{36}$$

then by (32), the Cauchy-Schwarz inequality and Proposition 3.3, it holds for all $i \in \mathcal{I}$ that

$$\begin{aligned}
L_2^i(\tilde{f}, f_\diamond) &= 2\mathbb{E}_{M(i)}\left[(\tilde{f}(X) - f(X))(f(X) - f_\diamond(X))\right] \\
&= 2\mathbb{E}_{\tilde{M}(i)}\left[(\tilde{f}(X) - f(X))(f(X) - f_\diamond(X))\right] \\
&= -2\mathbb{E}_{\tilde{M}(i)}\left[(\tilde{f}(X) - f(X))^2\right] + 2\mathbb{E}_{\tilde{M}(i)}\left[(\tilde{f}(X) - f(X))(\tilde{f}(X) - f_\diamond(X))\right] \\
&\geq -8\delta^2K^2 - 2\sqrt{4\delta^2K^2}\sqrt{4\,\mathrm{Var}_M(\xi_Y)} \\
&= -8\delta^2K^2 - 8\delta K\sqrt{\mathrm{Var}_M(\xi_Y)},
\end{aligned} \tag{37}$$

where, in the third equality, we have added and subtracted the term $2\mathbb{E}_{\tilde{M}(i)}\left[(\tilde{f}(X) - f(X))\tilde{f}(X)\right]$. Now let $\mathcal{S} := \{f_\diamond \in \mathcal{F} : \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f_\diamond(X))^2\right] \leq \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - \tilde{f}(X))^2\right]\}$ be the set of all functions satisfying (36). Due to (33), (34), (35) and (37) we have the following lower bound of interest

$$\begin{aligned}
&\inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f_\diamond(X))^2\right] \\
&= \inf_{f_\diamond\in\mathcal{S}} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f_\diamond(X))^2\right] \\
&= \inf_{f_\diamond\in\mathcal{S}} \sup_{i\in\mathcal{I}} \left\{\mathbb{E}_{M(i)}\left[(Y - f_\diamond(X))^2\right] + L_1^i(\tilde{f}) + L_2^i(\tilde{f}, f_\diamond) + L_3^i(\tilde{f})\right\} \\
&\geq \inf_{f_\diamond\in\mathcal{S}} \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\left[(Y - f_\diamond(X))^2\right] - 8\delta^2K^2 - 8\delta K\sqrt{\mathrm{Var}_M(\xi_Y)} - 4\delta K\sqrt{\mathrm{Var}_M(\xi_Y)} \\
&\geq \inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\left[(Y - f_\diamond(X))^2\right] - 8\delta^2K^2 - 12\delta K\sqrt{\mathrm{Var}_M(\xi_Y)}.
\end{aligned} \tag{38}$$

Next, we construct the aforementioned upper bound of interest. To that end, note that

$$\begin{aligned}
&\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f^*(X))^2\right] \\
&\quad = \sup_{i\in\mathcal{I}} \left\{\mathbb{E}_{M(i)}\left[(Y - f^*(X))^2\right] + L_1^i(\tilde{f}) + L_2^i(\tilde{f}, f^*) + L_3^i(\tilde{f})\right\},
\end{aligned} \tag{39}$$

by (33). We have already established upper bounds for $L_1^i(\tilde{f})$ and $L_3^i(\tilde{f})$ in (34) and (35), respectively. In order to control $L_2^i(\tilde{f}, f^*)$ we introduce an auxiliary function. Let $\bar{f}^* \in \mathcal{F}$ satisfy

$$\sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\left[(Y - \bar{f}^*(X))^2)\right] \leq \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\left[(Y - f(X))^2)\right], \tag{40}$$

and

$$\left|\sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\left[(Y - \bar{f}^*(X))^2\right] - \inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\left[(Y - f_\diamond(X))^2\right]\right| \leq \varepsilon. \tag{41}$$

Choosing such a $\bar{f}^* \in \mathcal{F}$ is always possible. If f is an ε-minimax solution, i.e., it satisfies (41), then choose $\bar{f}^* = f$. Otherwise, if f is not a ε-minimax solution, then choose any $\bar{f}^* \in \mathcal{F}$ that is an ε-minimax solution (which is always possible). In this case we have that

$$\sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\left[(Y - \bar{f}^*(X))^2\right] - \inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\left[(Y - f_\diamond(X))^2\right] \leq \varepsilon,$$

and

$$\sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f(X))^2\big] - \inf_{f_\diamond \in \mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f_\diamond(X))^2\big] \geq \varepsilon,$$

which implies that (40) is satisfied. We can now construct an upper bound on $L_2^i(\tilde{f}, f^*)$ in terms of $L_2^i(\tilde{f}, \bar{f}^*)$ by noting that for all $i \in \mathcal{I}$

$$\begin{aligned}
\big|L_2^i(\tilde{f}, f^*)\big| &= 2\big|\mathbb{E}_{M(i)}\big[(\tilde{f}(X) - f(X))(f(X) - f^*(X))\big]\big| \\
&\leq 2\big|\mathbb{E}_{M(i)}\big[(\tilde{f}(X) - f(X))(f(X) - \bar{f}^*(X))\big]\big| \\
&\qquad + 2\mathbb{E}_{M(i)}\big|(\tilde{f}(X) - f(X))(\bar{f}^*(X) - f^*(X))\big| \\
&= \big|L_2^i(\tilde{f}, \bar{f}^*)\big| + 2\mathbb{E}_{M(i)}\big|(\tilde{f}(X) - f(X))(\bar{f}^*(X) - f^*(X))\big| \\
&\leq \big|L_2^i(\tilde{f}, \bar{f}^*)\big| + 2\sqrt{\mathbb{E}_{M(i)}\left[(\tilde{f}(X) - f(X))^2\right]\mathbb{E}_{M(i)}\left[(\bar{f}^*(X) - f^*(X))^2\right]} \\
&\leq \big|L_2^i(\tilde{f}, \bar{f}^*)\big| + 4\delta K\sqrt{\mathbb{E}_{M(i)}\left[(\bar{f}^*(X) - f^*(X))^2\right]},
\end{aligned} \tag{42}$$

where we used the triangle inequality, Cauchy-Schwarz inequality and (32) . Furthermore, (32) and (40) together with Proposition 3.3 yield the following bound

$$\begin{aligned}
|L_2^i(\tilde{f}, \bar{f}^*)| &= 2\big|\mathbb{E}_{M(i)}\big[(\tilde{f}(X) - f(X))(f(X) - \bar{f}^*(X))\big]\big| \\
&= 2\sqrt{\mathbb{E}_{M(i)}\big[(\tilde{f}(X) - f(X))^2\big]\mathbb{E}_{M(i)}\big[(f(X) - \bar{f}^*(X))^2\big]} \\
&\leq 2\sqrt{4\delta^2K^2}\sqrt{4\operatorname{Var}_M(\xi_Y)} \\
&= 8\delta K\sqrt{\operatorname{Var}_M(\xi_Y)},
\end{aligned} \tag{43}$$

for any $i \in \mathcal{I}$. Thus, it suffices to construct an upper bound on the second term in the final expression in (42). Direct computation leads to

$$\begin{aligned}
\mathbb{E}_{M(i)}\big[(Y - f^*(X))^2\big] = \mathbb{E}_{M(i)}\big[(Y - \bar{f}^*(X))^2\big] \\
+ \mathbb{E}_{M(i)}\big[(\bar{f}^*(X) - f^*(X))^2\big] \\
+ 2\mathbb{E}_{M(i)}\big[(Y - \bar{f}^*(X))(\bar{f}^*(X) - f^*(X))\big].
\end{aligned}$$

Rearranging the terms and applying the triangle inequality and Cauchy-Schwarz results in

$$\begin{aligned}
&\mathbb{E}_{M(i)}\big[(\bar{f}^*(X) - f^*(X))^2\big] \\
&\quad = \mathbb{E}_{M(i)}\big[(Y - f^*(X))^2\big] - \mathbb{E}_{M(i)}\big[(Y - \bar{f}^*(X))^2\big] \\
&\qquad - 2\mathbb{E}_{M(i)}\big[(Y - \bar{f}^*(X))(\bar{f}^*(X) - f^*(X))\big] \\
&\quad \leq \big|\mathbb{E}_{M(i)}\big[(Y - f^*(X))^2\big] - \inf_{f_\diamond \in \mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f_\diamond(X))^2\big]\big| \\
&\qquad + \big|\inf_{f_\diamond \in \mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f_\diamond(X))^2\big] - \mathbb{E}_{M(i)}\big[(Y - \bar{f}^*(X))^2\big]\big| \\
&\qquad + 2\mathbb{E}_{M(i)}\big|(Y - \bar{f}^*(X))(\bar{f}^*(X) - f^*(X))\big| \\
&\quad \leq 2\varepsilon + 2\sqrt{\mathbb{E}_{M(i)}\big[(Y - \bar{f}^*(X))^2\big]}\sqrt{\mathbb{E}_{M(i)}\big[(\bar{f}^*(X) - f^*(X))^2\big]} \\
&\quad \leq 2\varepsilon + 2\sqrt{\operatorname{Var}_M(\xi_Y)}\sqrt{\mathbb{E}_{M(i)}\big[(\bar{f}^*(X) - f^*(X))^2\big]},
\end{aligned}$$

for any $i \in \mathcal{I}$. Here, we used that both f^* and $\bar{f}^*$ are ε-minimax solutions with respect to M and that $\bar{f}^*$ satisfies (40) which implies that

$$\mathbb{E}_{M(i)}\big[(Y - \bar{f}^*(X))^2)\big] \leq \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f(X))^2)\big] = \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\big[\xi_Y^2\big] = \operatorname{Var}_M(\xi_Y),$$

for any $i \in \mathcal{I}$, as ξ_Y is unaffected by an intervention on X. Thus, $\mathbb{E}_{M(i)}\big[(\bar{f}^*(X) - f^*(X))^2\big]$ must satisfy $\ell(\mathbb{E}_{M(i)}\big[(\bar{f}^*(X) - f^*(X))^2\big]) \leq 0$, where $\ell : [0, \infty) \to \mathbb{R}$ is given by $\ell(z) =$

$z - 2\varepsilon - 2\sqrt{\mathrm{Var}_M(\xi_Y)}\sqrt{z}$. The linear term of ℓ grows faster than the square root term, so the largest allowed value of $\mathbb{E}_{M(i)}\big[(\bar{f}^*(X) - f^*(X))^2\big]$ coincides with the largest root of $\ell(z)$. The largest root is given by

$$C^2 := 2\varepsilon + 2\,\mathrm{Var}_M(\xi_Y) + 2\sqrt{\mathrm{Var}_M(\xi_Y)^2 + 2\varepsilon\,\mathrm{Var}_M(\xi_Y)},$$

where $(\cdot)^2$ refers to the square of C. Hence, for any $i \in \mathcal{I}$ it holds that

$$\mathbb{E}_{M(i)}\big[(\bar{f}^*(X) - f^*(X))^2\big] \leq C^2. \tag{44}$$

Hence by (42), (43) and (44) we have that the following upper bound is valid for any $i \in \mathcal{I}$.

$$\big|L_2^i(\tilde{f}, f^*)\big| \leq 8\delta K\sqrt{\mathrm{Var}_M(\xi_Y)} + 4\delta K C. \tag{45}$$

Thus, using (39) with (34), (35) and (45), we get the following upper bound

$$\begin{aligned}
&\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f^*(X))^2\big] \\
&\quad\leq \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f^*(X))^2\big] + 4\delta^2K^2 + 4\delta K C + 12\delta K\sqrt{\mathrm{Var}_M(\xi_Y)}.
\end{aligned} \tag{46}$$

Finally, by combining the bounds (38) and (46) together with (31) we get that

$$\begin{aligned}
&\Big|\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f^*(X))^2\big] - \inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f_\diamond(X))^2\big]\Big| \\
&\quad\leq \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f^*(X))^2\big] - \inf_{f_\diamond\in\mathcal{F}} \sup_{i\in\mathcal{I}} \mathbb{E}_{M(i)}\big[(Y - f_\diamond(X))^2\big] \\
&\qquad\quad + 4\delta^2K^2 + 4\delta K C + 12\delta K\sqrt{\mathrm{Var}_M(\xi_Y)} \\
&\qquad\quad + 8\delta^2K^2 + 12\delta K\sqrt{\mathrm{Var}_M(\xi_Y)} \\
&\quad\leq \varepsilon + 12\delta^2K^2 + 24\delta K\sqrt{\mathrm{Var}_M(\xi_Y)} + 4\delta K C.
\end{aligned} \tag{47}$$

Using that all terms are positive, we get that

$$C = \sqrt{\mathrm{Var}_M(\xi_Y)} + \sqrt{\mathrm{Var}_M(\xi_Y) + 2\varepsilon} \leq 2\sqrt{\mathrm{Var}_M(\xi_Y)} + \sqrt{2\varepsilon}$$

Hence, (47) is bounded above by

$$\varepsilon + 12\delta^2K^2 + 32\delta K\sqrt{\mathrm{Var}_M(\xi_Y)} + 4\sqrt{2}\delta K\sqrt{\varepsilon}.$$

This completes the proof of Proposition 4.6. □

E.11 Proof of Proposition 4.7

Proof. Let $\bar{f} \in \mathcal{F}$ and $c > 0$. By assumption, $\mathcal{I}$ is a well-behaved set of support-extending interventions on X. Since $\mathrm{supp}^M_{\mathcal{I}}(X) \setminus \mathrm{supp}^M(X)$ has non-empty interior, there exists an intervention $i_0 \in \mathcal{I}$ and $\varepsilon > 0$ such that $\mathbb{P}_{M(i_0)}(X \in B) \geq \varepsilon$, for some open subset $B \subsetneq \bar{B}$, such that $\mathrm{dist}(B, \mathbb{R}^d \setminus \bar{B}) > 0$, where $\bar{B} := \mathrm{supp}^M_{\mathcal{I}}(X) \setminus \mathrm{supp}^M(X)$. Let $\tilde{f}$ be any continuous function satisfying that, for all $x \in B \cup (\mathbb{R}^d \setminus \bar{B})$,

$$\tilde{f}(x) = \begin{cases} \bar{f}(x) + \gamma, & x \in B \\ f(x), & x \in \mathbb{R}^d \setminus \bar{B}, \end{cases}$$

where $\gamma := \varepsilon^{-1/2}\left\{(2\mathbb{E}_{\tilde{M}}[\xi_Y^2] + c)^{1/2} + (\mathbb{E}_{\tilde{M}}[\xi_Y^2])^{1/2}\right\}$.

Consider a secondary model $\tilde{M} = (\tilde{f}, g, h_1, h_2, Q) \in \mathcal{M}$. Then, by Assumption 1, it holds that $\mathbb{P}_M = \mathbb{P}_{\tilde{M}}$. Since $\mathcal{I}$ only consists of interventions on X, it holds that $\mathbb{P}_{M(i_0)}(X \in$

$B) = \mathbb{P}_{\tilde{M}(i_0)}(X \in B)$ (this holds since all components of $\tilde{M}$ and M are equal, except for the function f, which is not allowed to enter in the intervention on X). Therefore,

$$
\begin{aligned}
\mathbb{E}_{\tilde{M}(i_0)}\big[(Y - \bar{f}(X))^2\big] &\geq \mathbb{E}_{\tilde{M}(i_0)}\big[(Y - \bar{f}(X))^2 \mathbb{1}_B(X)\big] \\
&= \mathbb{E}_{\tilde{M}(i_0)}\big[(\gamma + \xi_Y)^2 \mathbb{1}_B(X)\big] \\
&\geq \gamma^2 \varepsilon + 2\gamma \mathbb{E}_{\tilde{M}(i_0)}\big[\xi_Y \mathbb{1}_B(X)\big] \\
&\geq \gamma^2 \varepsilon - 2\gamma \left(\mathbb{E}_{\tilde{M}}\big[\xi_Y^2\big]\varepsilon\right)^{1/2} \\
&= c + \mathbb{E}_{\tilde{M}}[\xi_Y^2], \qquad (48)
\end{aligned}
$$

where the third inequality follows from Cauchy–Schwarz. Further, by the definition of the infimum it holds that

$$
\inf_{f_\diamond \in \mathcal{F}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f_\diamond(X))^2\big] \leq \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - \tilde{f}(X))^2\big] = \mathbb{E}_{\tilde{M}}[\xi_Y^2]. \qquad (49)
$$

Therefore, combining (48) and (49), the claim follows. □

E.12 Proof of Proposition 4.8

Proof. We prove the result by showing that under Assumption 3 it is possible to express interventions on A as confounding-preserving interventions on X and applying Propositions 4.3 and 4.4. To avoid confusion, we will throughout this proof denote the true model by $M^0 = (f^0, g^0, h_1^0, h_2^0, Q^0)$. Fix an intervention $i \in \mathcal{I}$. Since it is an intervention on A, there exist ψ^i and I^i such that for any $M = (f, g, h_1, h_2, Q) \in \mathcal{M}$, the intervened SCM $M(i)$ is of the form

$$
A^i := \psi^i(I^i, \varepsilon_A^i), \quad H^i := \varepsilon_H^i, \quad X^i := g(A^i) + h_2(H^i, \varepsilon_X^i), \quad Y^i := f(X^i) + h_1(H^i, \varepsilon_Y^i),
$$

where $(\varepsilon_X^i, \varepsilon_Y^i, \varepsilon_A^i, \varepsilon_H^i) \sim Q$. We now define a confounding-preserving intervention j on X, such that, for all models $\tilde{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, the distribution of (X, Y) under $\tilde{M}(j)$ coincides with that under $\tilde{M}(i)$. To that end, define the intervention function

$$
\bar{\psi}^j(h_2, A^j, H^j, \varepsilon_X^j, I^j) := g^0(\psi^i(I^j, A^j)) + h_2(H^j, \varepsilon_X^j),
$$

where g^0 is the fixed function corresponding to model M, and therefore not an argument of $\bar{\psi}^j$. Let now j be the intervention on X satisfying that, for all $M = (f, g, h_1, h_2, Q) \in \mathcal{M}$, the intervened model $M(j)$ is given as

$$
A^j := \varepsilon_A^j, \quad H^j := \varepsilon_H^j, \quad X^j := \bar{\psi}^j(h_2, A^j, H^j, \varepsilon_X^j, I^j), \quad Y^j := f(X^j) + h_1(H^j, \varepsilon_Y^j),
$$

where $(\varepsilon_X^j, \varepsilon_Y^j, \varepsilon_A^j, \varepsilon_H^j) \sim Q$ and where I^j is chosen such that $I^j \overset{\mathrm{d}}{=} I^i$. By definition, j is a confounding-preserving intervention. Let now $\tilde{M} = (\tilde{f}, \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q})$ be such that $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, and let $(\tilde{X}^i, \tilde{Y}^i)$ and $(\tilde{X}^j, \tilde{Y}^j)$ be generated under $\tilde{M}(i)$ and $\tilde{M}(j)$, respectively. By Assumption 3, it holds for all $a \in \operatorname{supp}(A) \cup \operatorname{supp}_{\mathcal{I}}(A)$ that $\tilde{g}(a) = g^0(a)$. Hence, we get that

$$
\begin{aligned}
(\tilde{X}^i, \tilde{Y}^i) &\overset{\mathrm{d}}{=} (\tilde{g}(\psi^i(I^i, \tilde{\varepsilon}_A^i)) + \tilde{h}_2(\tilde{\varepsilon}_H^i, \tilde{\varepsilon}_X^i), \tilde{f}(\tilde{g}(\psi^i(I^i, \tilde{\varepsilon}_A^i)) + \tilde{h}_2(\tilde{\varepsilon}_H^i, \tilde{\varepsilon}_X^i)) + \tilde{h}_1(\tilde{\varepsilon}_H^i, \tilde{\varepsilon}_Y^i)) \\
&= (g^0(\psi^i(I^i, \tilde{\varepsilon}_A^i)) + \tilde{h}_2(\tilde{\varepsilon}_H^i, \tilde{\varepsilon}_X^i), \tilde{f}(g^0(\psi^i(I^i, \tilde{\varepsilon}_A^i)) + \tilde{h}_2(\tilde{\varepsilon}_H^i, \tilde{\varepsilon}_X^i)) + \tilde{h}_1(\tilde{\varepsilon}_H^i, \tilde{\varepsilon}_Y^i)) \\
&\overset{\mathrm{d}}{=} (g^0(\psi^i(I^j, \tilde{\varepsilon}_A^j)) + \tilde{h}_2(\tilde{\varepsilon}_H^j, \tilde{\varepsilon}_X^j), \tilde{f}(g^0(\psi^i(I^j, \tilde{\varepsilon}_A^j)) + \tilde{h}_2(\tilde{\varepsilon}_H^j, \tilde{\varepsilon}_X^j)) + \tilde{h}_1(\tilde{\varepsilon}_H^j, \tilde{\varepsilon}_Y^j)) \\
&\overset{\mathrm{d}}{=} (\bar{\psi}^j(\tilde{h}_2, \tilde{\varepsilon}_A^j, \tilde{\varepsilon}_H^j, \tilde{\varepsilon}_X^j, I^j), \tilde{f}(\bar{\psi}^j(\tilde{h}_2, \tilde{\varepsilon}_A^j, \tilde{\varepsilon}_H^j, \tilde{\varepsilon}_X^j, I^j)) + \tilde{h}_1(\tilde{\varepsilon}_H^j, \tilde{\varepsilon}_Y^j)) \\
&\overset{\mathrm{d}}{=} (\tilde{X}^j, \tilde{Y}^j),
\end{aligned}
$$

as desired. Since $i \in \mathcal{I}$ was arbitrary, we have now shown that there exists a mapping π from $\mathcal{I}$ into a set $\mathcal{J}$ of confounding-preserving (and hence a well-behaved set) of interventions on X, such that for all $\tilde{M}$ with $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$, $\mathbb{P}_{\tilde{M}(i)}^{(X,Y)} = \mathbb{P}_{\tilde{M}(\pi(i))}^{(X,Y)}$. Hence, we can rewrite Equation (2) in Definition 2.1 in terms of the set $\mathcal{J}$. The result now follows from Propositions 4.3 and 4.4. □

E.13 Proof of Proposition 4.9

Proof. Let $b \in \mathbb{R}^d$ be such that $f(x) = b^\top x$ for all $x \in \mathbb{R}^d$. We start by characterizing the error $\mathbb{E}_{\tilde{M}(i)}\big[(Y - f_\diamond(X))^2\big]$. Let us consider models of the form $\tilde{M} = (f, \tilde{g}, h_1, h_2, Q) \in \mathcal{M}$ for some function $\tilde{g} \in \mathcal{G}$ with $\tilde{g}(a) = g(a)$ for all $a \in \mathrm{supp}_M(A)$. Clearly, any such model satisfies that $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$. For every $a \in \mathcal{A}$, let $i_a \in \mathcal{I}$ denote the corresponding hard intervention on A. For every $a \in \mathcal{A}$ and $b_\diamond \in \mathbb{R}^d$, we then have

$$
\begin{aligned}
&\mathbb{E}_{\tilde{M}(i_a)}\big[(Y - b_\diamond^\top X)^2\big] \\
&= \mathbb{E}_{\tilde{M}(i_a)}\big[(b^\top X + \xi_Y - b_\diamond^\top X)^2\big] \\
&= (b - b_\diamond)^\top \mathbb{E}_{\tilde{M}(i_a)}[XX^\top](b - b_\diamond) + 2(b - b_\diamond)^\top \mathbb{E}_{\tilde{M}(i_a)}[X\xi_Y] + \mathbb{E}_{\tilde{M}(i_a)}\big[\xi_Y^2\big] \qquad (50) \\
&= (b - b_\diamond)^\top \underbrace{(\tilde{g}(a)\tilde{g}(a)^\top + \mathbb{E}_M[\xi_X\xi_X^\top])}_{=:K_{\tilde{M}}(a)}(b - b_\diamond) + 2(b - b_\diamond)^\top \mathbb{E}_M[\xi_X\xi_Y] + \mathbb{E}_M\big[\xi_Y^2\big],
\end{aligned}
$$

where we have used that, under i_a, the distribution of (ξ_X, ξ_Y) is unaffected. We now show that, for any $\tilde{M}$ with the above form, the causal function f does not minimize the worst-case risk across interventions in $\mathcal{I}$. The idea is to show that the worst-case risk (50) strictly decreases at $b_\diamond = b$ in the direction $u := \mathbb{E}_M[\xi_X\xi_Y]/\|\mathbb{E}_M[\xi_X\xi_Y]\|_2$. For every $a \in \mathcal{A}$ and $s \in \mathbb{R}$, define

$$
\ell_{\tilde{M},a}(s) := \mathbb{E}_{\tilde{M}(i_a)}\big[(Y - (b + su)^\top X)^2\big] = u^\top K_{\tilde{M}}(a)u \cdot s^2 - 2u^\top \mathbb{E}_M[\xi_X\xi_Y] \cdot s + \mathbb{E}_M\big[\xi_Y^2\big].
$$

For every a, $\ell'_{\tilde{M},a}(0) = -2\|\mathbb{E}_M[\xi_X\xi_Y]\|_2 < 0$, showing that $\ell_{\tilde{M},a}$ is strictly decreasing at $s = 0$ (with a derivative that is bounded away from 0 across all $a \in \mathcal{A}$). By boundedness of $\mathcal{A}$ and by the continuity of $a \mapsto \ell''_{\tilde{M},a}(0) = 2u^\top K_{\tilde{M}}(a)u$, it further follows that $\sup_{a\in\mathcal{A}}|\ell''_{\tilde{M},a}(0)| < \infty$. Hence, we can find $s_0 > 0$ such that for all $a \in \mathcal{A}$, $\ell_{\tilde{M},a}(0) > \ell_{\tilde{M},a}(s_0)$. It now follows by continuity of $(a, s) \mapsto \ell_{\tilde{M},a}(s)$ that

$$
\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - b^\top X)^2\big] = \sup_{a\in\mathcal{A}} \ell_{\tilde{M},a}(0) > \sup_{a\in\mathcal{A}} \ell_{\tilde{M},a}(s_0) = \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - (b + s_0u)^\top X)^2\big],
$$

showing that $b + s_0u$ attains a lower worst-case risk than b.

We now show that all functions other than f may result in an arbitrarily large error. Let $\bar{b} \in \mathbb{R}^d \setminus \{b\}$ be given, and let $j \in \{1, \dots, d\}$ be such that $b_j \neq \bar{b}_j$. The idea is to construct a function $\tilde{g} \in \mathcal{G}$ such that, under the corresponding model $\tilde{M} = (f, \tilde{g}, h_1, h_2, Q) \in \mathcal{M}$, some hard interventions on A result in strong shifts of the jth coordinate of X. Let $a \in \mathcal{A}$. Let $e_j \in \mathbb{R}^d$ denote the jth unit vector, and assume that $\tilde{g}(a) = ne_j$ for some $n \in \mathbb{N}$. Using (50), it follows that

$$
\begin{aligned}
&\mathbb{E}_{\tilde{M}(i_a)}\big[(Y - \bar{b}^\top X)^2\big] \\
&\qquad = n^2(\bar{b}_j - b_j)^2 + (\bar{b} - b)^\top \mathbb{E}_M[\xi_X\xi_X^\top](\bar{b} - b) + 2(\bar{b} - b)^\top \mathbb{E}_M[\xi_X\xi_Y] + \mathbb{E}_M\big[\xi_Y^2\big].
\end{aligned}
$$

By letting $n \to \infty$, we see that the above error may become arbitrarily large. Given any $c > 0$, we can therefore construct $\tilde{g}$ such that $\mathbb{E}_{\tilde{M}(i_a)}\big[(Y - \bar{b}^\top X)^2\big] \geq c + \mathbb{E}_M\big[\xi_Y^2\big]$. By carefully choosing $a \in \mathrm{int}(\mathcal{A} \setminus \mathrm{supp}_M(A))$, this can be done such that $\tilde{g}$ is continuous and $\tilde{g}(a) = g(a)$ for all $a \in \mathrm{supp}_M(A)$, ensuring that $\mathbb{P}_{\tilde{M}} = \mathbb{P}_M$. It follows that

$$
\begin{aligned}
c &\leq \mathbb{E}_{\tilde{M}(i_a)}\big[(Y - \bar{b}^\top X)^2\big] - \mathbb{E}_M\big[\xi_Y^2\big] \\
&= \mathbb{E}_{\tilde{M}(i_a)}\big[(Y - \bar{b}^\top X)^2\big] - \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - b^\top X)^2\big] \\
&\leq \mathbb{E}_{\tilde{M}(i_a)}\big[(Y - \bar{b}^\top X)^2\big] - \inf_{b_\diamond\in\mathbb{R}^d} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - b_\diamond^\top X)^2\big] \\
&\leq \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - \bar{b}^\top X)^2\big] - \inf_{b_\diamond\in\mathbb{R}^d} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - b_\diamond^\top X)^2\big],
\end{aligned}
$$

which completes the proof of Proposition 4.9. □

E.14 Proof of Proposition 5.1

Proof. By assumption, $\mathcal{I}$ is a set of interventions on X or A of which at least one is confounding-removing. Now fix any

$$\tilde{M} = (f_{\eta_0}(x; \tilde{\theta}), \tilde{g}, \tilde{h}_1, \tilde{h}_2, \tilde{Q}) \in \mathcal{M},$$

with $\mathbb{P}_M = \mathbb{P}_{\tilde{M}}$. By Proposition 3.1, we have that a minimax solution is given by the causal function. That is,

$$\inf_{f_\diamond \in \mathcal{F}_{\eta_0}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f_\diamond(X))^2\big] = \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f_{\eta_0}(X; \tilde{\theta}))^2\big] = \mathbb{E}_M[\xi_Y^2],$$

where we used that ξ_Y is unaffected by an intervention on X. By the support restriction $\mathrm{supp}^M(X) \subseteq (a, b)$ we know that

$$f_{\eta_0}(x; \theta^0) = B(x)^\top \theta^0, \quad f_{\eta_0}(x; \tilde{\theta}) = B(x)^\top \tilde{\theta}, \quad f_{\eta_0}(x; \hat{\theta}^n_{\lambda_n^\star, \eta_0, \mu}) = B(x)^\top \hat{\theta}^n_{\lambda_n^\star, \eta_0, \mu},$$

for all $x \in \mathrm{supp}^M(X)$. Furthermore, as $Y = B(X)^\top \theta^0 + \xi_Y$ $\mathbb{P}_M$-almost surely, we have that

$$\mathbb{E}_M\left[C(A)Y\right] = \mathbb{E}_M\left[C(A)B(X)^\top \theta^0\right] + \mathbb{E}_M\left[C(A)\xi_Y\right] = \mathbb{E}_M\left[C(A)B(X)^\top\right]\theta^0, \tag{51}$$

where we used the assumptions that $\mathbb{E}\left[\xi_Y\right] = 0$ and $A \perp\!\!\!\perp \xi_Y$ by the exogeneity of A. Similarly,

$$\mathbb{E}_{\tilde{M}}\left[C(A)Y\right] = \mathbb{E}_{\tilde{M}}\left[C(A)B(X)^\top\right]\tilde{\theta}.$$

As $\mathbb{P}_M = \mathbb{P}_{\tilde{M}}$, we have that $\mathbb{E}_M[C(A)Y] = \mathbb{E}_{\tilde{M}}[C(A)Y]$ and $\mathbb{E}_M[C(A)B(X)^\top] = \mathbb{E}_{\tilde{M}}[C(A)B(X)^\top]$, hence

$$\mathbb{E}_M\left[C(A)B(X)^\top\right]\tilde{\theta} = \mathbb{E}_M\left[C(A)B(X)^\top\right]\theta^0 \iff \tilde{\theta} = \theta^0,$$

by assumption (B2), which states that $\mathbb{E}[C(A)B(X)^\top]$ is of full rank (bijective). In other words, the causal function parameterized by θ^0 is identified from the observational distribution. Assumptions 1 and 2 are therefore satisfied. Furthermore, we also have that

$$\begin{aligned}
&\sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f_{\eta_0}(X; \hat{\theta}^n_{\lambda_n^\star, \eta_0, \mu}))^2\big] \\
&\quad = \sup_{i \in \mathcal{I}} \big\{\mathbb{E}_{\tilde{M}(i)}\big[(f_{\eta_0}(X; \theta^0) - f_{\eta_0}(X; \hat{\theta}^n_{\lambda_n^\star, \eta_0, \mu}))^2\big] + \mathbb{E}_{\tilde{M}(i)}\big[\xi_Y^2\big] \\
&\qquad\quad + 2\mathbb{E}_{\tilde{M}(i)}\big[\xi_Y(f_{\eta_0}(X; \theta^0) - f_{\eta_0}(X; \hat{\theta}^n_{\lambda_n^\star, \eta_0, \mu}))\big]\big\} \\
&\quad \leq \sup_{i \in \mathcal{I}} \big\{\mathbb{E}_{\tilde{M}(i)}\big[(f_{\eta_0}(X; \theta^0) - f_{\eta_0}(X; \hat{\theta}^n_{\lambda_n^\star, \eta_0, \mu}))^2\big] + \mathbb{E}_{\tilde{M}(i)}\big[\xi_Y^2\big] \\
&\qquad\quad + 2\sqrt{\mathbb{E}_{\tilde{M}(i)}\big[\xi_Y^2\big]\mathbb{E}_{\tilde{M}(i)}\big[(f_{\eta_0}(X; \theta^0) - f_{\eta_0}(X; \hat{\theta}^n_{\lambda_n^\star, \eta_0, \mu}))^2\big]}\big\} \\
&\quad \leq \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(f_{\eta_0}(X; \theta^0) - f_{\eta_0}(X; \hat{\theta}^n_{\lambda_n^\star, \eta_0, \mu}))^2\big] + \mathbb{E}_M\big[\xi_Y^2\big] \\
&\qquad\quad + 2\sqrt{\mathbb{E}_M\big[\xi_Y^2\big]\sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(f_{\eta_0}(X; \theta^0) - f_{\eta_0}(X; \hat{\theta}^n_{\lambda_n^\star, \eta_0, \mu}))^2\big]},
\end{aligned}$$

by Cauchy-Schwarz inequality, where we additionally used that $\mathbb{E}_{\tilde{M}(i)}[\xi_Y^2] = \mathbb{E}_M[\xi_Y^2]$ as ξ_Y is unaffected by interventions on X. Thus,

$$\begin{aligned}
&\big|\sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f_{\eta_0}(X; \hat{\theta}^n_{\lambda_n^\star, \eta_0, \mu}))^2\big] - \inf_{f_\diamond \in \mathcal{F}_{\eta_0}} \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(Y - f_\diamond(X))^2\big]\big| \\
&\quad \leq \sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(f_{\eta_0}(X; \theta^0) - f_{\eta_0}(X; \hat{\theta}^n_{\lambda_n^\star, \eta_0, \mu}))^2\big] \\
&\qquad\quad + 2\sqrt{\mathbb{E}_M\big[\xi_Y^2\big]\sup_{i \in \mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\big[(f_{\eta_0}(X; \theta^0) - f_{\eta_0}(X; \hat{\theta}^n_{\lambda_n^\star, \eta_0, \mu}))^2\big]}.
\end{aligned}$$

For the next few derivations let $\hat{\theta} = \hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}$ for notational simplicity. Note that, for all $x \in \mathbb{R}$,

$$
\begin{aligned}
(f_{\eta_0}(x;\theta^0) - f_{\eta_0}(x;\hat{\theta}))^2 \leq (\theta^0 - \hat{\theta})^\top B(x) B(x)^\top (\theta^0 - \hat{\theta}) \\
+ (B(a)^\top (\theta^0 - \hat{\theta}) + B'(a)^\top (\theta^0 - \hat{\theta})(x-a))^2 \\
+ (B(b)^\top (\theta^0 - \hat{\theta}) + B'(b)^\top (\theta^0 - \hat{\theta})(x-b))^2.
\end{aligned}
$$

The second term has the following upper bound

$$
\begin{aligned}
&(B(a)^\top (\theta^0 - \hat{\theta}) + B'(a)^\top (\theta^0 - \hat{\theta})(x-a))^2 \\
&\quad = (\theta^0 - \hat{\theta})^\top B(a) B(a)^\top (\theta^0 - \hat{\theta}) \\
&\qquad + (x-a)^2 (\theta^0 - \hat{\theta})^\top B'(a) B'(a)^\top (\theta^0 - \hat{\theta}) \\
&\qquad + 2(x-a)(\theta^0 - \hat{\theta})^\top B'(a) B(a)^\top (\theta^0 - \hat{\theta}) \\
&\quad \leq \lambda_{\max}(B(a)B(a)^\top) \|\theta^0 - \hat{\theta}\|_2^2 \\
&\qquad + (x-a)^2 \lambda_{\max}(B'(a)B'(a)^\top) \|\theta^0 - \hat{\theta}\|_2^2 \\
&\qquad + 2(x-a)\lambda_{\max}((B'(a)B(a)^\top + B(a)B'(a)^\top)/2) \|\theta^0 - \hat{\theta}\|_2^2,
\end{aligned}
$$

where $\lambda_{\max}$ denotes the maximum eigenvalue. An analogous upper bound can be constructed for the third term. Thus, by combining these two upper bounds with a similar upper bound for the first term, we arrive at

$$
\begin{aligned}
&\mathbb{E}_{\tilde{M}(i)}\left[(f_{\eta_0}(X;\theta^0) - f_{\eta_0}(X;\hat{\theta}))^2\right] \\
&\quad \leq \lambda_{\max}(\mathbb{E}_{\tilde{M}(i)}[B(X)B(X)^\top]) \|\theta^0 - \hat{\theta}\|_2^2 \\
&\qquad + \lambda_{\max}(B(a)B(a)^\top) \|\theta^0 - \hat{\theta}\|_2^2 \\
&\qquad + \mathbb{E}_{\tilde{M}(i)}[(X-a)^2] \lambda_{\max}(B'(a)B'(a)^\top) \|\theta^0 - \hat{\theta}\|_2^2 \\
&\qquad + 2\mathbb{E}_{\tilde{M}(i)}[X-a] \lambda_{\max}((B'(a)B(a)^\top + B(a)B'(a)^\top)/2) \|\theta^0 - \hat{\theta}\|_2^2 \\
&\qquad + \lambda_{\max}(B(b)B(b)^\top) \|\theta^0 - \hat{\theta}\|_2^2 \\
&\qquad + \mathbb{E}_{\tilde{M}(i)}[(X-b)^2] \lambda_{\max}(B'(b)B'(b)^\top) \|\theta^0 - \hat{\theta}\|_2^2 \\
&\qquad + 2\mathbb{E}_{\tilde{M}(i)}[X-b] \lambda_{\max}((B'(b)B(b)^\top + B(b)B'(b)^\top)/2) \|\theta^0 - \hat{\theta}\|_2^2.
\end{aligned}
$$

Assumption (B1) imposes that $\sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}[X^2]$ and $\sup_{i\in\mathcal{I}} \lambda_{\max}(\mathbb{E}_{\tilde{M}(i)}[B(X)B(X)^\top])$ are finite. Hence, the supremum of each of the above terms is finite. That is, there exists a constant $c > 0$ such that

$$
\begin{aligned}
&\Big| \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f_{\eta_0}(X;\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}))^2\right] - \inf_{f_\diamond \in \mathcal{F}_{\eta_0}} \sup_{i\in\mathcal{I}} \mathbb{E}_{\tilde{M}(i)}\left[(Y - f_\diamond(X))^2\right] \Big| \\
&\quad \leq c \|\theta^0 - \hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}\|_2^2 + 2\sqrt{\mathbb{E}_M\left[\xi_Y^2\right] c}\, \|\theta^0 - \hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}\|_2.
\end{aligned}
$$

It therefore suffices to show that

$$
\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu} \xrightarrow[n\to\infty]{P} \theta^0,
$$

with respect to the distribution induced by M. To simplify notation, we henceforth drop the M subscript in the expectations and probabilities. Note that by the rank conditions in (B2), and the law of large numbers, we may assume that the corresponding sample product moments satisfy the same conditions. That is, for the purpose of the following arguments, it suffices that the sample product moment only satisfies these rank conditions asymptotically with probability one.

Let $B := B(X)$, $C := C(A)$, let $\mathbf{B}$ and $\mathbf{C}$ be row-wise stacked i.i.d. copies of $B(X)^\top$ and $C(A)^\top$, and recall the definition $\mathbf{P}_\delta := \mathbf{C}\left(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M}\right)^{-1}\mathbf{C}^\top$. By convexity of the objective function we can find a closed form expression for our estimator of θ^0 by solving the corresponding normal equations. The closed form expression is given by

$$\begin{aligned}\hat{\theta}^n_{\lambda,\eta,\mu} &:= \underset{\theta\in\mathbb{R}^k}{\operatorname{argmin}} \|\mathbf{Y}-\mathbf{B}\theta\|_2^2 + \lambda\|\mathbf{P}_\delta(\mathbf{Y}-\mathbf{B}\theta)\|_2^2 + \gamma\theta^\top\mathbf{K}\theta, \\ &= \left(\frac{\mathbf{B}^\top\mathbf{B}}{n} + \lambda_n^\star\frac{\mathbf{B}^\top\mathbf{P}_\delta\mathbf{P}_\delta\mathbf{B}}{n} + \frac{\gamma\mathbf{K}}{n}\right)^{-1}\left(\frac{\mathbf{B}^\top\mathbf{Y}}{n} + \lambda_n^\star\frac{\mathbf{B}^\top\mathbf{P}_\delta\mathbf{P}_\delta\mathbf{Y}}{n}\right),\end{aligned}$$

where we used that $\lambda_n^\star \in [0,\infty)$ almost surely by (C2). Consequently (using standard convergence arguments and that $n^{-1}\gamma\mathbf{K}$ and $n^{-1}\delta\mathbf{M}$ converges to zero in probability), if $\lambda_n^\star$ diverges to infinity in probability as n tends to infinity, then

$$\begin{aligned}\hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu} &\xrightarrow{P} \left(\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}\left[CB^\top\right]\right)^{-1}\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}\left[CY\right] \\ &= \theta^0.\end{aligned}$$

Here, we also used that the terms multiplied by $\lambda_n^\star$ are the only asymptotically relevant terms. These are the standard arguments that the K-class estimator (with minor penalized regression modifications) is consistent as long as the parameter $\lambda_n^\star$ converges to infinity, or, equivalently, $\kappa_n^\star = \lambda_n^\star/(1+\lambda_n^\star)$ converges to one in probability.

We now consider two cases: *(i)* $\mathbb{E}[B\xi_Y] \neq 0$ and *(ii)* $\mathbb{E}[B\xi_Y] = 0$, corresponding to the case with unmeasured confounding and without, respectively. For *(i)* we show that $\lambda_n^\star$ converges to infinity in probability and for *(ii)* we show consistency by other means (as $\lambda_n^\star$ might not converge to infinity in this case).

Case (i): The confounded case $\mathbb{E}[B\xi_Y] \neq 0$. It suffices to show that

$$\lambda_n^\star := \inf\{\lambda \geq 0 : T_n(\hat{\theta}^n_{\lambda,\eta_0,\mu}) \leq q(\alpha)\} \xrightarrow[n\to\infty]{P} \infty.$$

To that end, note that for fixed $\lambda \geq 0$ we have that

$$\hat{\theta}^n_{\lambda,\eta_0,\mu} \xrightarrow[n\to\infty]{P} \theta_\lambda, \tag{52}$$

where

$$\begin{aligned}\theta_\lambda := &\left(\mathbb{E}\left[BB^\top\right] + \lambda\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}\left[CB^\top\right]\right)^{-1} \\ &\times\left(\mathbb{E}\left[BY\right] + \lambda\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}\left[CY\right]\right).\end{aligned} \tag{53}$$

Recall that (51) states that $\mathbb{E}\left[CY\right] = \mathbb{E}\left[CB^\top\right]\theta^0$. Using (51) and that $Y = B^\top\theta^0 + \xi_Y$ $\mathbb{P}_M$-almost surely, we have that the latter factor of (53) is given by

$$\begin{aligned}&\mathbb{E}\left[BY\right] + \lambda\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}\left[CY\right] \\ &\quad= \mathbb{E}\left[BB^\top\right]\theta^0 + \mathbb{E}\left[B\xi_Y\right] + \lambda\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}\left[CB^\top\right]\theta^0 \\ &\quad= \left(\mathbb{E}\left[BB^\top\right] + \lambda\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}\left[CB^\top\right]\right)\theta^0 + \mathbb{E}\left[B\xi_Y\right]\end{aligned}$$

Inserting this into (53) we arrive at the following representation of θ_λ

$$\theta_\lambda = \theta^0 + \left(\mathbb{E}\left[BB^\top\right] + \lambda\mathbb{E}\left[BC^\top\right]\mathbb{E}\left[CC^\top\right]^{-1}\mathbb{E}\left[CB^\top\right]\right)^{-1}\mathbb{E}\left[B\xi_Y\right]. \tag{54}$$

Since $\mathbb{E}\left[B\xi_Y\right] \neq 0$ by assumption, the above yields that

$$\forall\lambda \geq 0: \qquad \theta^0 \neq \theta_\lambda. \tag{55}$$

Now we prove that $\lambda_n^\star$ diverges to infinity in probability as n tends to infinity. That is, for any $\lambda \geq 0$ we will prove that

$$\lim_{n\to\infty} \mathbb{P}(\lambda_n^\star \leq \lambda) = 0.$$

We fix an arbitrary $\lambda \geq 0$. By (55) we have that $\theta^0 \neq \theta_\lambda$. This implies that there exists an $\varepsilon > 0$ such that $\theta^0 \notin \overline{B(\theta_\lambda, \varepsilon)}$, where $\overline{B(\theta_\lambda, \varepsilon)}$ is the closed ball in $\mathbb{R}^k$ with center θ_λ and radius ε. By the consistency result (52), we know that the sequence of events $(A_n)_{n\in\mathbb{N}}$, for every $n \in \mathbb{N}$, given by

$$A_n := (|\hat{\theta}^n_{\lambda,\eta_0,\mu} - \theta_\lambda| \leq \varepsilon) = (\hat{\theta}^n_{\lambda,\eta_0,\mu} \in \overline{B(\theta_\lambda, \varepsilon)}),$$

satisfies $\mathbb{P}(A_n) \to 1$ as $n \to \infty$. By assumption (C3) we have that

$$\tilde{\lambda} \mapsto T_n(\theta^n_{\tilde{\lambda},\eta_0,\mu}), \qquad \text{and} \qquad \theta \mapsto T_n(\theta),$$

are weakly decreasing and continuous, respectively. Together with the continuity of $\tilde{\lambda} \mapsto \hat{\theta}^n_{\tilde{\lambda},\eta_0,\mu}$, this implies that also the mapping $\tilde{\lambda} \mapsto T_n(\hat{\theta}^n_{\tilde{\lambda},\eta_0,\mu})$ is continuous. It now follows from Assumption (C2) (stating that $\lambda_n^\star$ is almost surely finite) that for all $n \in \mathbb{N}$, $\mathbb{P}(T_n(\hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu}) \leq q(\alpha)) = 1$. Furthermore, since $\tilde{\lambda} \mapsto T_n(\theta^n_{\tilde{\lambda},\eta_0,\mu})$ is weakly decreasing, it follows that

$$\begin{aligned}
\mathbb{P}(\lambda_n^\star \leq \lambda) &= \mathbb{P}(\{\lambda_n^\star \leq \lambda\} \cap \{T_n(\hat{\theta}^n_{\lambda_n^\star,\eta_0,\mu}) \leq q(\alpha)\}) \\
&\leq \mathbb{P}(\{\lambda_n^\star \leq \lambda\} \cap \{T_n(\hat{\theta}^n_{\lambda,\eta_0,\mu}) \leq q(\alpha)\}) \\
&= \mathbb{P}(\{\lambda_n^\star \leq \lambda\} \cap \{T_n(\hat{\theta}^n_{\lambda,\eta_0,\mu}) \leq q(\alpha)\} \cap A_n) \\
&\qquad + \mathbb{P}(\{\lambda_n^\star \leq \lambda\} \cap \{T_n(\hat{\theta}^n_{\lambda,\eta_0,\mu}) \leq q(\alpha)\} \cap A_n^c) \\
&\leq \mathbb{P}(\{\lambda_n^\star \leq \lambda\} \cap \{T_n(\hat{\theta}^n_{\lambda,\eta_0,\mu}) \leq q(\alpha)\} \cap \{|\hat{\theta}^n_{\lambda,\eta_0,\mu} - \theta_\lambda| \leq \varepsilon\}) + \mathbb{P}(A_n^c).
\end{aligned}$$

It now suffices to show that the first term converges to zero, since $\mathbb{P}(A_n^c) \to 0$ as $n \to \infty$. We have

$$\begin{aligned}
&\mathbb{P}(\{\lambda_n^\star \leq \lambda\} \cap \{T_n(\hat{\theta}^n_{\lambda,\eta_0,\mu}) \leq q(\alpha)\} \cap \{|\hat{\theta}^n_{\lambda,\eta_0,\mu} - \theta_\lambda| \leq \varepsilon\}) \\
&\leq \mathbb{P}\Big(\{\lambda_n^\star \leq \lambda\} \cap \Big\{ \inf_{\theta \in \overline{B(\theta_\lambda,\varepsilon)}} T_n(\theta) \leq q(\alpha)\Big\} \cap \{|\hat{\theta}^n_{\lambda,\eta_0,\mu} - \theta_\lambda| \leq \varepsilon\}\Big) \\
&\leq \mathbb{P}\Big(\inf_{\theta \in \overline{B(\theta_\lambda,\varepsilon)}} T_n(\theta) \leq q(\alpha)\Big) \\
&\xrightarrow{P} 0,
\end{aligned}$$

as $n \to \infty$, since $\overline{B(\theta_\lambda, \varepsilon)}$ is a compact set not containing θ^0. Here, we used that the test statistic (T_n) is assumed to have compact uniform power (C1). Hence, $\lim_{n\to\infty} \mathbb{P}(\lambda_n^\star \leq \lambda) = 0$ for any $\lambda \geq 0$, proving that $\lambda_n^\star$ diverges to infinity in probability, which ensures consistency.

Case (ii): the unconfounded case $\mathbb{E}[B(X)\xi_Y] = 0$. Recall that

$$\begin{aligned}
\hat{\theta}^n_{\lambda,\eta_0,\mu} &:= \operatorname*{argmin}_{\theta \in \mathbb{R}^k} \|\mathbf{Y} - \mathbf{B}\theta\|_2^2 + \lambda \|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2 + \gamma \theta^\top \mathbf{K}\theta \\
&= \operatorname*{argmin}_{\theta \in \mathbb{R}^k} l^n_{\text{OLS}}(\theta) + \lambda l^n_{\text{TSLS}}(\theta) + \gamma l_{\text{PEN}}(\theta),
\end{aligned} \tag{56}$$

where we defined $l^n_{\text{OLS}}(\theta) := n^{-1}\|\mathbf{Y} - \mathbf{B}\theta\|_2^2$, $l^n_{\text{TSLS}}(\theta) := n^{-1}\|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\theta)\|_2^2$, and $l_{\text{PEN}}(\theta) := n^{-1}\theta^\top \mathbf{K}\theta$. For any $0 \leq \lambda_1 < \lambda_2$ we have

$$\begin{aligned}
&l^n_{\text{OLS}}(\hat{\theta}^n_{\lambda_1,\eta_0,\mu}) + \lambda_1 l^n_{\text{TSLS}}(\hat{\theta}^n_{\lambda_1,\eta_0,\mu}) + \gamma l_{\text{PEN}}(\hat{\theta}^n_{\lambda_1,\eta_0,\mu}) \\
&\quad \leq l^n_{\text{OLS}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}) + \lambda_1 l^n_{\text{TSLS}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}) + \gamma l_{\text{PEN}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}) \\
&\quad = l^n_{\text{OLS}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}) + \lambda_2 l^n_{\text{TSLS}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}) + \gamma l_{\text{PEN}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}) + (\lambda_1 - \lambda_2) l^n_{\text{TSLS}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}) \\
&\quad \leq l^n_{\text{OLS}}(\hat{\theta}^n_{\lambda_1,\eta_0,\mu}) + \lambda_2 l^n_{\text{TSLS}}(\hat{\theta}^n_{\lambda_1,\eta_0,\mu}) + \gamma l_{\text{PEN}}(\hat{\theta}^n_{\lambda_1,\eta_0,\mu}) + (\lambda_1 - \lambda_2) l^n_{\text{TSLS}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}),
\end{aligned}$$

where we used (56). Rearranging this inequality and dividing by $(\lambda_1 - \lambda_2)$ yields

$$l^n_{\text{TSLS}}(\hat{\theta}^n_{\lambda_1,\eta_0,\mu}) \geq l^n_{\text{TSLS}}(\hat{\theta}^n_{\lambda_2,\eta_0,\mu}),$$

proving that $\lambda \mapsto l^n_{\text{TSLS}}(\hat{\theta}^n_{\lambda,\eta_0,\mu})$ is weakly decreasing. Thus, since $\lambda^\star_n \geq 0$ almost surely, we have that

$$l^n_{\text{TSLS}}(\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}) \leq l^n_{\text{TSLS}}(\hat{\theta}^n_{0,\eta_0,\mu}) = n^{-1}(\mathbf{Y} - \mathbf{B}\hat{\theta}^n_{0,\eta_0,\mu})^\top \mathbf{P}_\delta \mathbf{P}_\delta (\mathbf{Y} - \mathbf{B}\hat{\theta}^n_{0,\eta_0,\mu}). \tag{57}$$

Furthermore, recall from (52) that

$$\hat{\theta}^n_{0,\eta_0,\mu} \xrightarrow[n\to\infty]{P} \theta_0 = \theta^0, \tag{58}$$

where the last equality follows from (54) using that we are in the unconfounded case $\mathbb{E}[B(X)\xi_Y] = 0$. By expanding and deriving convergence statements for each term, we get

$$\begin{aligned}
&(\mathbf{Y} - \mathbf{B}\hat{\theta}^n_{0,\eta_0,\mu})^\top \mathbf{P}_\delta \mathbf{P}_\delta (\mathbf{Y} - \mathbf{B}\hat{\theta}^n_{0,\eta_0,\mu}) \\
&\xrightarrow[n\to\infty]{P} (\mathbb{E}[YC^\top] - \theta_0 \mathbb{E}[BC^\top])\mathbb{E}[C^\top C]^{-1}(\mathbb{E}[CY] - \mathbb{E}[CB^\top]\theta_0) \\
&= 0,
\end{aligned} \tag{59}$$

where we used Slutsky's theorem, the weak law of large numbers, (58) and (51). Thus, by (57) and (59) it holds that

$$l^n_{\text{TSLS}}(\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}) = n^{-1}\|\mathbf{P}_\delta(\mathbf{Y} - \mathbf{B}\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu})\|_2^2 \xrightarrow[n\to\infty]{P} 0.$$

For any $z \in \mathbb{R}^n$ we have that

$$\begin{aligned}
\|\mathbf{P}_\delta z\|_2^2 &= z^\top \mathbf{C}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}\mathbf{C}^\top\mathbf{C}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}\mathbf{C}^\top z \\
&= z^\top \mathbf{C}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}(\mathbf{C}^\top\mathbf{C})^{1/2}(\mathbf{C}^\top\mathbf{C})^{1/2}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}\mathbf{C}^\top z \\
&= \|(\mathbf{C}^\top\mathbf{C})^{1/2}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}\mathbf{C}^\top z\|_2^2,
\end{aligned}$$

hence

$$\begin{aligned}
\|H_n - G_n\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu}\|_2^2 &= \|n^{-1/2}(\mathbf{C}^\top\mathbf{C})^{1/2}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}\mathbf{C}^\top(\mathbf{Y} - \mathbf{B}\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu})\|_2^2 \\
&\xrightarrow{P} 0,
\end{aligned} \tag{60}$$

where for each $n \in \mathbb{N}$, $G_n \in \mathbb{R}^{k\times k}$ and $H_n \in \mathbb{R}^{k\times 1}$ are defined as

$$\begin{aligned}
G_n &:= n^{-1/2}(\mathbf{C}^\top\mathbf{C})^{1/2}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}\mathbf{C}^\top\mathbf{B}, \text{ and} \\
H_n &:= n^{-1/2}(\mathbf{C}^\top\mathbf{C})^{1/2}(\mathbf{C}^\top\mathbf{C} + \delta\mathbf{M})^{-1}\mathbf{C}^\top\mathbf{Y}.
\end{aligned}$$

Using the weak law of large numbers, the continuous mapping theorem and Slutsky's theorem, it follows that, as $n \to \infty$,

$$\begin{aligned}
G_n &\xrightarrow{P} G := E[CC^\top]^{1/2}E[CC^\top]^{-1}E[CB^\top], \text{ and} \\
H_n &\xrightarrow{P} H := E[CC^\top]^{1/2}E[CC^\top]^{-1}E[CY] \\
&\qquad = E[CC^\top]^{1/2}E[CC^\top]^{-1}E[CB^\top]\theta^0 \\
&\qquad = G\theta^0,
\end{aligned}$$

where the second to last equality follows from (51). Together with (60), we now have that

$$\|G_n\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu} - G\theta^0\|_2^2 \leq \|G_n\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu} - H_n\|_2^2 + \|H_n - G\theta^0\|_2^2 \xrightarrow[n\to\infty]{P} 0.$$

Furthermore, by the rank assumptions in (B2) we have that $G_n \in \mathbb{R}^{k\times k}$ is of full rank (with probability tending to one), hence

$$\begin{aligned}\|\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu} - \theta^0\|_2^2 &= \|G_n^{-1}G_n(\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu} - \theta^0)\|_2^2 \\ &\leq \|G_n^{-1}\|_{\text{op}}^2\|G_n(\hat{\theta}^n_{\lambda^\star_n,\eta_0,\mu} - \theta^0)\|_2^2 \\ &\xrightarrow{P} \|G^{-1}\|_{\text{op}}^2 \cdot 0 \\ &= 0,\end{aligned}$$

as $n \to \infty$, proving the proposition.

□

References

[1] S. S. Abadeh, P. M. M. Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pages 1576–1584, 2015.

[2] J. Aldrich. Autonomy. *Oxford Economic Papers*, 41:15–34, 1989.

[3] T. Amemiya. The nonlinear two-stage least-squares estimator. *Journal of Econometrics*, 2:105–110, 1974.

[4] T. W. Anderson and H. Rubin. Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, 20(1): 46–63, 1949.

[5] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *ArXiv e-prints (1907.02893)*, 2019.

[6] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[7] J. A. Bagnell. Robust supervised learning. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, page 714. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

[8] P. L. Bartlett, V. Dani, T. Hayes, S. Kakade, A. Rakhlin, and A. Tewari. High-probability regret bounds for bandit online linear optimization. In *21st Annual Conference on Learning Theory (COLT)*, 2008.

[9] J. Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

[10] S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(Sep):2137–2155, 2009.

[11] J. Blanchet, Y. Kang, K. Murthy, and F. Zhang. Data-driven optimal transport cost selection for distributionally robust optimization. In *2019 Winter Simulation Conference (WSC)*, pages 3740–3751. IEEE, 2019.

[12] K. A. Bollen. *Structural Equations with Latent Variables.* John Wiley & Sons, New York, 1989.

[13] S. Bongers, J. Peters, B. Schölkopf, and J. M. Mooij. Foundations of structural causal models with cycles and latent variables. *ArXiv e-prints (1611.06221v3)*, 2016.

[14] R. J. Bowden and D. A. Turkington. *Instrumental Variables.* Econometric Society Monographs. Cambridge University Press, 1985.

[15] R. Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

[16] H. Chen, Y. Wang, R. Li, and K. Shear. A note on a nonparametric regression test through penalized splines. *Statistica Sinica*, 24:1143, 2014.

[17] G. Csurka. Domain adaptation for visual applications: A comprehensive survey. In G. Csurka, editor, *Domain Adaptation in Computer Vision Applications*. Springer, 2017.

[18] S. Darolles, Y. Fan, J.-P. Florens, and E. Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.

[19] H. Daume III and D. Marcu. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26:101–126, 2006.

[20] S. B. David, T. Lu, T. Luu, and D. Pal. Impossibility theorems for domain adaptation. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 129–136. PMLR, 2010.

[21] L. El Ghaoui, G. R. G. Lanckriet, and G. Natsoulis. Robust classification with interval data. Technical report, 2003.

[22] P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

[23] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and D. Song. Robust physical-world attacks on deep learning models. *ArXiv e-prints (1707.08945)*, 2017.

[24] L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. *Regression: models, methods and applications*. Springer Science & Business Media, 2013.

[25] F. M. Fisher. *The identification problem in econometrics*. McGraw-Hill, New York, NY, USA, 1966.

[26] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *ArXiv e-prints (1412.6572)*, 2014.

[27] W. H. Greene. *Econometric analysis*. Pearson Education, Upper Saddle River, New Jersey, USA, 2003.

[28] T. Haavelmo. The probability approach in econometrics. *Econometrica*, 12:S1–S115 (supplement), 1944.

[29] A. R. Hall. *Generalized method of moments*. Oxford University Press, 2005.

[30] C. Heinze-Deml and N. Meinshausen. Conditional variance penalties and domain shift robustness. *ArXiv e-prints (1710.11469)*, 2017.

[31] J. L. Horowitz. Applied nonparametric instrumental variables estimation. *Econometrica*, 79(2):347–394, 2011.

[32] P. Hoyer, S. Shimizu, A. Kerminen, and M. Palviainen. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49:362–378, 2008.

[33] Z. Hu and L. J. Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 2013.

[34] M. E. Jakobsen and J. Peters. Distributional robustness of K-class estimators and the PULSE. *arXiv preprint arXiv:2005.03353*, 2020.

[35] D. Janzing, J. Peters, J. M. Mooij, and B. Schölkopf. Identifying confounders using additive noise models. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 249–257. AUAI Press, 2009.

[36] D. W. Jorgenson and J.-J. Laffont. Efficient estimation of nonlinear simultaneous equations with additive disturbances. In *Annals of Economic and Social Measurement, Volume 3, number 4*, pages 615–640. National Bureau of Economic Research, Inc, 1974.

[37] H. H. Kelejian. Two-stage least squares and econometric systems linear in parameters but nonlinear in the endogenous variables. *Journal of the American Statistical Association*, 66(334):373–374, 1971.

[38] S.-J. Kim, A. Magnani, and S. Boyd. Robust fisher discriminant analysis. In *Advances in Neural Information Processing Systems 18 (NIPS)*, pages 659–666, 2006.

[39] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[40] S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 10846–10856. Curran Associates, Inc., 2018.

[41] Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems 21 (NIPS)*, pages 1041–1048. Curran Associates, Inc., 2009.

[42] R. S. Mariano. Simultaneous equation model estimators: Statistical properties and practical implications. In B. H. Baltagi, editor, *A Companion to Theoretical Econometrics*, chapter 6, pages 122–43. Blackwell Publishing Ltd, NJ, USA, 2001.

[43] N. Meinshausen. Causality from a distributional robustness point of view. In *IEEE Data Science Workshop*, pages 6–10, 2018.

[44] N. Meinshausen and P. Bühlmann. Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, 43(4):1801–1830, 2015.

[45] K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18, 2013.

[46] W. K. Newey. Nonparametric instrumental variables estimation. *American Economic Review*, 103(3):550–56, 2013.

[47] W. K. Newey and J. L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

[48] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345 – 1359, 2010.

[49] J. Pearl. *Causality.* Cambridge University Press, New York, USA, 2nd edition, 2009.

[50] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

[51] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms.* MIT Press, Cambridge, MA, 2017.

[52] N. Pfister, S. Bauer, and J. Peters. Learning stable and predictive structures in kinetic systems. *Proceedings of the National Academy of Sciences*, 116(51):25405–25411, 2019.

[53] N. Pfister, E. G. Williams, J. Peters, R. Aebersold, and P. Bühlmann. Stabilizing variable selection and regression. *arXiv preprint arXiv:1911.01850*, 2019.

[54] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

[55] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

[56] J. S. Racine and T. Hayfield. *np: Nonparametric Kernel Smoothing Methods for Mixed Data Types*, 2018. URL https://CRAN.R-project.org/package=np. R package version 0.60–10.

[57] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Causal transfer in machine learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

[58] D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: heterogeneous data meets causality. *ArXiv e-prints (1801.06229)*, 2018.

[59] N. Sani, J. Lee, and I. Shpitser. Identification and estimation of causal effects defined by shift interventions. In *Proceedings of the 36th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2020.

[60] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1255–1262. Omnipress, 2012.

[61] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227 – 244, 2000.

[62] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

[63] A. Sinha, H. Namkoong, and J. Duchi. Certifying some distributional robustness with principled adversarial training. *ArXiv e-prints (1710.10571)*, 2017.

[64] M. Sugiyama and K. Müller. Generalization error estimation under covariate shift. In *Workshop on Information-Based Induction Sciences*, 2005.

[65] M. Sugiyama, S. Nakajima, H. Kashima, P. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems 21 (NIPS)*, pages 1433 – 1440, 2008.

[66] H. Theil. *Economic forecasts and policy*. North-Holland, Amsterdam, Netherlands, 1958.

[67] R. Volpi, P. Morerio, S. Savarese, and V. Murino. Adversarial feature augmentation for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5495–5504, 2018.

[68] J. M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT Press, Cambridge, MA, 2010.

[69] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[70] K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, Corvallis, Oregon, USA, 2009. AUAI Press.