

# Counterfactual Cocycles: A Framework for Robust and Coherent Counterfactual Transports

**Hugh Dance**

*Gatsby Computational Neuroscience Unit  
University College London  
London, United Kingdom*

HUGH.DANCE.15@UCL.AC.UK

**Benjamin Bloem-Reddy**

*Department of Statistics  
University of British Columbia  
Vancouver, Canada*

BENBR@STAT.UBC.CA

## Abstract

Estimating joint distributions, or couplings, over counterfactual outcomes is central to personalized decision-making and treatment risk assessment. Two emergent frameworks with identifiability guarantees are: (i) bijective structural causal models (SCMs), which are flexible but brittle to mis-specified latent noise; and (ii) optimal-transport (OT) methods, which avoid latent noise assumptions but can produce incoherent counterfactual transports which fail to identify higher-order couplings. In this work, we bridge the gap with *counterfactual cocycles*: a framework for counterfactual transports that uses algebraic structure to provide coherence and identifiability guarantees. Every counterfactual cocycle corresponds to an equivalence class of SCMs, however the cocycle is invariant to the latent noise distribution, enabling us to sidestep various mis-specification problems. We characterize the structure of all identifiable counterfactual cocycles; propose flexible model parameterizations; introduce a novel cocycle estimator that avoids any distributional assumptions; and derive mis-specification robustness properties of the resulting counterfactual inference method. We demonstrate state-of-the-art performance and noise-robustness of counterfactual cocycles across synthetic benchmarks and a 401(k) eligibility study.

**Keywords:** causal inference, counterfactuals, structural causal models, normalizing flows, optimal transport

## 1 Introduction

In many fields such as medicine, economics and public policy, decision-makers need to predict outcomes under different actions. Common examples include estimating how a higher drug dose would affect a patient’s recovery; forecasting today’s inflation if last year’s interest rates had been higher; or inferring the effect of tax relief on poverty levels. This gap between observed data and counterfactual scenarios lies at the heart of identification in causal inference: we see outcomes under one (observational) regime, but want to know what would have happened under another (counterfactual) regime.

Over recent decades, two complementary frameworks have provided principled ways to identify causal quantities from observables. The *potential outcomes* framework (Rubin, 1974) posits latent counterfactual variables whose statistical links to observed outcomes encode causal assumptions. The *causal graphical model* framework (Pearl, 2009a; Spirtes

(et al., 2000) represents those assumptions in a directed acyclic graph and derives identification by  $d$ -separation and the *do*-calculus. Both formalisms yield identification results for average treatment effects, marginal counterfactual distributions, and many other causal targets, by expressing them as functionals of the observed data distribution.

Yet a fundamentally harder class of causal targets remains: those that demand an explicit joint distribution across counterfactual outcomes, known as a *counterfactual coupling*. To understand the distinction, consider estimating the effect of a binary treatment  $X \in \{0, 1\}$  on outcomes  $Y := (Y_1, \dots, Y_p) \in \mathbb{R}^p$ . The *average treatment effect* (ATE) is

$$\text{ATE} := \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)].$$

where  $Y(x)$  is the potential outcome under treatment state  $x \in \{0, 1\}$ . Since this factorizes into treated and control means, in a randomized controlled trial (RCT) it can be estimated by sample averages. Now, consider instead the *treatment harm rate* (THR) (Shen et al., 2013), which measures the probability that treatment worsens outcomes:

$$\text{THR} := \mathbb{P}(Y(1) \preceq Y(0)) = \mathbb{E}[\mathbf{1}\{Y(1) \preceq Y(0)\}]. \quad (1)$$

Here  $\mathbf{1}\{Y(1) \preceq Y(0)\} = 1$  if  $Y_j(1) \leq Y_j(0)$  for some index  $j \in \{1, \dots, p\}$ , and 0 otherwise. The THR is an important criterion for assessing whether a treatment passes the “*first, do no harm*” principle in medical science (Young et al., 2015). In contrast to the ATE, the THR cannot be recovered from the marginal distributions  $\mathbb{P}_{Y(1)}$  and  $\mathbb{P}_{Y(0)}$ , instead requiring an expectation under the *joint distribution*  $\mathbb{P}_{Y(1), Y(0)}$ , which we call the *counterfactual coupling*. Unfortunately, one cannot identify the counterfactual coupling from the marginals alone, as infinitely many joint distributions can admit the same marginal distributions (Villani, 2021). It also cannot be directly estimated from data, since at most one of  $Y(1)$ ,  $Y(0)$  are ever observed per sample or unit. Hence, additional modeling assumptions are needed. The THR is just one motivation for the need for counterfactual couplings. Other situations include individualized treatment and decision-making (Imai et al., 2010), algorithmic fairness (Kusner et al., 2017), and other treatment risk assessments (Kallus, 2023).

A predominant approach to identifying and estimating counterfactual couplings is to use the framework of *Structural Causal Models (SCMs)* (Pearl, 2009a; Peters et al., 2017). In the present setting with no confounding, the idea is to posit a structural model

$$Y =_{\text{a.s.}} f(X, \xi), \quad \xi \sim \mathbb{P}_\xi, \quad \xi \perp\!\!\!\perp X,$$

where  $\xi$  captures all unobserved factors affecting  $Y$  (in practice, one may augment  $X$  to include measured covariates  $Z$ ). The coupling is then characterized by  $f$  and  $\mathbb{P}_\xi$  as

$$(Y(1), Y(0)) =_{\text{a.s.}} (f(1, \xi), f(0, \xi)), \quad \xi \sim \mathbb{P}_\xi.$$

To ensure the pair  $(f, \mathbb{P}_\xi)$  can be identified up to model-specific automorphisms, present theory requires  $f(x, \cdot)$  to be *bijective* (Xi and Bloem-Reddy, 2023; Javaloy et al., 2023). An established, state-of-the-art approach to modeling bijective SCMs is to use *causal normalizing flows* (or flow-based SCMs) (Pawlowski et al., 2020; Khemakhem et al., 2021; Nasr-Esfahany et al., 2023; Javaloy et al., 2023). This involves specifying a simple base distribution (e.g.,  $\widehat{\mathbb{P}}_\xi = N(0, I)$ ) and learning  $f$  using flexible classes of conditional diffeomorphisms parameterized by deep neural networks (Papamakarios et al., 2021). However,

as we establish in Section 2, this approach is brittle to the choice of base distribution. For instance, if the tails or support of this distribution are mis-specified, the true flow  $f$  can be extremely complex (Jaini et al., 2020), may not exist (Cornish et al., 2020), and the resulting estimator  $\hat{f}$  can fail to converge (see e.g., Example 1 in Section 2.1). Such limitations are known in the normalizing flows literature, but existing solutions either fail to fully address the problems, or sacrifice bijectivity of  $f$ , losing any identifiability guarantees.

A recent line of work has instead turned to *optimal transport* (OT) methods (Charpentier et al., 2023; De Lara et al., 2024; Balakrishnan et al., 2025), appealing to the notion that counterfactual worlds should be as similar as possible while satisfying the desired change (Lewis, 1973). The basic idea is to specify transports between counterfactuals—e.g.,

$$Y(1) = T_{1,0}(Y(0)) ,$$

and estimate them by minimizing a transport cost while preserving the marginal distributions. This approach avoids specifying latent noise distributions and guarantees identifiability for continuous variables (Villani, 2021). However, when there are more than two treatment values (e.g.,  $x \in \{0, 1, 2\}$ ), there are multiple ways to couple the counterfactuals (e.g.,  $Y(0) \mapsto (Y(1), Y(2))$  or  $Y(0) \mapsto Y(1) \mapsto Y(2)$ ). Since OT maps are not generally closed under composition (i.e.,  $T_{2,0} \neq T_{2,1} \circ T_{1,0}$ ), each path may induce a different coupling, leading to model incoherence and an identifiability problem (see Example 2 in Section 2.2).

## 1.1 Our Contributions

In this work, we propose a transport-based framework for modeling counterfactual couplings that is free of the noise mis-specification biases of SCMs and incoherence of OT methods. Below we summarize the main contributions and plan of the paper.

In Section 2, we cover the necessary background and demonstrate the limitations of SCMs and OT in detail. These limitations serve as motivation for our approach.

In Section 3, we develop our framework in a simplified confounding-free setting. We start by asking what properties a set of transports  $\{T_{x,x'}\}_{x,x' \in \mathbb{X}}$  must satisfy to be able to couple a set of counterfactuals (a.k.a. potential outcomes) under different treatment levels:

$$Y(x) =_{\text{a.s.}} T_{x,x'}(Y(x')) \quad \forall x, x' \in \mathbb{X} . \quad (2)$$

It turns out that the necessary and sufficient properties are precisely

$$T_{x,x} = \text{id} \quad (\text{Identity}), \quad \text{and} \quad T_{x'',x'} \circ T_{x',x} = T_{x'',x} \quad (\text{Path Independence}) , \quad (3)$$

on the support of the counterfactuals, along with a marginal-matching property (Theorem 2). These properties make the function  $T : (x, x', y) \mapsto T_{x,x'}(y)$  a *cocycle* in dynamical systems theory (Arnold, 1998). We hence call (2) a *counterfactual cocycle* model. The path independence property is the key ingredient missing from OT methods.

In Theorem 3, we show that any counterfactual cocycle can be expressed using a family of injective functions  $(f_x)_{x \in \mathbb{X}}$  with left inverses  $(f_x^+)_{x \in \mathbb{X}}$ , in the sense that for each  $x$  and  $x'$ ,

$$T_{x,x'} = f_x \circ f_{x'}^+ \quad \text{with probability one under } \mathbb{P}_{Y(x')} .$$

This gives us a general recipe to construct flexible counterfactual cocycles: (i) specify each  $f_x : \mathbb{Y} \rightarrow \mathbb{Y}$  using any parameterized bijection (e.g., a normalizing flow or invertible neural network), and (ii) set  $T(x, x', y') := f_x \circ f_{x'}^{-1}(y')$ . We provide general conditions under

which cocycles are identifiable (Theorem 5) and show how to specify cocycles that enforce these conditions using knowledge of the causal ordering of outcomes (Theorem 6). We conclude by showing in Theorem 7 that, under standard potential outcomes assumptions, each counterfactual cocycle is equivalent to an SCM with independent noise:

$$Y(x) =_{\text{a.s.}} f_x \circ f_{x'}^+(Y(x')) \iff Y =_{\text{a.s.}} f(X, \xi), \quad \xi \perp\!\!\!\perp X, \quad \text{where } f_x = f(x, \bullet).$$

Importantly, the cocycle  $T(x, x', y') = f_x \circ f_{x'}^+(y')$  does not depend on the distribution of the noise  $\mathbb{P}_\xi$ . This suggests a robustness to directly modeling counterfactual cocycles. In particular, a set of bijections  $\mathcal{F}$  is well-specified for the cocycle as long as there exists *some* distribution  $\mathbb{P}_\xi^*$  and functions  $\{\hat{f}_x\}_{x \in \mathbb{X}} \subseteq \mathcal{F}$  such that  $\hat{f}_x(\xi^*) \sim \mathbb{P}_{Y(x)}$  for all  $x$ , where  $\xi^* \sim \mathbb{P}_\xi^*$ . By contrast, using  $\mathcal{F}$  to model an SCM with *fixed* base distribution  $\widehat{\mathbb{P}}_\xi$  requires the same property, but the noise must have distribution  $\widehat{\mathbb{P}}_\xi$ . If  $\widehat{\mathbb{P}}_\xi$  poorly matches the true  $\mathbb{P}_\xi$ , the resulting maps  $\{\hat{f}_x\}_{x \in \mathbb{X}}$  may be highly complex or non-bijective and so may lie outside  $\mathcal{F}$ .

In Section 4, we extend counterfactual cocycle models to a more general setup that can handle confounding, and cover our high-level approach to estimating causal quantities with cocycles. The basic idea is to target a distributional discrepancy of the form

$$\ell(T) = \mathbb{E}_{x, x' \sim \mathbb{P}_X} D(\mathbb{P}_{Y(x)}, (T_{x, x'})_\# \mathbb{P}_{Y(x')})^2, \quad (4)$$

and use the cocycle to compute counterfactuals  $\hat{Y}^{(i)}(x)$  of observations  $\{X^{(i)}, Y^{(i)}\}_{i=1}^n$ . From here, one can empirically estimate causal quantities. For instance, for the THR:

$$\widehat{\text{THR}} = \widehat{\mathbb{P}}(Y(1) \preceq Y(0)) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{\hat{Y}^{(i)}(1) \preceq \hat{Y}^{(i)}(0)\}.$$

In Section 5, we derive a tractable cocycle estimation procedure and show that it is asymptotically equivalent to minimizing (4) (Theorem 8 and Theorem 9). The resulting cocycle estimators are consistent under general conditions (Theorem 10) and consistency is unaffected by the noise distribution of the (true) SCM. (see Theorem 11 and Fig. 7). We also show  $\sqrt{n}$ -consistency for parametric classes under regularity conditions (Theorem 12).

In Section 6, we analyze the robustness of modeling counterfactual cocycles versus SCMs in more detail, by studying an invariance under reparameterizations. Specifically, any bijective SCM can be reparameterized using an invertible transformation  $g$ :

$$Y = f(X, \xi) = f(X, g \circ g^{-1}(\xi)) = f^{(g)}(X, \xi^{(g)})$$

Importantly, although the choice of  $g$  affects  $f^{(g)}$ , the associated cocycle is unchanged:

$$T_{x, x'}^{(g)} = f_x^{(g)} \circ (f_{x'}^{(g)})^{-1} = f_x \circ g \circ g^{-1} \circ f_{x'}^{-1} = T_{x, x'}.$$

Thus, while an SCM implicitly commits to a particular  $g$  through the noise parameterization, one is free to model the cocycle using the parameterization  $f^{(g^*)}$  that has the lowest functional complexity (e.g., Sobolev norm). As we prove in Theorem 13 and show in Examples 3 and 4, this property permits using smaller model classes  $\mathcal{F}$  for the cocycle than for an SCM with a fixed base distribution, while remaining well-specified. We also show our method is robust to the dependence structure of the noise  $\mathbb{P}_\xi$ , unlike recently proposed quantile methods for SCMs (Plečko and Meinshausen, 2020; Machado et al., 2024).

In Section 7, we discuss the implementation details of counterfactual cocycle modeling. In Section 8, we implement counterfactual cocycles in several simulations and a real experiment, demonstrating state-of-the-art performance and robustness to noise assumptions.

## 2 Background and Limitations of Existing Methods

To set the stage, we first present the basic setting of interest and formally introduce counterfactual couplings. We then review existing approaches to modeling couplings and discuss their limitations, which motivate our work.

**Causal Inference and Counterfactuals** Let  $\mathbf{V} := (V_1, \dots, V_d) \in \prod_j^d \mathbb{V}_j \subset \mathbb{R}^d$ , be observed random variables with distribution  $\mathbb{P}_{\mathbf{V}}$ . For most of this work we assume  $\mathbf{V} = (Z, X, Y)$ , where  $X := (X_1, \dots, X_q)$  are treatment variables we wish to manipulate (e.g., doses of different medications),  $Y := (Y_1, \dots, Y_p)$  are outcomes of interest (e.g., patient blood pressure, resting heart rate), and  $Z := (Z_1, \dots, Z_l)$  are pre-treatment covariates (e.g., age, gender). We denote  $Y(x)$  as the counterfactual outcomes under a ('do') intervention that fixes the treatment level to  $x$  (often denoted  $\text{do}(X = x)$ ) (Pearl, 2009a; Chernozhukov et al., 2025). We assume the observed outcomes correspond to the counterfactuals under the received treatment level:  $Y =_{\text{a.s.}} Y(X)$  (i.e., the consistency property (Rubin, 1974)). When dealing with an i.i.d. dataset of observations, we index the samples as  $(\mathbf{V}^{(i)})_{i=1}^n \sim_{\text{iid}} \mathbb{P}_{\mathbf{V}}$ .

**Counterfactual Couplings** In this work, we focus not only on the problem of identifying and estimating the marginal distribution of each counterfactual  $Y(x)$ , but on the harder problem of recovering a joint distribution  $\pi$  over *collections* of counterfactuals, e.g.,

$$\pi_I := \mathcal{L}(\{Y(x)\}_{x \in I}), \quad I \text{ is finite.}$$

Here  $\mathcal{L}(\{Y(x)\}_{x \in I})$  denotes the joint law (distribution) of the variables  $\{Y(x)\}_{x \in I}$ . Such couplings are necessary to identify many distributional effects of treatment. For instance, in medical settings, if  $X \in \{0, 1\}$  is a treatment and  $Y \in \mathbb{R}$  a health outcome, we may wish to quantify the extent of adverse effects caused by treatment via the THR as in (1). In finance, we may wish to compute the Conditional Value at Risk (CVaR) (Rockafellar and Uryasev, 2002) to assess the risk of a binary investment decision (Kallus, 2023):

$$\text{CVaR}_{\alpha}(Y(0) - Y(1)) := \mathbb{E}[Y(0) - Y(1) \mid Y(0) - Y(1) \geq q_{\alpha}],$$

where  $q_{\alpha}$  is the  $\alpha$ -quantile (VaR) of the return differential. In economic policy, one may wish to analyze whether a reform primarily benefits those who would already be well-off under the status quo, by assessing how the policy effect varies with status-quo outcomes:

$$\mu(\alpha) := \mathbb{E}[Y(1) - Y(0) \mid Y(0) = q_{\alpha}].$$

When  $X$  is continuous, the same effect measures can be used to assess the dose effects by replacing  $Y(1) - Y(0)$  with the contrast  $Y(x) - Y(0)$ , or the effect of the current treatment policy via the contrast  $Y(X) - Y(0)$ . Note when there are multiple outcomes these quantities may be computed per dimension, or in aggregate as in (1).

Unfortunately, standard causal inference frameworks for identifying the marginal distribution of each  $Y(x)$ , such as causal graphs (Pearl, 2009b) and potential outcomes (Rubin, 1974), cannot identify counterfactual couplings without further assumptions, since there can be infinitely many couplings  $\pi$  that admit the same marginals over  $\{Y(x)\}_{x \in \mathbb{X}}$ . Even in the fully randomized (unconfounded) setting, we only ever observe at most one counterfactual outcome per unit, so one can never directly learn the coupling from data.

Below we review two popular approaches for identifying counterfactual couplings and their limitations: structural causal models, and optimal transport methods.

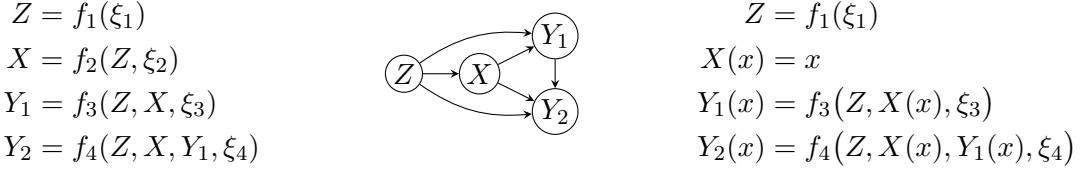


Figure 1: Left: SCM over  $(Z, X, Y_1, Y_2)$ . Middle: corresponding maximal DAG. Right: modified SCM after hard intervention  $\text{do}(X = x)$ . Equalities hold almost surely.

## 2.1 Structural Causal Models

Structural Causal Models (SCMs) (Spirtes et al., 2000; Pearl, 2009a) specify the full causal mechanism on  $\mathbf{V} = (V_1, \dots, V_d)$  using a set of independent (exogenous) noise variables  $\boldsymbol{\xi} := (\xi_1, \dots, \xi_d) \in \mathcal{E}$  with distribution  $\mathbb{P}_{\boldsymbol{\xi}} := \prod_{j=1}^d \mathbb{P}_{\xi_j}$  and a structural map  $F : \mathbb{V} \times \mathcal{E} \rightarrow \mathbb{V}$ ,

$$\mathbf{V} =_{\text{a.s.}} F(\mathbf{V}, \boldsymbol{\xi}). \quad (5)$$

For identifiability, one usually assumes the variables  $\mathbf{V}$  admit a known *causal ordering*  $V_1 \prec \dots \prec V_d$ . This concept, dating back to Simon (1953), formalizes the idea that  $V_i$  may cause  $V_j$  but not vice versa when  $i < j$ . Here we assume the variables are already ordered, so no permutation is necessary. The ordering is used to restrict the SCM to be *acyclic*:

$$V_j =_{\text{a.s.}} f_j(V_{<j}, \xi_j), \quad \forall j \in \{1, \dots, d\}. \quad (6)$$

Here  $f_j : \mathbb{V}_{<j} \times \mathcal{E}_j \rightarrow \mathbb{V}_j$  is the  $j^{\text{th}}$  coordinate of  $F$  that determines how  $V_j$  depends on its predecessors and the noise  $\xi_j$ . In this case we can write  $\mathbf{V} = F(\boldsymbol{\xi})$  with  $F$  lower-triangular.

Every acyclic SCM can be associated with a causal directed acyclic graph (DAG)  $\mathcal{G} = (\mathbf{V}, E)$ , with nodes  $\{V_1, \dots, V_d\}$  and edges  $E = \{V_i \rightarrow V_j : i < j\}$  encoding all possible direct effects consistent with the ordering. This *maximal DAG* includes every forward edge allowed by the ordering. However, since each  $f_j$  may not necessarily depend on all  $V_1, \dots, V_{j-1}$ , many sparser DAGs may also be consistent with the acyclic SCM. In practice, when a specific DAG is known, the SCM can be explicitly restricted to

$$V_j =_{\text{a.s.}} f_j(V_{\text{pa}(j)}, \xi_j),$$

where  $V_{\text{pa}(j)} \subseteq V_{<j}$  are the ‘parents’ of  $V_j$  in the causal DAG, denoted by the index function  $\text{pa}(\bullet)$  which satisfies  $j \in \text{pa}(i)$  if and only if  $V_j \rightarrow V_i$  in  $\mathcal{G}$  (Pearl, 2009a; Peters et al., 2017).

The counterfactuals under a do-intervention  $\text{do}(X = x)$  for any subset  $X \subset \mathbf{V}$  are determined by a modified SCM  $\mathbf{V}(x) =_{\text{a.s.}} F_x(\boldsymbol{\xi})$ , which sets the coordinate functions for  $X$  as  $X(x) = x$ . Fig. 1 shows an example SCM, corresponding maximal causal DAG, and SCM modification for a four variable example where  $Z \prec X \prec Y_1 \prec Y_2$  and we set  $\text{do}(X = x)$ .

Given an SCM, statistical functionals of any counterfactuals (e.g.,  $(\mathbf{V}(x), \mathbf{V}(x'))$ ) conditioned on any subset of observed variables  $\mathbf{W} \subset \mathbf{V}$  can be estimated by the familiar three-step abduct-act-predict recipe (Pearl, 2009a):

**Abduct.** Update noise distribution to condition on evidence:  $\widehat{\mathbb{P}}_{\boldsymbol{\xi}} \leftarrow \mathbb{P}_{\boldsymbol{\xi}|\mathbf{W}=\mathbf{w}}$ .

**Act.** Modify structural equations to  $F_x, F_{x'}$ .

**Predict.** Propagate noise:  $\mathbb{E}[h(\mathbf{V}(x), \mathbf{V}(x')) \mid \mathbf{W} = \mathbf{w}] = \mathbb{E}_{\boldsymbol{\xi} \sim \widehat{\mathbb{P}}_{\boldsymbol{\xi}}} [h(F_x(\boldsymbol{\xi}), F_{x'}(\boldsymbol{\xi}))]$ .

Table 1: Autoregressive normalizing flows, their transforms, and conditions for Lipschitz continuity.  $M$  is a rational-quadratic spline (Gregory and Delbourgo, 1982).

Model	Autoregressive transform $f_j(v_{<j}, \xi_j)$	Lipschitz Condition
NICE (Dinh et al., 2014)	$\xi_j + \mu_j(v_{<j})1_{k \notin [j]}$	$\mu_j$ Lipschitz
MAF (Papamakarios et al., 2017)	$\xi_j \mapsto \sigma_j(v_{<j}) \cdot \xi_j + (1 - \sigma_j(v_{<j}))\mu_j(v_{<j})$	$\sigma_j$ bounded
IAF (Kingma et al., 2016)	$\xi_j \mapsto \exp(\lambda_j(v_{<j})) \xi_j + \mu_j(v_{<j})$	$\lambda_j$ bounded, $\mu_j$ Lipschitz
Real-NVP (Dinh et al., 2016)	$\xi_j \mapsto \exp(\lambda_j(v_{<j})1_{k \notin [j]}) \xi_j + \mu_j(v_{<j})1_{k \notin [j]}$	$\lambda_j$ bounded, $\mu_j$ Lipschitz
Glow (Kingma and Dhariwal, 2018)	$\xi_j \mapsto \sigma_j(v_{<j})\xi_j + \mu_j(v_{<j})1_{k \notin [j]}$	$\sigma_j$ bounded, $\mu_j$ Lipschitz
NAF (Huang et al., 2018)	$\xi_j \mapsto \sigma^{-1}(w(v_{<j}) \cdot \sigma(\sigma_j(v_{<j})\xi_j + \mu_j(v_{<j})))$	Always (logistic mixture CDF)
NSF (Durkan et al., 2019)	$\xi_j \mapsto v_j 1_{v_j \notin [-B, B]} + M_j(\xi_j; v_{<j})1_{v_j \in [-B, B]}$	Always (linear outside $[-B, B]$ )

**Identifiability via Bijective Causal Models** Various classes of SCMs have been proposed in recent years, primarily relying on the flexibility of deep neural networks (Pawlowski et al., 2020; Sanchez-Martin et al., 2021; Khemakhem et al., 2021; Geffner et al., 2022; Nasr-Esfahany et al., 2023; Javaloy et al., 2023). The most flexible models with identifiability guarantees for counterfactual couplings are *bijective causal models (BCMs)*, which constrain each  $f_j$  to be bijective on the noise  $\xi_j$ . If two BCMs  $(F, \mathbb{P}_\xi)$ ,  $(\tilde{F}, \tilde{\mathbb{P}}_\xi)$  induce the same distribution  $\mathbb{P}_V$ , and each structural function  $f_j$  and  $\tilde{f}_j$  is monotone increasing on  $\xi_j$ , they produce the same counterfactual couplings (Nasr-Esfahany et al., 2023). This functional restriction subsumes many traditional SCM classes such as additive-noise models ( $V_i = f(V_{\text{pa}(i)}) + \xi_i$ ) and location-scale models ( $V_i = f(V_{\text{pa}(i)}) + h(V_{\text{pa}(i)})\xi_i$ ). BCMs are also natural counterparts to our proposed methods, and so we focus on them within the broader SCM framework.

A popular and state-of-the-art (SOTA) modeling approach for flexible BCMs is to use *normalizing flows* (Khemakhem et al., 2021; Javaloy et al., 2023; Nasr-Esfahany et al., 2023), which specify  $F$  as a sequence of invertible and differentiable transformations from a simple base distribution (e.g.,  $\hat{\mathbb{P}}_\xi = \mathcal{N}(0, I)$ ) (Papamakarios et al., 2021). As discussed above, the causal order restricts  $F$  to be lower-triangular, for which *autoregressive flows* are used (Table 1). Such flows can be composed for increased expressiveness while respecting the causal ordering and enabling fast and exact maximum likelihood training. We refer to such BCMs as *flow-based SCMs*.

While flow-based SCMs have achieved SOTA performance on causal inference tasks, they suffer from a key practical limitation: if the base distribution  $\hat{\mathbb{P}}_\xi$  poorly matches the target  $\mathbb{P}_V$ , the required flow may be extremely complex or may not even exist. Although these problems have been recognized in the normalizing flows literature, their effects on flow-based SCMs have not been systematically studied. Alternative estimation approaches for BCMs have been proposed based on quantile methods (Plečko and Meinshausen, 2020; Machado et al., 2024). However, they too rely on a fixed noise distribution and can therefore suffer from related mis-specification problems, as discussed in Section 6.4. Below we focus on two primary problems that occur in flow-based SCMs: tail and support mis-specification.

**Tail Mis-specification** It is known that if the base distribution  $\hat{\mathbb{P}}_\xi$  lies in a different tail class to the observational distribution  $\mathbb{P}_V$  (e.g., exponential vs. logarithmic tail decay), then no regular class of bi-Lipschitz flows can match the tails of  $\mathbb{P}_V$  (e.g., see Theorem 3.2 and Corollary 3.3 in Liang et al., 2022 and Theorem 3 in Jaini et al., 2020). This is problematic, since most flows are bi-Lipschitz, in some cases by design (see Table 1). Recent

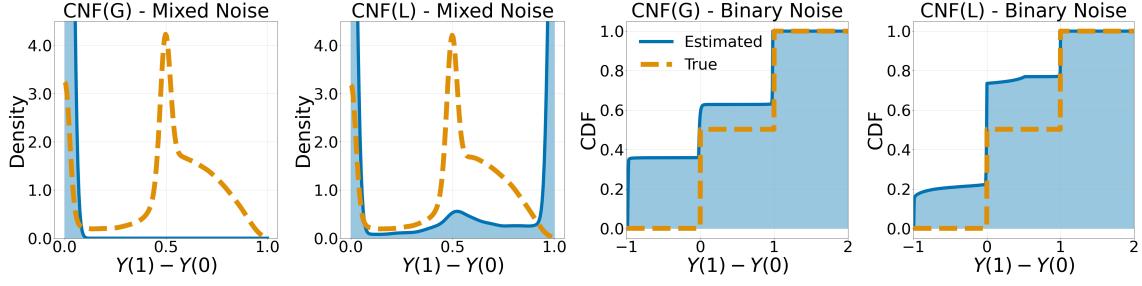


Figure 2: True (orange) vs. estimated (blue) distribution of the treatment effect  $Y(1) - Y(0)$  for the SCM  $Y(x) = (x + 1)\xi$ , under two different noise laws: (LHS) mixed-tailed  $\xi$  (density plots) and (RHS) binary  $\xi$  (CDF plots). Estimates are obtained with flow-based SCMs using Gaussian (CNF(G)) and Laplace (CNF(L)) base distributions  $\hat{\mathbb{P}}_\xi$ , each trained on  $n = 2000$  samples. For the mixed-tailed case,  $Y(1) - Y(0)$  is transformed to  $[0, 1]$  via  $\sigma(y) = 1/(1 - e^{-y})$  for visualization. Both estimated flows fail to recover the effect distributions.

work in the normalizing flows literature has mitigated the problem using tail-adaptive base distributions (Jaini et al., 2020; Liang et al., 2022; Laszkiewicz et al., 2022). However, those approaches still impose isotropic tails per dimension and so can fail to learn distributions where e.g., some  $V_j$  has a heavier right tail than left tail, which occurs frequently in finance and climate applications (Verhoeven and McAleer, 2004; Deidda et al., 2023). This mis-specification can also result in estimation pathologies due to high-variance log-likelihood gradients (Amiri et al., 2024). In the extreme case that  $\mathbb{E}\|V\|_2 = \infty$  and  $\hat{\mathbb{P}}_\xi = \mathcal{N}(\mathbf{0}, I)$  is chosen, the gradient may even be infinite and so the minimizer  $F_{\theta^*}$  can diverge as  $n \rightarrow \infty$ .

**Support Mis-specification** If there is no differentiable function that transports  $\hat{\mathbb{P}}_\xi$  to  $\mathbb{P}_V$ , then no normalizing flow can map between them (since normalizing flows are diffeomorphisms). This arises whenever a continuous base is chosen with  $\text{supp}(\hat{\mathbb{P}}_\xi) = \mathbb{R}^d$ , while the true BCM noise distribution has disconnected support (e.g.,  $\xi_j \sim \alpha\mathbb{P}_0 + (1 - \alpha)\mathbb{P}_1$ ) or lies on a lower-dimensional manifold (e.g., discrete  $\xi_j$ ). Such noise structures have been assumed in medical settings, where patients with latent discrete characteristics may respond differently to treatment (Yin et al., 2018; Loh and Kim, 2022). In the normalizing flows literature, solutions to this relax the bijectivity of  $F$  (Tanielian et al., 2020; Cornish et al., 2020). However, this would sacrifice any identifiability guarantees for counterfactual inference. Additionally, this mis-specification can also lead to estimation identifiability problems. The expected log-likelihood depends only on the inverse flow and its derivatives over the support of  $\mathbb{P}_V$ . As a result, if the image of  $F^{-1}(\mathbf{V})$  does not cover  $\text{supp}(\hat{\mathbb{P}}_\xi)$ , the solution may be undetermined and different minimizers can yield different distributions.

**Example 1.** To demonstrate the severity of these problems, consider the structural model,

$$Y(x) = (x + 1)\xi \quad , \quad X \sim \text{Bern}(1/2)$$

Suppose the estimation target is the distribution of the treatment effect  $Y(1) - Y(0) = \xi$  and the noise has a heavy left tail and light right tail  $\xi \sim \frac{1}{2}|\mathcal{N}(0, 1)| - \frac{1}{2}|\text{NBP}(0.1, 0.1)|$ .<sup>1</sup>

1. Here,  $\text{NBP}(\alpha, \beta)$  is the heavy-tailed Normal-Beta-Prime distribution on  $\mathbb{R}$  (Bai and Ghosh, 2021).

Fig. 2 (left + middle left) shows the true density of  $Y(1) - Y(0)$ , rescaled to  $[0, 1]$  via  $\sigma(y) = 1/(1+\exp(-y))$  for better visualization, alongside estimated densities via flow-based SCMs with Gaussian (left) and Laplace (middle left) base distributions. Since  $x \in \{0, 1\}$  is binary, we specify separate flows  $f_x$  for each  $x \in \{0, 1\}$ , rather than parameter sharing across  $x$ . Each flow architecture was specified as a Neural Spline Flow (NSF) (Durkan et al., 2019) with a single (8-bin) spline layer sandwiched between two affine layers. The flow parameters were trained by maximum likelihood on  $n = 2000$  samples  $\{X^{(i)}, Y^{(i)}\}_{i=1}^n$  and the exact training and optimization routine matches that used in Section 8.1. In order to match the heavy left tail, both estimated flows move most probability mass there, but as a consequence fail to recover the rest of the density. Note the large mass *near* zero of the estimated densities under the transformation  $\sigma$  does not indicate a learned heavy left tail.

We next re-implement the same flow-based models with binary noise,  $\xi \sim \text{Rad}(1/2)$ , this time showing the learned CDFs (Fig. 2, middle right and right). While we acknowledge flows are not designed for discrete outcomes  $Y$ , the example serves to highlight the problems that arise when the support of  $\mathbb{P}_{Y(x)}$  lies on a lower-dimensional manifold than that of  $\mathbb{P}_\xi$ . Since each model only evaluates the likelihood on a set of measure zero under its base distribution, there are infinitely many flows that can achieve the same likelihood. As a consequence, both implemented models fail to learn the size and placement of the modes of the distribution (as indicated by the jumps in the learned vs. true CDF).

## 2.2 Optimal Transport Methods for Counterfactual Inference

Several recent works estimate counterfactual couplings with *optimal transport* (OT) (Charpentier et al., 2023; De Lara et al., 2024; Balakrishnan et al., 2025), providing an alternative to SCMs, albeit in a restricted setting. Given discrete treatments  $X$  and continuous outcomes  $Y$ , the idea is to model the coupling  $\mathcal{L}(\{Y(x)\}_{x \in \mathbb{X}})$  via a collection of transports  $\{T_{x,x'}\}_{x \in \mathbb{X}}$  that model deterministic counterfactuals under treatment changes,

$$Y(x) =_{\text{a.s.}} T_{x,x'}(Y(x')) , \quad x, x' \in \mathbb{X} . \quad (7)$$

Since many sets of transports can give rise to the same marginal counterfactual distributions, the transports are identified via the principle of Lewis (1973): out of all couplings, choose that which induces the most “similar” counterfactual worlds. This is formalized by choosing  $T_{x,x'}$  to minimize the cost of transporting mass from  $\mathbb{P}_{Y(x')}$  to  $\mathbb{P}_{Y(x)}$ ,

$$T_{x,x'}^* = \arg \min_{T_{x,x'}: T_{x,x'}(Y(x')) =_d Y(x)} \mathbb{E} [\|Y(x) - T_{x,x'}(Y(x'))\|_2^2] \quad (8)$$

Since quadratic-cost OT problems between continuous distributions have a unique solution (known as the Brenier map) (Villani, 2021), this guarantees identifiability when  $\mathbb{P}_{Y(x)}$  is absolutely continuous for all  $x$ . This approach is appealing, as it avoids the need to specify a full generative process (e.g., noise distributions) or causal ordering when there are multiple outcomes  $Y := (Y_1, \dots, Y_p)$ . Under no confounding, one has  $\mathbb{P}_{Y(x)} = \mathbb{P}_{Y|X=x}$ . In practice, given i.i.d. data  $\{Y^{(i)}(x)\}_{i=1}^n \sim \mathbb{P}_{Y|X=x}$ , one can plug the empirical analogues  $\widehat{\mathbb{P}}_{Y(x)} = \frac{1}{n_x} \sum_{i=1}^{n_x} \delta_{Y^{(i)}(x)}$  of the conditionals into (8) and solve the resulting problem (either in closed form or using specialized solvers). When there are measured confounders  $Z$ , the transports are specified between  $\{Y(x, z)\}_{x,z \in \mathbb{X} \times \mathbb{Z}}$  instead (Balakrishnan et al., 2025).

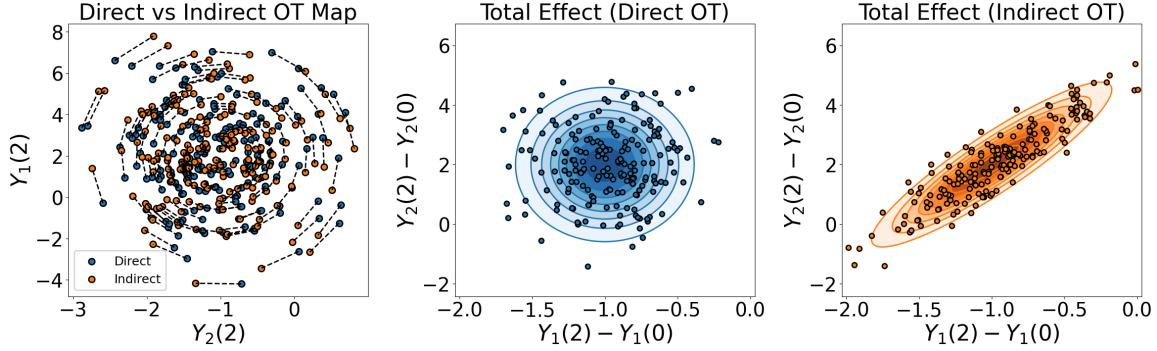


Figure 3: Transport of 200 samples  $Y(0) \sim \mathbb{P}_0$  to  $\mathbb{P}_1$  and  $\mathbb{P}_2$  using OT maps. Blue: direct route  $(Y(1), Y(2)) = (T_{1,0}(Y(0)), T_{2,0}(Y(0)))$ . Orange: indirect route  $(T_{1,0}(Y(0)), T_{2,1} \circ T_{1,0}(Y(0)))$ . Left: Imputed  $Y(2)$  under each method; dashed lines highlight the inconsistency. Middle & right: Density of  $Y(2) - Y(0)$  depends on which maps are used (direct vs. indirect), so cannot be identified.

When  $Y$  is scalar, the Brenier map is the quantile transform (Santambrogio, 2015) and can be estimated via the CDF (Balakrishnan et al., 2025). However, when  $Y$  is multivariate, one needs to solve the OT problem separately for each pair  $(x, x')$ , typically using a specialized solver, which makes it challenging to handle continuous treatments and covariates.

A more serious issue overlooked in the literature is that, except in special cases, Brenier maps will not yield a valid joint coupling when  $\dim(\mathbb{Y}) > 1$  and  $|\mathbb{X}| > 2$ . The solution to (8) is the gradient of a convex function,  $T_{x,x'}^* = \nabla \varphi_{x,x'}$  (Villani, 2021), but the composition of two such maps is generally not the gradient of a convex function when  $\dim(\mathbb{Y}) > 1$  (Santambrogio, 2015). Consequently, if  $T_{x'',x}^* \neq T_{x'',x'}^* \circ T_{x',x}^*$  on a set of positive  $\mathbb{P}_{Y(x)}$ -measure, the transports cannot satisfy (7) for any set of counterfactuals with the same marginal distributions. Below we demonstrate this incompatibility and resulting identifiability issues.

**Example 2.** Consider solving the quadratic-cost OT problem for transporting between bivariate Gaussian distributions  $\mathbb{P}_0, \mathbb{P}_1, \mathbb{P}_2$  with the following means and covariances.

$$(\mu_0, \Sigma_0) = (\mathbf{0}, I_2), \quad (\mu_1, \Sigma_1) = \left( \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & \frac{-9}{10} \\ \frac{-9}{10} & 1 \end{pmatrix} \right), \quad (\mu_2, \Sigma_2) = \left( \begin{pmatrix} -1 \\ 2 \end{pmatrix}, \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 5 \end{pmatrix} \right)$$

The Brenier map from  $\mathbb{P}_x$  to  $\mathbb{P}_{x'}$  is the affine map  $T_{x',x}(y) = \mu_{x'} + A_{x',x}(y - \mu_x)$ , with

$$A_{x',x} = \Sigma_x^{-1/2} (\Sigma_x^{1/2} \Sigma_{x'} \Sigma_x^{1/2})^{1/2} \Sigma_x^{-1/2}, \quad x, x' \in \{0, 1, 2\}.$$

However, as  $\Sigma_1$  and  $\Sigma_2$  do not commute, the maps cannot be valid counterfactual transports in the sense of (7), since using them to impute counterfactuals results in a contradiction:

$$Y(2) - Y(2) =_{\text{a.s.}} T_{2,0}(Y(0)) - T_{2,1} \circ T_{1,0}(Y(0)) = \underbrace{(\Sigma_2^{\frac{1}{2}} - A_{2,1}\Sigma_1^{\frac{1}{2}})(Y(0) - \mu_0)}_{\neq 0} \neq_{\text{a.s.}} 0.$$

To illustrate the identifiability issues that arise from this, suppose we draw  $n = 200$  samples  $\{Y^{(i)}(0)\}_{i=1}^n \sim \mathbb{P}_0$  and impute their counterfactuals  $\{Y^{(i)}(1), Y^{(i)}(2)\}_{i=1}^n$ . There are two

natural ways of using the OT maps to do this: (i) *directly*,  $y_0 \mapsto (T_{1,0}(y_0), T_{2,0}(y_0))$ ; and (ii) *indirectly*,  $y_0 \mapsto (T_{1,0}(y_0), T_{2,1} \circ T_{1,0}(y_0))$ . Each approach uses two out of three OT maps to induce a coupling over  $\{Y(0), Y(1), Y(2)\}$ . Fig. 3 shows that the direct map  $T_{0,2}$  and the composition  $T_{1,2} \circ T_{0,1}$  result in different counterfactual predictions for  $Y(2)$  (left) and different distributions of the total effect  $Y(2) - Y(0)$  (middle and right). Thus, the induced coupling depends on the subset of transports used, and is not identifiable from these maps.

### 3 Transport-based Counterfactual Couplings in a Simplified Setting

Our goal is to develop a framework for modeling counterfactual couplings that avoids the limitations of SCMs and OT methods. The basic idea is to directly model counterfactual transports (as in OT), but in a way that guarantees coherence and identifiability under general conditions (as in SCMs). To that end, in this section we analyze the algebraic properties that a set of transports must satisfy to induce a valid coupling between counterfactual outcomes, in a simplified confounding-free setting. We show that these properties are precisely those of a *cocycle*, and so call any system of transports satisfying them a counterfactual cocycle. We then show that any counterfactual cocycle can be represented using a family of injective functions. This lets us characterize conditions for cocycle identifiability and provide a general recipe for parameterizing flexible, identifiable cocycle classes. Finally, we show that every counterfactual cocycle corresponds to an equivalence class of injective SCMs. This suggests that modeling counterfactual cocycles can avoid the mis-specification problems of SCMs, and underpins the estimation approach we later develop.

#### 3.1 Counterfactual Cocycles as Admissible Transports

We focus on a simple setting where  $X \in \mathbb{X} \subset \mathbb{R}$  is a randomized treatment,  $Y := (Y_1, \dots, Y_p) \in \mathbb{Y} \subset \mathbb{R}^p$  are a set of outcomes of interest, and  $\{Y(x)\}_{x \in \mathbb{X}} \in \mathbb{Y}$  are counterfactuals under different treatment levels. Throughout, we assume all spaces (here  $\mathbb{X}, \mathbb{Y}$ ) are standard Borel and all maps are Borel measurable. To formalize counterfactuals in this setting, we work under the standard potential outcome assumptions (Rubin, 1974).

**Assumption 1.** 1. *Consistency:*  $Y =_{\text{a.s.}} Y(X)$

2. *Exchangeability:*  $\{Y(x)\}_{x \in \mathbb{X}} \perp\!\!\!\perp X$

Assumption 1 is consistent with a “coarse-grained” causal DAG  $X \rightarrow Y$  (Richardson and Robins, 2013) and identifies the marginal counterfactual distribution as  $Y(x) \sim \mathbb{P}_{Y|X=x}$ .

To recover a coupling over  $\{Y(x)\}_{x \in \mathbb{X}}$ , we follow previous transport-based approaches and start from the existence of a collection of transport maps between counterfactual outcomes under different treatment levels.

**Assumption 2.** There exist a collection of (Borel measurable) transport maps  $\{T_{x,x'} : \mathbb{Y} \rightarrow \mathbb{Y} \mid x, x' \in \mathbb{X}\}$  that satisfy the **Counterfactual Coupling (CC)** property:

$$Y(x) =_{\text{a.s.}} T_{x,x'}(Y(x')) , \quad \text{for all } x, x' \in \mathbb{X} . \quad (\text{CC})$$

In contrast to previous transport-based methods, our approach to modeling the transports is motivated by the insight that, as is easily checked, (CC) can only hold if the transports satisfy the following properties:

1. **Identity:** For each  $x \in \mathbb{X}$ ,

$$T_{x,x} = \text{id}, \quad \mathbb{P}_{Y|X=x}\text{-a.s.} \quad (\text{ID})$$

2. **Path Independence:** For each  $x, x', x'' \in \mathbb{X}$ ,

$$T_{x'',x'} \circ T_{x',x} = T_{x'',x}, \quad \mathbb{P}_{Y|X=x}\text{-a.s.} \quad (\text{PI})$$

3. **Distribution Adaptedness:** For each  $x, x' \in \mathbb{X}$ ,

$$(T_{x',x})_\# \mathbb{P}_{Y|X=x} = \mathbb{P}_{Y|X=x'} \quad (\text{DA})$$

Properties (ID) and (PI) together make the map

$$T : (x, x', y) \mapsto T_{x,x'}(y),$$

a **cocycle** (Arnold, 1998). In classical dynamics, a cocycle describes how a dynamical system state evolves as time flows. In the present context, “time” is replaced by the treatment value  $x$ , the “state” is the outcome value  $y(x)$ , and the cocycle  $T$  encodes how  $y(x)$  changes as treatment values change  $x \rightarrow x'$ . If a cocycle also satisfies (DA) we call it a  $\mathbb{P}_{Y|X}$ -adapted **cocycle**. Lastly, if a cocycle  $T : \mathbb{X}^2 \times \mathbb{Y} \rightarrow \mathbb{Y}$  satisfies (CC) then we call it a **counterfactual cocycle**. We thus also refer to (CC) as the *counterfactual cocycle property*.

**Importance of Cocycle Properties** (DA) alone ensures that (CC) holds in *distribution*,

$$Y(x) =_d T_{x,x'}(Y(x')).$$

However, without (ID) and (PI), the transports  $T_{x,x'}$  cannot describe a valid coupling between a set of counterfactuals as in (CC). For example, if (PI) fails and we try to impose (CC), we can arrive at logical impossibilities in the counterfactual outcomes:

$$Y(x_2) =_{\text{a.s.}} T_{x_2,x_0}(Y(x_0)) \neq_{\text{a.s.}} T_{x_2,x_1} \circ T_{x_1,x_0}(Y(x_0)) =_{\text{a.s.}} Y(x_2).$$

Fig. 4 illustrates this failure. This issue arises in recent transport-based models (De Lara et al., 2024; Charpentier et al., 2023; Torous et al., 2024; Balakrishnan et al., 2025), where  $T_{x,x'}$  is defined as the OT map from  $\mathbb{P}_{Y|X=x}$  to  $\mathbb{P}_{Y|X=x'}$ . These maps satisfy (ID), (DA) and an invertibility property ( $T_{x,x'} \circ T_{x',x}(Y(x)) =_{\text{a.s.}} Y(x)$ ), but, as demonstrated in Section 2.2 can fail (PI) when  $|\mathbb{X}| > 2$  and  $\dim(\mathbb{Y}) > 1$ , leading to identifiability problems.

The properties (ID), (PI), and (DA) are purely mathematical and so are necessary but not sufficient to guarantee (CC)—the latter requires that the counterfactual variables  $\{Y(x)\}_{x \in \mathbb{X}}$  exist on a common probability space and are actually linked by the transports. However, these properties are necessary and sufficient to ensure (CC) holds w.r.t. *some* set of variables  $\{\tilde{Y}(x)\}_{x \in \mathbb{X}}$ . Thus, these properties characterize all sets of transports which can induce admissible couplings over counterfactuals. This is formalized below.

**Definition 1.** A set of transports  $\{T_{x,x'} : \mathbb{Y} \rightarrow \mathbb{Y}\}_{x,x' \in \mathbb{X}}$  are **admissible** w.r.t.  $\mathbb{P}_{Y|X}$  if there exists a collection of random variables  $\{\tilde{Y}(x)\}_{x \in \mathbb{X}}$  such that  $\tilde{Y}(x) \sim \mathbb{P}_{Y|X=x}$  and  $\tilde{Y}(x) =_{\text{a.s.}} T_{x,x'}(\tilde{Y}(x'))$  for every  $x, x' \in \mathbb{X}$ .

**Theorem 2** (Cocycle Equivalence to Admissible Transports).  $\{T_{x,x'} : \mathbb{Y} \rightarrow \mathbb{Y}\}_{x,x' \in \mathbb{X}}$  satisfy (ID), (PI), and (DA) w.r.t.  $\mathbb{P}_{Y|X}$  if and only if they are admissible w.r.t.  $\mathbb{P}_{Y|X}$ .

Our aim is therefore to develop a framework for modeling and estimating counterfactual cocycles. In the rest of this section we focus on the modeling aspects.

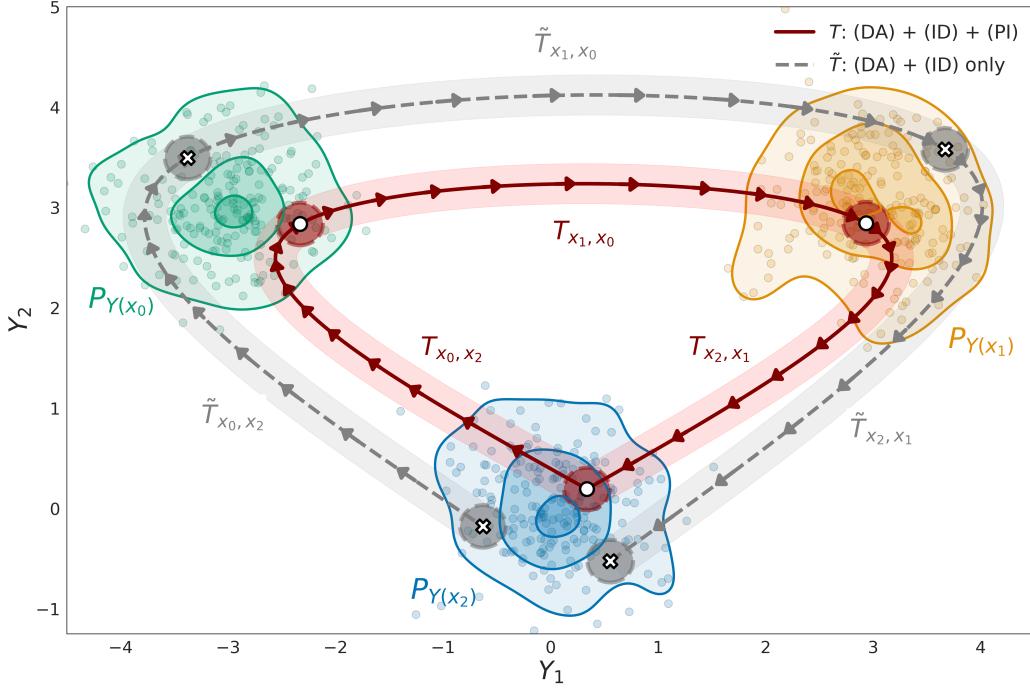


Figure 4: Illustration of cocycle property importance for coherent counterfactual transports. Shaded regions show trajectories of different values of  $Y(x_2) \sim \mathbb{P}_{Y(x_2)}$  as we change the treatment level  $x_2 \rightarrow x_0 \rightarrow x_1 \rightarrow x_2$ , under transport collections  $T$  (maroon) and  $\tilde{T}$  (gray). Lines indicate trajectory for single value  $y(x_2)$  within each region.  $\{T_{x,x'}\}_{x,x' \in \mathbb{X}}$  satisfies (DA) and (ID) but fails (PI), leading to the logical impossibility that  $Y(x_2) \neq Y(x_2)$  on a set of strictly positive measure (see mismatch in gray regions of  $\mathbb{P}_{Y(x_2)}$ ). The transports therefore do not define a valid counterfactual model. In contrast,  $\{T_{x,x'}\}_{x,x' \in \mathbb{X}}$  satisfy all cocycle properties and so cohere with one another, inducing a valid counterfactual model (Theorem 2).

### 3.2 Structure and Identifiability of Counterfactual Cocycles

The previous results tell us that modeling admissible counterfactual transports reduces to targeting the properties (ID), (PI), and (DA). However, it remains unclear how to construct transports that satisfy these properties in general, or under what conditions the resulting system of transports is uniquely determined. Both questions must be answered in order to apply these properties in practice to estimate counterfactual transports.

**Characterizing Cocycles via Factorization** The following result shows that every cocycle has a special structure: it can always be constructed using a family of injective functions. This provides a clear path to modeling transports that satisfy (ID) and (PI).

**Theorem 3** (Cocycle Factorization). *Let  $\{T_{x,x'}\}_{x,x' \in \mathbb{X}}$  satisfy (ID), (PI), and (DA) w.r.t.  $\mathbb{P}_{Y|X}$ . Then there is a set  $\mathbb{Y}_0 \subseteq \mathbb{Y}$  and, for each  $x \in \mathbb{X}$ , a function  $f_x : \mathbb{Y}_0 \rightarrow \mathbb{Y}$  with left-inverse  $f_x^+ : \mathbb{Y} \rightarrow \mathbb{Y}_0$ , such that for every  $x' \in \mathbb{X}$ :*

$$T_{x,x'} = f_x \circ f_{x'}^+, \quad \mathbb{P}_{Y|X=x'}\text{-a.s.} \quad (9)$$

The proof is straightforward: one essentially defines  $f_x := T_{x,x_0}$  and  $f_x^+ := T_{x_0,x}$  for any  $x_0 \in \mathbb{X}$  and applies (ID) and (PI) to verify (9). To ensure  $f_x^+$  is a left inverse, we restrict the functions to map to and from an appropriate subset  $\mathbb{Y}_0 \subset \mathbb{Y}$  of full  $\mathbb{P}_{Y|X=x_0}$ -measure.

Theorem 3 implies that specifying a parameterized family of cocycles reduces to specifying a parameterized family of functions  $f : \mathbb{X} \times \mathbb{Y}_0 \rightarrow \mathbb{Y}$  injective in their second argument,

$$\mathcal{F} \subseteq \mathcal{I} := \{ f : \mathbb{X} \times \mathbb{Y}_0 \rightarrow \mathbb{Y} \mid f_x := f(x, \cdot) \text{ is injective } \forall x \in \mathbb{X} \}.$$

Indeed, any element  $\hat{f} \in \mathcal{F}$  can be used to construct a candidate cocycle via,

$$\hat{T}(x, x', y') := \hat{f}_x \circ \hat{f}_{x'}^+(y'), \quad (10)$$

where  $\hat{f}_x^+$  is a left inverse of  $\hat{f}_x$ . To avoid the need to separately model the left inverse (which is only unique on the image of  $\hat{f}_x$ ) a convenient approach is to take  $\mathbb{Y}_0 := \mathbb{Y}$  and parameterize  $\hat{f}_x : \mathbb{Y} \rightarrow \mathbb{Y}$  using *bijective* maps such as normalizing flows (NFs) (Papamakarios et al., 2021) or invertible neural networks (INNs) (Ishikawa et al., 2023), with parameters conditioned on  $x$ . In this case the left inverse is unique and is an exact inverse  $\hat{f}_x^+ = \hat{f}_x^{-1}$ . As a consequence, the construction (10) is uniquely determined by  $\hat{f}$  and automatically satisfies (ID) and (PI) on all of  $\mathbb{Y}$ :

$$\hat{T}_{x,x} = \hat{f}_x \circ \hat{f}_x^{-1} = \text{id}, \quad \hat{T}_{x,x'} \circ \hat{T}_{x',x''} = \hat{f}_x \circ \hat{f}_{x'}^{-1} \circ \hat{f}_{x'} \circ \hat{f}_{x''}^{-1} = \hat{T}_{x,x''}$$

We note specializing to bijections is not overly restrictive. The constructed  $f_x$  in Theorem 3 is already bijective on  $\mathbb{Y}_0 \rightarrow \mathbb{Y}_x := f_x(\mathbb{Y}_0)$ . When the sets  $\{\mathbb{Y}_x\}_{x \in \mathbb{X}}$  are *Borel-isomorphic* (e.g., full-dimensional Borel subsets of  $\mathbb{Y}^2$ ), classical results guarantee a measurable bijective extension  $\hat{f}_x : \mathbb{Y} \rightarrow \mathbb{Y}$  of  $f_x$  (Kechris, 2012, Thm. 15.6). Although the extension is not always continuous, NFs and INNs remain highly expressive, and will be used to specify cocycle models in Section 7. Cocycles of the form  $T(x, x', y') = f_x \circ f_{x'}^{-1}(y')$  are known as **coboundaries** (Varadarajan, 1968) and so we call the function  $f$  a **coboundary map**.

**General Conditions for Identifiability** Theorem 3 provides a general recipe for parameterizing counterfactual cocycles, but does not tell us how to restrict  $\mathcal{F}$  so that any cocycle constructed by functions in  $\mathcal{F}$  that satisfies (ID), (PI) and (DA) is (almost surely) unique. Indeed, many sets of transports may satisfy these properties, but they can only all satisfy (CC) for the same set of counterfactuals if they are identical up to null sets.

To demonstrate this point, a collection of OT maps  $\{T_{x,x'}^{(OT)}\}_{x \in \mathbb{X}}$  can in principle be used as coboundary maps to construct a cocycle that satisfies (ID), (PI) and (DA). One can simply define  $f_x := T_{x,x_0}^{(OT)}$  and  $f_x := T_{x_0,x}^{(OT)}$  for some  $x_0$ , and define a  $\mathbb{P}_{Y|X}$ -adapted cocycle  $\tilde{T}(x, x', y') := f_x \circ f_{x'}^+(y')$ . However, since OT maps do not necessarily satisfy (PI), the constructed cocycle  $\tilde{T}$  will depend on the reference point  $x_0$ , and so there are potentially  $|\mathbb{X}|$ -many different construction choices. We already saw this in Example 2 and Fig. 3: the direct imputation approach used reference  $x_0 = 0$ , while the indirect approach used  $x_0 = 1$ .

The factorization structure of cocycles enables us to provide general conditions for identifiability when working with bijective cocycles (a.k.a. coboundaries) defined by transformation groups.<sup>3</sup> Formally, a *transformation group*  $\mathbb{G}$  on  $\mathbb{Y}$  is a set of bijective transformations

2. Note, we do *not* require that  $\mathbb{Y}$  is a full-dimensional subset of  $\mathbb{R}^p$ .

3. For simplicity, we characterize identifiability under the assumption that the cocycles can be characterized using bijections, but a nearly identical version holds in the more general injective case using a more complicated algebraic construction involving monoids instead of groups.

$g : \mathbb{Y} \rightarrow \mathbb{Y}$  that contains the identity and is closed under compositions and inverses. We say that a cocycle  $T$  is **G-valued** if its coboundary map  $f$  satisfies  $f(x, \bullet) := f_x \in \mathbb{G}$  for every  $x \in \mathbb{X}$ . For example, any bijective  $f_x : \mathbb{Y} \rightarrow \mathbb{Y}$  belongs to the automorphism group of  $\mathbb{Y}$ , denoted  $\text{Aut}(\mathbb{Y})$ . If  $f_x$  is also bi-continuous, it belongs to the group of self-homeomorphisms,  $\text{Homeo}(\mathbb{Y})$ . If  $f_x$  and  $f_x^{-1}$  also have continuous derivatives (e.g. if modeling each  $f_x$  using typical parameterizations of the flows in Table 1), it lies in the diffeomorphism group,  $\text{Diff}^1(\mathbb{Y})$ . The set of all  $\mathbb{G}$ -valued coboundary maps is denoted

$$\mathcal{F}_{\mathbb{G}} := \{f : \mathbb{X} \times \mathbb{Y} \rightarrow \mathbb{Y} \mid f(x, \bullet) \in \mathbb{G}\}.$$

Using these concepts, we have the following formal notion of cocycle identifiability.

**Definition 4** (Identifiability). *Let  $\mathcal{F}$  be a set of  $\mathbb{G}$ -valued coboundary maps  $f : \mathbb{X} \times \mathbb{Y} \rightarrow \mathbb{Y}$  and define for each  $f \in \mathcal{F}$  the cocycle  $T_f : (x, x', y') \mapsto f_x \circ f_{x'}^{-1}(y')$ . A  $\mathbb{P}_{Y|X}$ -adapted cocycle  $T$  is identifiable in  $\mathcal{F}$  if*

1.  $\exists f^* \in \mathcal{F}$  with  $T = T_{f^*}$   $(\mathbb{P}_X \otimes \mathbb{P}_{X,Y})$ -a.s.
2. For any  $f \in \mathcal{F}$ , if  $T_f$  is a  $\mathbb{P}_{Y|X}$ -adapted cocycle satisfying (ID), (PI) and (DA), then  $T_f = T$   $(\mathbb{P}_X \otimes \mathbb{P}_{X,Y})$ -a.s.

Lastly, we denote  $\text{Aut}(P) := \{g \in \text{Aut}(\mathbb{Y}) : g_\# P = P\}$  as the set of transformations that leave distribution  $P \in \mathcal{P}(\mathbb{Y})$  unchanged,  $\text{Aut}(P)|_{\mathbb{G}} := \text{Aut}(P) \cap \mathbb{G}$  its restriction to  $\mathbb{G}$ , and  $[f]_P$  as the set of functions equal to  $f$ ,  $P$ -almost surely. With these concepts, we can now state our general identifiability result.

**Theorem 5** (Identifiability of Counterfactual Cocycle). *Let  $T$  be a  $\mathbb{P}_{Y|X}$ -adapted,  $\mathbb{G}$ -valued cocycle.  $T$  is identifiable in  $\mathcal{F}_{\mathbb{G}}$  if and only if  $\text{Aut}(\mathbb{P}_{Y|X=x_0})|_{\mathbb{G}} \subseteq [\text{id}]_{\mathbb{P}_{Y|X=x_0}}$  for some  $x_0 \in \mathbb{X}$ .*

Intuitively, the result is a consequence of the fact that (i) any  $\mathbb{G}$ -valued  $\mathbb{P}_{Y|X}$ -adapted cocycle has an almost sure coboundary representation  $T_{x,x'} = f_x \circ f_{x'}^{-1}$  with  $\{f_x\}_{x \in \mathbb{X}} \subseteq \mathbb{G}$ , and (ii) any other cocycle  $\tilde{T}$  differs only by a twist—i.e.,  $\tilde{T}_{x,x'} = f_x \circ b_x \circ b_{x'}^{-1} \circ f_{x'}^{-1}$  with  $b_x \in \text{Aut}(\mathbb{P}_{Y|X=x_0})|_{\mathbb{G}}$ . Thus, identifiability requires that the elements of  $\text{Aut}(\mathbb{P}_{Y|X=x_0})|_{\mathbb{G}}$  behave like the identity map. Any non-trivial elements of  $\text{Aut}(\mathbb{P}_{Y|X=x_0})|_{\mathbb{G}}$  can lead to indeterminacy. This parallels identifiability results in latent variable models and nonlinear ICA (Khemakhem et al., 2020; Xi and Bloem-Reddy, 2023), where partially identified models can still be useful for some estimands (Syrota et al., 2025). Note that the choice of  $x_0$  in Theorem 5 is arbitrary: if  $g \neq_{\text{a.s.}} \text{id}$  is an automorphism of  $\mathbb{P}_{Y|X=x_0}$  and  $g \in \mathbb{G}$ , then, for any  $x \in \mathbb{X}$ ,  $\hat{g}_x := T_{x,x_0} \circ g \circ T_{x,x_0}^{-1} \neq_{\text{a.s.}} \text{id}$  is an automorphism of  $\mathbb{P}_{Y|X=x}$  and  $\hat{g}_x \in \mathbb{G}$ .

**Identifiable Parameterizations via TMI Maps** In general, specifying models with a smaller group  $\mathbb{G}$  preserves identifiability for a larger set of conditional distributions  $\mathbb{P}_{Y|X}$ , but risks violating (DA). A practical way to achieve identifiability without risking (DA) for continuous outcomes  $Y$ , is to constrain  $\mathbb{G}$  using knowledge of the causal ordering of variables. In particular, suppose  $Y = (Y_1, \dots, Y_p) \in \mathbb{R}^p$  admit a known causal order  $Y_1 \prec \dots \prec Y_p$  (here we assume the variables are already permuted to reflect this order). To preserve the causal order, it is natural to enforce a lower triangular structure in the transports via  $f_x$ :

$$f_x(y) = (f_{x,1}(y_1), f_{x,2}(y_1, y_2), \dots, f_{x,p}(y_1, \dots, y_p)).$$

Requiring that each  $f_{x,j}$  is strictly increasing in  $y_j$  makes  $f_x$  a *triangular monotone increasing* (TMI) map. The set of such maps with full support forms a transformation group,  $\mathbb{G}_{\text{TMI}}$ . It is known that, for any two absolutely continuous distributions  $P, Q \in \mathcal{P}(\mathbb{R}^p)$ , there is a (a.s.) unique TMI map  $g \in \mathbb{G}_{\text{TMI}}$  such that  $P = g\#Q$  (Bogachev et al., 2005). Below we use this result to prove that, if  $T$  is a  $\mathbb{G}_{\text{TMI}}$ -valued cocycle adapted to *any*  $\mathbb{P}_{Y|X}$  (i.e., not necessarily absolutely continuous) and  $\mathbb{Y} \subset \mathbb{R}^p$ , then it is identifiable within  $\mathcal{F}_{\mathbb{G}_{\text{TMI}}}$ .

**Theorem 6** (Identifiability under TMI maps). *Let  $T$  be a  $\mathbb{P}_{Y|X}$ -adapted,  $\mathbb{G}_{\text{TMI}}$ -valued cocycle and  $\mathbb{Y} \subseteq \mathbb{R}^p$ . Then  $T$  is identifiable in  $\mathcal{F}_{\mathbb{G}_{\text{TMI}}}$ .*

Since any family of *autoregressive flows* (e.g., Table 1) lie in  $\mathcal{F}_{\mathbb{G}_{\text{TMI}}}$ , these architectures are a natural choice to model the coboundary map  $f$  whilst guaranteeing identifiability, by specifying  $f(x, \cdot)$  as a flow with parameters conditioned on  $x$ . We cover implementation details in Section 7. Although such architectures lead to mis-specification problems in SCMs in Section 2.1, we will see that using them to model the cocycle avoids those problems.

### 3.3 Connection to SCMs

The factorization structure of cocycles and their viable parameterization using the same function classes as flow-based SCMs suggests a close connection between counterfactual cocycles and SCMs. This connection is formalized below and has important implications.

**Theorem 7** (Cocycle Equivalence to Structural Model). *A collection of counterfactual variables  $\{Y(x)\}_{x \in \mathbb{X}}$  satisfies Assumption 1 and Assumption 2 with cocycle  $T : \mathbb{X}^2 \times \mathbb{Y} \rightarrow \mathbb{Y}$  if and only if there is a function  $f : \mathbb{X} \times \mathbb{Y}_0 \rightarrow \mathbb{Y}$  injective on its second argument, such that*

$$Y =_{\text{a.s.}} f(X, \xi_Y), \quad \xi_Y \in \mathbb{Y}_0 \subseteq \mathbb{Y}, \quad \xi_Y \perp\!\!\!\perp X.$$

The proof is straightforward up to some measurability technicalities. To illustrate the first direction, since (CC) implies (ID), (PI) and (DA), by Theorem 3 we have  $Y(x) =_{\text{a.s.}} f_x \circ f_{x'}^+(Y(x'))$  for injective  $f_x$  with left inverse  $f_x^+$ . Thus, one can set  $\xi_Y := f_{x_0}^+(Y(x_0)) \in \mathbb{Y}_0$  and apply (CC) to get  $Y(x) =_{\text{a.s.}} f_x(\xi_Y)$ . Defining  $f : (x, y) \mapsto f_x(y)$ , Assumption 1 gives  $Y =_{\text{a.s.}} f(X, \xi_Y)$  by consistency.  $\xi_Y \perp\!\!\!\perp X$  holds since  $Y(x_0) \perp\!\!\!\perp X \implies f_{x_0}^+(Y(x_0)) \perp\!\!\!\perp X$ .

A direct consequence of Theorem 7 is that every counterfactual cocycle model (CC) corresponds to an equivalence class of SCMs of the form

$$\mathbf{V} := \begin{pmatrix} X \\ Y \end{pmatrix} =_{\text{a.s.}} \begin{pmatrix} \xi_X \\ f(\xi_X, \xi_Y) \end{pmatrix} =: F(\boldsymbol{\xi}), \quad (11)$$

where  $\xi_X \perp\!\!\!\perp \xi_Y$  and  $F : \mathbb{V} \rightarrow \mathbb{V}$  is injective by the injectivity of  $f(x, \cdot)$ . Each member of the equivalence class is defined by a different noise distribution  $\mathbb{P}_{\xi} := \mathbb{P}_{\xi_X} \otimes \mathbb{P}_{\xi_Y}$ . Naturally, when  $f(x, \cdot)$  is bijective (or has a bijective extension), the equivalence is to a class of BCMs.

Note that when  $Y$  is multivariate, (11) shows only a partial factorization of  $F$ . However, under the TMI restriction in Section 3.2,  $f$  further decomposes into coordinate maps,

$$Y_j =_{\text{a.s.}} f_j(X, Y_{<j}, \xi_{Y,j}), \quad \xi_{Y,j} \perp\!\!\!\perp X,$$

where each coordinate function  $f_j$  strictly increasing in  $\xi_{Y,j}$ . This is precisely the identifiability restriction used in BCMs (Nasr-Esfahany et al., 2023; Javaloy et al., 2023) and so is a natural counterpart to the TMI restriction proposed in the previous subsection. However, we do *not* need to assume  $\xi_{Y,j} \perp\!\!\!\perp \xi_{Y,i}$  for  $i \neq j$  when deriving an SCM from a cocycle.

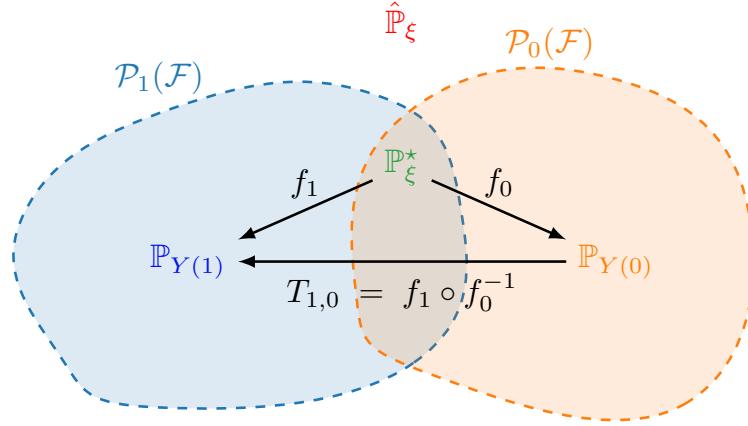


Figure 5: Illustration of robustness of counterfactual cocycle modeling in a randomized control trial (i.e.,  $Y(x) \sim \mathbb{P}_{Y|X=x}$  and  $x \in 0, 1$ ). Each dashed region,  $\mathcal{P}_1(\mathcal{F})$  and  $\mathcal{P}_0(\mathcal{F})$ , denotes the set of distributions reachable via pushforwards of  $\mathbb{P}_{Y(1)}$  and  $\mathbb{P}_{Y(0)}$  under inverses of elements of  $\mathcal{F} \subset \{f : \mathbb{Y} \rightarrow \mathbb{Y}\}$ . Flow-based SCMs require committing to a fixed base distribution  $\hat{\mathbb{P}}_\xi$ ; the class  $\mathcal{F}$  is well-specified if and only if  $\hat{\mathbb{P}}_\xi$  lies in the intersection  $\mathcal{P}_1(\mathcal{F}) \cap \mathcal{P}_0(\mathcal{F})$ . By contrast, cocycle models require only that the intersection is non-empty. In this case, there exists some  $\mathbb{P}_\xi^* \in \mathcal{P}_1(\mathcal{F}) \cap \mathcal{P}_0(\mathcal{F})$  and  $f_0, f_1 \in \mathcal{F}$  such that  $(f_x)_\# \mathbb{P}_\xi^* = \mathbb{P}_{Y(x)}$  for  $x = 0, 1$ , which gives the counterfactual cocycle  $T(x, x', y) = f_x \circ f_{x'}^{-1}(y)$ . Note the existence of  $(\mathbb{P}_\xi^*, f_0, f_1)$  does not depend on which true noise distribution generated the data.

**Noise Invariance of Counterfactual Cocycles** A key difference between counterfactual cocycles and SCMs is that cocycles are *noise-invariant*. To demonstrate, let  $Y(x) := f_x(\xi)$  where each  $f_x := f(x, \cdot)$  is bijective. Notice that each counterfactual transport

$$T_{x,x'} := f_x \circ f_{x'}^{-1}, \quad (T_{x,x'})_\# : \mathbb{P}_{Y|X=x} \mapsto \mathbb{P}_{Y|X=x'}, \quad (12)$$

remains unchanged if we modify the SCM by changing the noise law  $\mathbb{P}_\xi$ . This invariance suggests a practical robustness to directly targeting a cocycle that solves (12). In particular, the flow-based SCM approach is to model the system in (11) by fixing some base distribution  $\hat{\mathbb{P}}_{\xi_Y}$  and class of functions  $\mathcal{F}$  to model  $f$  (note one can always set  $\hat{\mathbb{P}}_{\xi_X} = \hat{\mathbb{P}}_X = \frac{1}{n} \sum_i \delta_{X^{(i)}}$  (see Nasr-Esfahany et al. (2023))). As we showed in Example 1 in Section 2.1, whether there is a function  $\hat{f} \in \mathcal{F}$  that correctly models the data distribution, i.e.,

$$(\hat{f}_x)_\# \hat{\mathbb{P}}_{\xi_Y} = \mathbb{P}_{Y|X=x}, \quad \forall x \in \mathbb{X}, \quad (13)$$

depends on whether the properties of the *chosen* noise distribution  $\hat{\mathbb{P}}_{\xi_Y}$  (e.g., support and tails) matches the *true* noise distribution  $\mathbb{P}_{\xi_Y}$ . In contrast, the cocycle-based approach requires there to exist a function  $\hat{f} \in \mathcal{F}$  that solves (12) for every  $x, x' \in \mathbb{X}$ . This is true as long as there exists *some* noise distribution  $\mathbb{P}_\xi^* \in \mathcal{P}(\mathbb{Y})$  such that (13) holds:

$$\exists \mathbb{P}_\xi^* \in \mathcal{P}(\mathbb{Y}) : (\hat{f}_x)_\# \mathbb{P}_\xi^* = \mathbb{P}_{Y|X=x}, \quad \forall x \in \mathbb{X}. \quad (14)$$

To see this, note that (14) implies  $(\hat{f}_x^{-1})_{\#}\mathbb{P}_{Y|x=x} =_d (\hat{f}_{x'}^{-1})_{\#}\mathbb{P}_{Y|x=x'}$  for every  $x$  and  $x'$ , which implies (12). (14) is a much weaker requirement than (13) and, for instance, is satisfied in Example 1 for some  $\hat{f} \in \mathcal{F}$  when using most of the flow classes in Table 1. Moreover, whether (14) is satisfied does not depend on the true noise distribution of the corresponding SCM. Fig. 5 illustrates this point in a simple binary treatment setting. These properties are analyzed in greater detail in Section 6 and underpin our approach to estimation.

## 4 Counterfactual Cocycles: General Framework Under Confounding

Having established a viable approach to modeling counterfactual transports in a simple set-up, we now present our framework under a more general setting, cover our approach to model parameterization, and how we propose to estimate causal quantities with cocycles.

### 4.1 Counterfactual Cocycle Models Under Partial Orderings and Confounding

Let  $\mathbf{V} := (V_1, \dots, V_d) \in \prod_j^d \mathbb{V}_j \subset \mathbb{R}^d$  now denote the full set of observed variables. We retain focus on a structured setting where  $X := (X_1, \dots, X_q) \subset \mathbf{V}$  are a set of ('treatment') variables we wish to manipulate,  $Y := (Y_1, \dots, Y_p) \subset \mathbf{V}$  are a set of outcomes of interest that may be affected by the treatments, and  $Z := (Z_1, \dots, Z_l) \subset \mathbf{V}$  are a collection of 'pre-treatment' covariates which may confound the effect of any  $X_j$  on  $Y_i$ . That is, we assume the variable sets  $(Z, X, Y)$  satisfy the following *partial* causal ordering,

$$Z \prec X \prec Y, \quad (15)$$

where  $Z \prec X \implies Z_i \prec X_j$  for every  $(i, j) \in \{1, \dots, l\} \times \{1, \dots, q\}$ . Fig. 6 shows several causal DAGs consistent with this ordering. Although we are yet to discuss unobserved confounding, the examples in Fig. 6 are compatible with the framework we develop here. Note that, when  $X \in \mathbf{V}$  is a single treatment, one can trivially find sets  $Z, Y \subset \mathbf{V}$  that satisfy (15) for *any* causal DAG over  $\mathbf{V}$ . The only notable exclusion is the dynamic treatment regime, where intermediate outcomes can affect subsequent treatments. However, we do not envisage any fundamental challenge extending to that setting in future work.

The partial ordering (15) lets us straightforwardly extend the framework laid out in Section 3. The basic idea is to specify Assumption 1 and Assumption 2 but on a set of counterfactuals  $\{Y(x, z)\}_{x, z \in \mathbb{X} \times \mathbb{Z}}$  under both levels of treatments  $X$  and covariates  $Z$ . These counterfactuals relate to the counterfactuals of interest  $\{Y(x)\}_{x \in \mathbb{X}}$  via an additional 'nested consistency' property, which is commonly used when combining counterfactuals with graphs (Richardson and Robins, 2013; Shpitser and Tchetgen Tchetgen, 2016; Malinsky et al., 2019). In particular, our counterfactual assumptions are as follows.

**Assumption 3** (Counterfactual Cocycles with Covariates). *Let  $\{Y(x, z)\}_{(x, z) \in \mathbb{X} \times \mathbb{Z}}$  satisfy:*

1. **Consistency:** (i)  $Y(X, Z) =_{\text{a.s.}} Y$  and (ii)  $Y(x, Z) =_{\text{a.s.}} Y(x)$
2. **Exchangeability:**  $\{Y(x, z)\}_{(x, z)}$   $\perp\!\!\!\perp (X, Z)$
3. **Counterfactual Cocycle:**  $Y(x, z) =_{\text{a.s.}} T_{(x, z), (x', z')}(Y(x', z')), \forall (x, x', z, z') \in \mathbb{X}^2 \times \mathbb{Z}^2$

We stress that these assumptions do not further restrict the causal structure implied by (15). They simply provide a compatible potential outcomes representation which lets us

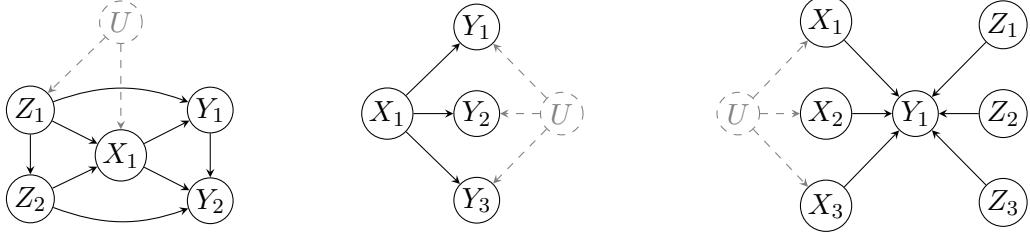


Figure 6: DAGs consistent with ordering in (15) and Assumption 3.  $U$  is unobserved.

formalize counterfactual transports in this setting. Note that Assumption 3.1 and 3.2 imply the well-known ‘strong ignorability’ criteria used in the potential outcomes literature,

$$Y(X) =_{\text{a.s.}} Y \quad \text{and} \quad \{Y(x)\}_{x \in \mathbb{X}} \perp\!\!\!\perp X \mid Z. \quad (\text{SI})$$

The assumptions identify  $\mathbb{P}_{Y(x)}$  via the adjustment formula:  $\mathbb{E}[h(Y(x))] = \mathbb{E}[\mathbb{E}[h(Y) \mid X = x, Z]]$  (Rubin, 1974). However Assumption 3.1 and 3.2 are stronger than (SI) as they additionally identify  $\mathbb{P}_{Y(x,z)} = \mathbb{P}_{Y|X=x, Z=z}$ . In terms of the implied causal graphs, this precludes any direct confounding of  $Z \leftrightarrow Y$ , but does allow confounding within each block  $Z, X, Y$  and between  $Z \leftrightarrow X$  (see Fig. 6). The main benefit of precluding confounding between  $Z \leftrightarrow Y$  is it lets us reduce the problem of recovering a coupling over counterfactuals  $\{Y(x)\}_{x \in \mathbb{X}}$ , to the problem of estimating transports  $\{T_{(x,z),(x',z')}\}_{(x,z) \in \mathbb{X} \times \mathbb{Z}}$  between conditional distributions  $(\mathbb{P}_{Y|X=x, Z=z})_{x,z \in \mathbb{X} \times \mathbb{Z}}$  (which we show how to do in Section 5). This is because, by the nested consistency Assumption 3.1.(ii) and counterfactual cocycle Assumption 3.3, the transports  $\{T_{(x,z),(x',z')}\}_{(x,z) \in \mathbb{X} \times \mathbb{Z}}$  determine the (stochastic) coupling over  $\{Y(x)\}_{x \in \mathbb{X}}$ :

$$Y(x) =_{\text{a.s.}} T_{(x,Z),(x',Z)}(Y(x')). \quad (16)$$

**Implied Causal Model** Since Assumption 3 is equivalent to Assumption 1 and Assumption 2 but on an augmented set of ‘treatments’  $\tilde{X} := (X, Z)$ , all results in Section 3 apply equivalently to the set of transports on  $\{Y(x, z)\}_{(x,z) \in \mathbb{X} \times \mathbb{Z}}$ . In particular, the set of admissible transports must satisfy (ID), (PI) and (DA) w.r.t.  $\mathbb{P}_{Y|X,Z}$ . This in turn implies that  $T_{(x,z),(x',z')} =_{\text{a.s.}} f_{x,z} \circ f_{x',z'}^+$  for some injective  $f_{x,z} : \mathbb{Y}_0 \rightarrow \mathbb{Y}$  and that  $Y =_{\text{a.s.}} f(X, Z, \xi_Y)$ ,  $\xi_Y \perp\!\!\!\perp (X, Z)$ . Therefore, any joint BCM over the variable blocks,

$$\mathbf{V} = \begin{pmatrix} Z \\ X \\ Y \end{pmatrix} =_{\text{a.s.}} \begin{pmatrix} \xi_Z \\ h(\xi_Z, \xi_X) \\ f(h(\xi_Z, \xi_X), \xi_Z, \xi_Y) \end{pmatrix} := F(\boldsymbol{\xi}), \quad (17)$$

is again consistent with the cocycle  $T$ . Different choices of  $h$  and reparameterizations of the noise  $(\xi_Z, \xi_X, \xi_Y)$  all correspond to the *same* cocycle by noise invariance. Thus, counterfactual cocycles provide the *minimal structure* needed to recover the required couplings and counterfactuals, without committing to a full SCM for  $(Z, X)$ , noise distribution  $\mathbb{P}_\xi$ , or enforcing the noise to factorize within the blocks  $(Z, X)$  and  $Y$ .

The SCM equivalence makes clear that even if  $Z \perp\!\!\!\perp X$  (i.e.,  $Z$  are not confounders, as in Fig. 6, right), if  $Z \not\perp\!\!\!\perp Y$  including them explicitly in the model formulation may still be required to satisfy the injectivity assumption implied by the cocycle. That is, we may only

have  $\dim(\text{supp}(\mathbb{P}_\xi)) \leq \dim(\text{supp}(\mathbb{P}_Y))$  after including enough measured causes  $Z$  of  $Y$ . This is an important distinction from estimating marginal causal effects, where conditioning on unnecessary covariates can increase estimation variance (Henckel et al., 2022).

## 4.2 Cocycle Parameterization and Refinement

**TMI Restrictions** Following the analysis in Section 3.2, for modeling practicality and identifiability we propose to parameterize cocycles using invertible TMI maps,

$$T_{(x,z),(x',z')}(y') := f_{x,z} \circ f_{x',z'}^{-1}(y') \quad \text{where } f_{x,z} \in \mathbb{G}_{\text{TMI}}. \quad (18)$$

In the present context, this essentially imposes a known partial causal ordering between the outcomes, i.e.,  $Y_1 \prec \dots \prec Y_p$  and implies an underlying acyclic SCM,

$$Y_j =_{\text{a.s.}} f_j(X, Z, Y_{<j}, \xi_{Y,j}), \quad \xi_{Y,j} \perp\!\!\!\perp X, Z. \quad (19)$$

While this is not the only possible restriction that can achieve identifiability, it is commonly used in the literature (Javaloy et al., 2023; Nasr-Esfahany et al., 2023; Machado et al., 2024) and, as discussed in Section 3.2, guarantees a well-defined transport whenever  $Y_1, \dots, Y_p$  are continuous on  $\mathbb{R}^p$ . In practice, we parameterize the coboundary map  $f(x, z, y) := f_{x,z}(y)$  using classes of autoregressive flows on  $y$  (Table 1), with flow parameters conditioned on the values of  $x, z$ . Implementation details are in Section 7.

**Incorporating Causal DAGs** If  $\mathcal{G}$  is a known causal DAG over  $\mathbf{V}$ , we can additionally restrict the transports to reflect the sparsity of the direct effects. In particular, if  $Y_{\text{pa}(j)} \subset (X, Z, Y_{<j})$  are the parents of  $Y_j$  in  $\mathcal{G}$ , then the SCM is of the form

$$Y_j =_{\text{a.s.}} f_j(Y_{\text{pa}(j)}, \xi_{Y,j}), \quad \xi_{Y,j} \perp\!\!\!\perp X, Z.$$

In Section 7, we discuss how we practically constrain the coboundary map to reflect this sparsity. The main idea is to use TMI maps with masks reflecting the structure of the DAG.

## 4.3 Estimating Causal Quantities with Cocycles

Any cocycle parameterized as in (18) trivially satisfies (ID) and (PI) on all of  $\mathbb{Y}$ . Thus, all that remains is to enforce (DA) w.r.t.  $\mathbb{P}_{Y|X,Z}$  and use the resulting cocycle to estimate causal quantities. Here we overview our proposed procedure, deferring details on how to estimate the cocycle itself to Section 5. In short, our estimation procedure is centered entirely around the cocycle and avoids ever specifying or referencing a latent noise distribution. This lets us take advantage of the invariances of counterfactual cocycles discussed in Section 3.3 and later in Section 6. In what follows, we define  $\tilde{X} := (X, Z)$  for convenience.

**Cocycle Estimation** Rather than specifying a base distribution  $\hat{\mathbb{P}}_\xi$  and learning a flow to the conditional as  $f_{\tilde{x}} : \hat{\mathbb{P}}_\xi \mapsto \mathbb{P}_{Y|\tilde{X}=\tilde{x}}$ , we directly target the transports between the conditionals, as  $T_{\tilde{x},\tilde{x}'} = f_{\tilde{x}} \circ f_{\tilde{x}'}^{-1} : \mathbb{P}_{Y|\tilde{X}=\tilde{x}} \mapsto \mathbb{P}_{Y|\tilde{X}=\tilde{x}'}$ . We do so by minimizing a tractable empirical analogue to a distributional discrepancy in the (DA) property,

$$\tilde{\ell}(T) := \mathbb{E}_{\tilde{x},\tilde{x}' \sim \mathbb{P}_{\tilde{X}}} D(\mathbb{P}_{Y|\tilde{X}=\tilde{x}}, (T_{\tilde{x},\tilde{x}'})^\# \mathbb{P}_{Y|\tilde{X}=\tilde{x}'})^2.$$

In Section 5 we derive this estimator in detail, and show that, under general conditions, the consistency of this estimator does not depend on any distributional assumptions that follow from the (true) latent noise  $\xi$  (see Theorem 11 in Section 5).

**Causal Quantity Estimation** Rather than using the abduct-act-predict procedure in Section 2.1, which may require sampling from a prior or posterior base distribution, we use the cocycles to directly impute the counterfactuals of interest. In particular, given an i.i.d. dataset  $\{Z^{(i)}, X^{(i)}, Y^{(i)}\}_{i=1}^n \sim \mathbb{P}_{Z,X,Y}$ , we can impute for each unit the counterfactual outcomes at treatment levels  $x_1, \dots, x_m$  using (16) and the consistency property:

$$\{\hat{Y}^{(i)}(x_1), \dots, \hat{Y}^{(i)}(x_m)\} = \{\hat{T}_{(x_1, Z^{(i)}), (X^{(i)}, Z^{(i)})}(Y^{(i)}), \dots, \hat{T}_{(x_m, Z^{(i)}), (X^{(i)}, Z^{(i)})}(Y^{(i)})\}$$

The imputed counterfactuals can then be used to estimate causal quantities via standard empirical and/or nonparametric techniques. Below are some examples using empirical averaging and kernel density estimation with smoothing kernel  $K_\lambda$ :

$$\text{Average Effect: } \widehat{\mathbb{E}}[Y(x) - Y(0)] = \frac{1}{n} \sum_{i=1}^n (\hat{Y}^{(i)}(x) - \hat{Y}^{(i)}(0)), \quad (20)$$

$$\text{True Harm Rate: } \widehat{\mathbb{P}}(Y(x) - Y(0) \leq 0) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{\hat{Y}^{(i)}(x) - \hat{Y}^{(i)}(0) \leq 0\} \quad (21)$$

$$\text{Density of Effect: } \widehat{p}_{Y(x)-Y(0)}(y) = \frac{1}{n} \sum_{i=1}^n K_\lambda(\hat{Y}^{(i)}(x) - \hat{Y}^{(i)}(0) - y) \quad (22)$$

One can also condition such quantities on covariates  $W \subset Z$  to examine effect heterogeneity. For this, one can replace the empirical averages in (20)-(22), with *weighted* averages estimated nonparametrically. For example, the conditional THR can be estimated as

$$\widehat{\mathbb{P}}(Y(x) - Y(0) \leq 0 \mid W = w) = \sum_{i=1}^n \hat{\alpha}_i(w) \mathbf{1}\{\hat{Y}^{(i)}(x) - \hat{Y}^{(i)}(0) \leq 0\} \quad (23)$$

where  $\hat{\alpha}_i(\cdot)$  are smoothing weights estimated via nonparametrically regressing  $\mathbf{1}\{\hat{Y}^{(i)}(x) - \hat{Y}^{(i)}(0) \leq 0\}$  on  $W^{(i)}$  using e.g., Nadaraya–Watson or RKHS regression. While we note that nonparametrics can suffer from slow rates of convergence in high-dimensions, in causal inference  $W$  is typically a low-dimensional (e.g., 1d or 2d) set of interest.

## 5 Cocycle Estimation

To leverage any benefits of directly modeling counterfactual cocycles, we require a tractable way to estimate them without making further assumptions on  $\mathbb{P}_{Y|X,Z}$ . This task is non-trivial for general flow classes beyond simple additive models, where one cannot recover the cocycle merely by regressing  $Y$  on functions of  $X$  and  $Z$ . In what follows, we develop our estimation procedure from first principles, establish its asymptotic properties, and demonstrate its performance on the problems in Example 1.

For notational simplicity, we henceforth write  $X$  in place of the augmented variable  $(X, Z)$ , since including covariates does not affect any of the results below.

### 5.1 Targeting (DA) via Distributional Discrepancy

Given a parameterized model  $\mathcal{F}$  for the coboundary map  $f : (x, y) \mapsto f_x(y)$  of a cocycle  $T$ , a natural estimation criterion is to minimize an expected distance that enforces (DA)

across all conditionals. In particular, denoting by  $D$  a divergence or metric on  $\mathcal{P}(\mathbb{Y})$ , the set of distributions on  $\mathbb{Y}$ , the criterion evaluated at  $f \in \mathcal{F}$  is

$$\ell_0(f) = \mathbb{E}_{X, X' \sim \mathbb{P}_X} D(\mathbb{P}_{Y|X}, (f_X \circ f_{X'}^{-1})_\# \mathbb{P}_{Y|X'})^2. \quad (24)$$

where  $\mathbb{P}_{Y|X} := \mathbb{P}(Y \in \cdot | X)$  is treated as a random probability measure on  $\mathbb{Y}$ , and  $f_X$  a random function  $\mathbb{Y} \rightarrow \mathbb{Y}$ . When the cocycle is identifiable from a model  $\mathcal{F}$  (e.g.,  $\mathcal{F} \subset \mathcal{F}_{\mathbb{G}_{\text{TMI}}}$ —see Section 3.2), then by definition  $\ell_0$  admits a minimizer that is almost-everywhere unique. Unfortunately, evaluating  $\ell_0$  requires knowledge of  $\mathbb{P}_{Y|X}$ , which defeats the purpose of modeling only the cocycle. Our aim is to modify this objective in a way that bypasses the need to estimate conditional distributions, without harming identifiability. To that end, we choose  $D$  to be the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012), which induces the following loss in terms of the cocycle  $T$

$$\ell_0(T) = \mathbb{E}_{X, X' \sim \mathbb{P}_X} \|\mathbb{E}[\psi(Y)|X] - \mathbb{E}[\psi(T_{X, X'}(Y'))|X, X']\|_{\mathcal{H}_k}^2. \quad (25)$$

Here  $(X, Y) \perp\!\!\!\perp (X', Y')$  are independent copies and  $\psi : \mathbb{Y} \rightarrow \mathcal{H}_k$  is a feature map to a reproducing kernel Hilbert space (RKHS)  $\mathcal{H}_k$  associated with a positive-definite kernel  $k : \mathbb{Y}^2 \rightarrow \mathbb{R}$ , in the sense that  $\psi(y) = k(y, \cdot)$ . Note, we assume the kernel  $k$  is *characteristic* so that the MMD is indeed a metric on  $\mathcal{P}(\mathbb{Y})$  (Sriperumbudur et al., 2011). Popular characteristic kernels include the Gaussian kernel  $k(y, y') = \exp(-\lambda \|y - y'\|_2^2)$ , and Laplace kernel  $k(y, y') = \exp(-\lambda \|y - y'\|_1^1)$ . While one could attempt to estimate  $\ell_0$  from data via nonparametric conditional mean embedding estimation (Li et al., 2022), we will modify the MMD objective so that it can be estimated using simple empirical averages.

**Modifying the Objective** We start by using the standard identity  $\mathbb{E}[\|W\|_{\mathcal{H}_k}^2] = \|\mathbb{E}[W]\|_{\mathcal{H}_k}^2 + \text{Tr}[\text{Cov}_{\mathcal{H}_k}[W]]$  for any RKHS-valued random variable  $W \in \mathcal{H}_k$  with  $\mathbb{E}\|W\|_{\mathcal{H}_k}^2 < \infty$  (Berlinet and Thomas-Agnan, 2011), where

$$\text{Cov}_{\mathcal{H}_k}[W] := \mathbb{E}[(W - \mathbb{E}[W]) \otimes (W - \mathbb{E}[W])] \in B_1(\mathcal{H}_k)$$

is the covariance operator,  $B_1(\mathcal{H}_k)$  is the set of trace-class operators on  $\mathcal{H}_k$ , and  $\text{Tr} : B_1(\mathcal{H}_k) \rightarrow \mathbb{R}$  is the standard trace functional. Applying this identity to the norm inside the outer expectation in (25) with  $W = \psi(Y) - \mathbb{E}[\psi(T_{X, X'}(Y'))|X = x, X' = x']$ , we get,

$$\tilde{\ell}(T) := \ell_0(T) - \text{Tr}[\text{Cov}[\psi(Y)]] = \mathbb{E}_{(X', X, Y) \sim \mathbb{P}_X \otimes \mathbb{P}_{X, Y}} \|\psi(Y) - \mathbb{E}[\psi(T_{X, X'}(Y'))|X, X']\|_{\mathcal{H}_k}^2 \quad (26)$$

Note that minimizing  $\tilde{\ell}$  is equivalent to minimizing  $\ell_0$ , so we can simply work with  $\tilde{\ell}$  instead. This removes one of the two conditional expectations in (25). Now, to remove the other conditional expectation, we pass the expectation over  $X'$  *inside* the norm, yielding

$$\ell(T) = \mathbb{E}_{(X, Y) \sim \mathbb{P}_{X, Y}} \|\psi(Y) - \mathbb{E}[\psi(T_{X, X'}(Y'))|X]\|_{\mathcal{H}_k}^2. \quad (27)$$

Using the reproducing property of  $k$  (i.e.,  $k(y, y') = \langle \psi(y), \psi(y') \rangle_{\mathcal{H}_k}$ ), the resulting loss function can be expressed as an expectation of a real-valued function,

$$\ell(T) = \mathbb{E}_{(X, Y), (X', Y'), (X'', Y'') \sim \mathbb{P}_{X, Y}} \left( k(Y, Y) + k(T_{X, X'}(Y'), T_{X, X''}(Y'')) - 2k(Y, T_{X, X'}(Y')) \right) \quad (28)$$

We note that exchanging the expectation over  $X'$  with the norm does not guarantee to preserve the minimizing set. However, below we prove that any minimizer  $T^*$  of  $\ell$  in a set of  $\mathbb{G}$ -valued cocycles satisfies (ID), (PI), and (DA) almost surely.

**Theorem 8** (CMMD Identifiability). *Let  $T$  satisfy (CC) w.r.t.  $\{Y(x)\}_{x \in \mathbb{X}}$  and  $T \in \mathcal{T}_{\mathbb{G}}$ , where  $\mathcal{T}_{\mathbb{G}} := \{T_f : f \in \mathcal{F}_{\mathbb{G}}\}$  is the set of cocycles constructed from coboundary maps in  $\mathcal{F}_{\mathbb{G}}$ . Let  $\ell$  be defined using characteristic kernel  $k : \mathbb{Y}^2 \rightarrow \mathbb{R}$ . Then, any  $T^* \in \operatorname{arginf}_{\tilde{T} \in \mathcal{T}_{\mathbb{G}}} \ell(\tilde{T})$  satisfies (ID), (PI) and (DA) for  $(\mathbb{P}_X \otimes \mathbb{P}_X)$ -almost all  $x, x'$ .*

The restriction to  $\mathbb{G}$ -valued cocycles can be relaxed, but as with earlier results lets us avoid certain complexities working with monoids. An immediate consequence of this result is that, whenever  $\mathcal{F} \subset \mathcal{F}_{\mathbb{G}_{\text{TMI}}}$  (or in any other situation in which  $\mathcal{F}$  is identifiable), the minimizer of  $\ell$  identifies the counterfactual cocycle  $T^*$ . Going forward, we refer to the objective (28) as the **CMMD loss** (short for Cocycle MMD). Intuitively, one can view CMMD as minimizing the (average) error between true and predicted counterfactuals in Hilbert space,  $\{\psi(Y(x^{(i)}))\}_{i=1}^n$ , where the predicted counterfactual at  $x^{(i)}$  is just the average transformed embedding  $\psi(\hat{Y}(x^{(i)})) := \frac{1}{n} \sum_{j \neq i} \psi(\hat{T}_{x^{(i)}, x^{(j)}}(Y(x^{(j)})))$ .

**Tractable Empirical Analogues** The only expectations in (28) are over  $\mathbb{P}_{X,Y}$ . Therefore, given data  $\mathcal{D}_n = \{(X^{(i)}, Y^{(i)})\} \sim_{iid} \mathbb{P}_{X,Y}$ , one can replace the population expectations with empirical ones. This gives rise to the following empirical V-statistic and U-statistic estimators for  $\ell$  (dropping all terms independent of  $T$ ):

$$\ell_n^V(T) = -\frac{2}{n^2} \sum_{i,j}^n k(Y^{(i)}, Y_T^{(i,j)}) + \frac{1}{n^3} \sum_{i,j,k}^n k(Y_T^{(i,j)}, Y_T^{(i,k)}) \quad (29)$$

$$\ell_n^U(T) = -\frac{2}{n(n-1)} \sum_{i \neq j}^n k(Y^{(i)}, Y_T^{(i,j)}) + \frac{1}{n(n-1)(n-2)} \sum_{i \neq j \neq k}^n k(Y_T^{(i,j)}, Y_T^{(i,k)}), \quad (30)$$

where  $Y_T^{(i,j)} := T_{X^{(i)}, X^{(j)}}(Y^{(j)})$ . Both loss functions  $\ell_n^V$  and  $\ell_n^U$  can be used to optimize flow-based cocycles via any gradient-based algorithm (e.g., ADAM) and model select between different flow classes. Implementation details are in Section 7.

## 5.2 Properties of CMMD Estimation

We now analyze the theoretical properties of CMMD estimation. We start off by verifying that both empirical analogues  $\ell_n^V, \ell_n^U$  converge to  $\ell$  at  $\sqrt{n}$ -rate under general conditions.

**Proposition 9.** *Let  $\{(X^{(i)}, Y^{(i)})\}_{i=1}^n \sim_{iid} \mathbb{P}_{X,Y}$ . For any cocycle  $T$  and bounded kernel  $k$ , we have  $\ell_n^V(T) = \ell(T) + c + \mathcal{O}_P(n^{-\frac{1}{2}})$ , where  $c \in \mathbb{R}$ . The same holds for  $\ell_n^U$ .*

We next turn to asymptotic analysis of the resulting estimators. For this, we work under standard parametric assumptions on the model class, and assume the cocycle is identifiable and that the kernel is sufficiently regular (as satisfied by typical kernel choices).

**Assumption 4.** (CMMD Consistency)

1. *Compactness:  $\Theta$  is a compact subset of  $\mathbb{R}^d$ .*
2. *Continuity:  $\theta \mapsto T_{\theta, x, x'}(y)$  is continuous for every  $(x, x', y) \in \mathbb{X}^2 \times \mathbb{Y}$ .*
3. *Identifiability:  $M = \arg \min_{\theta \in \Theta} \ell(\theta) \neq \emptyset$ .  $\theta_1, \theta_2 \in M \implies T_{\theta_1} = T_{\theta_2}$  ( $\mathbb{P}_X \otimes \mathbb{P}_{X,Y}$ )-a.s.*
4. *Kernel Regularity: The kernel  $k$  is continuous and bounded with  $\sup_{y, y'} |k(y, y')| \leq 1$ .*

Under these conditions, we have the following strong consistency result based on the U-statistic (30). By standard theory (e.g., Theorem 5.2.9 in De la Pena and Giné (2012)) analogous results hold for the V-statistic (29).

**Theorem 10** (CMMD Strong Consistency). *Let  $\{(X^{(i)}, Y^{(i)})\}_{i=1}^n \sim_{iid} \mathbb{P}_{X,Y}$  and  $\hat{\theta}_n \in \operatorname{argmin}_{\theta \in \Theta} \ell_n^U(\theta)$  be a measurable minimizer. If Assumption 4 holds, then*

$$\inf_{\theta \in M} \|\hat{\theta}_n - \theta\|_2 \xrightarrow[n \rightarrow \infty]{a.s.} 0 \text{ a.s.}$$

Moreover,  $\exists \theta_0 \in M$  such that  $T_{\theta_0} \xrightarrow{n \rightarrow \infty} T_{\theta_0}$  a.s., for  $\mathbb{P}_X \otimes \mathbb{P}_{X,Y}$  almost all  $(x, x', y')$ .

**Remark 11.** When the data are generated by a BCM  $Y = f(X, \xi)$ , the only condition in Assumption 4 which may depend on  $\mathbb{P}_\xi$  is the identifiability criterion Assumption 4.3. However, this condition only requires that the underlying cocycle  $T$  is almost-everywhere unique, rather than the parameterization  $\theta_0$ . Since there is at most one (a.s. unique) TMI map transporting between two distributions on  $\mathbb{R}^p$  (Theorem 6), as long as  $\mathcal{F} \subset \mathcal{F}_{\mathbb{G}_{TMI}}$  and the model is well-specified, Theorem 10 must hold for any BCM with function  $f$ . Thus, like the cocycle itself, our CMMD estimator also enjoys an invariance to the noise distribution. In contrast, likelihood-based estimators for the flow from a base distribution can fail to converge for certain noise distributions, as discussed in Section 2.1. Likewise, in Section 6.4 we will see conditional-quantile based SCM estimators can be biased under dependent noise.

Under the following additional regularity conditions we obtain  $\sqrt{n}$ -consistency of the U-statistic estimator to the minimizing set. We expect an analogous result for the V-statistic.

**Assumption 5.** (Additional Regularity for  $\sqrt{n}$ -rate)

1. *Lipschitz Cocycle:* There exists a measurable function  $L_T : \mathbb{X}^2 \times \mathbb{T} \rightarrow \mathbb{R}_{>0}$  with  $\mathbb{E}[L_T(X, X', Y')^2] < \infty$ , such that for all  $\theta, \theta' \in \Theta$ ,

$$\|T_{\theta, x, x'}(y') - T_{\theta, x, x'}(y')\|_2 \leq L_T(x, x', y') \|\theta - \theta'\|_2 \quad (\mathbb{P}_X \otimes \mathbb{P}_{X,Y})\text{-a.s.}$$

2. *Kernel Derivative Regularity:*  $\partial k : (y, y') \mapsto \nabla_y k(y, y')$  is continuous and bounded.
3. *Local Strong Convexity:* There exist  $c, \delta > 0$  such that whenever  $\inf_{\theta' \in M} \|\theta - \theta'\|_2 \leq \delta$ ,

$$\|\nabla_\theta \ell(\theta)\|_2 \geq c \inf_{\theta' \in M} \|\theta - \theta'\|_2.$$

**Theorem 12** ( $\sqrt{n}$ -Rate of CMMD). *Let  $\{(X^{(i)}, Y^{(i)})\}_{i=1}^n \sim_{iid} \mathbb{P}_{X,Y}$  and  $\hat{\theta}_n \in \operatorname{argmin}_{\theta \in \Theta} \ell_n^U(\theta)$  be a measurable minimizer. Then, under Assumption 4 and Assumption 5, we have*

$$\inf_{\theta \in M} \|\hat{\theta}_n - \theta\|_2 = O_p(n^{-1/2}).$$

We note the kernel regularity condition is satisfied by popular characteristic kernels such as the Gaussian kernel. The local strong-convexity condition plays an analogous role to the classical positive-definite Hessian assumption used to obtain  $\sqrt{n}$ -rates in the unique minimizer setting (i.e., that  $\nabla^2 \ell(\theta_0) \succeq \lambda_{\min} I_d$  when  $M = \{\theta_0\}$ ) (Van der Vaart, 2000). Indeed, in this case a first-order Taylor expansion of  $\nabla \ell$  around  $\theta_0$  yields the gradient bound  $\|\nabla \ell(\theta)\|_2 \geq \lambda_{\min} \|\theta - \theta_0\|_2$  for  $\theta$  near  $\theta_0$ . We assume the bound directly as it requires only first-order derivatives and is more natural when there are multiple minimizers.

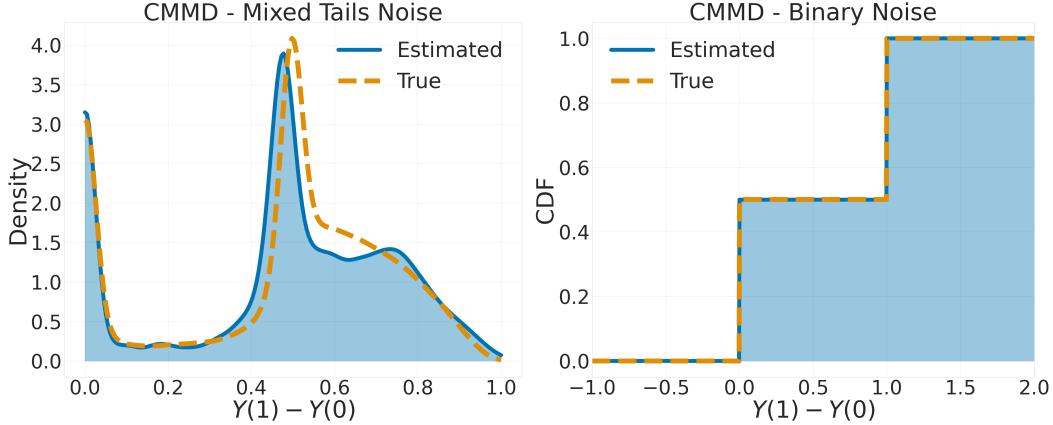


Figure 7: Estimated treatment effect distributions for Example 1 using flow-based cocycles with CMMMD. Left: density under mixed-tailed noise. Right: CDF under discrete noise. Cocycles accurately recover both, unlike flow-based SCMs in Fig. 2.

### 5.3 Demonstration on a Toy Example

We conclude by demonstrating the performance of counterfactual cocycles estimated with CMMMD on Example 1. We assume access to samples  $\{X^{(i)}, Y^{(i)}\}_{i=1}^{2000} \sim \mathbb{P}_{X,Y}$  and specify cocycle models using the same neural spline flow architecture as the flow-based SCMs in that example. We train each model using the CMMMD loss, impute the counterfactuals  $\{Y^{(i)}(0), Y^{(i)}(1)\}_{i=1}^n$  using the cocycles, and estimate the treatment effect density and CDF by (Gaussian) kernel smoothing and empirical averaging, as described in Section 4.3 (i.e., see equations (21) and (22)). CMMMD implementation details are in Section 7 and the exact optimization routine matches that in Section 8.1. Fig. 7 shows the estimated treatment effect density for the mixed-tailed noise case (left) and CDF for the binary noise case (right). In contrast to the poor performance of flow-based SCMs (see Fig. 2), our approach accurately recovers both distributions. This reflects that the CMMMD estimator does not require specifying a latent noise distribution, and that its consistency is generally robust to the (true) noise distribution—as established in the previous subsection.

## 6 Robustness and Simplicity of Counterfactual Cocycles Versus SCMs

In this section we analyze in detail the advantages of modeling counterfactual cocycles in contrast to SCMs, by studying a formal invariance of cocycles under model reparameterizations. This invariance means that centering the modeling and estimation process around the cocycle is more robust to mis-specification, can greatly simplify estimation, and in many cases lets us sidestep the problems with flow-based SCMs identified in Section 2.1. We also discuss advantages over recently proposed SCM-based transport methods using conditional quantile estimation (Plečko and Meinshausen, 2020; Machado et al., 2024).

In what follows, we restrict attention to the simplified setting

$$Y =_{\text{a.s.}} f(X, \xi), \quad \xi \perp\!\!\!\perp X, \quad f(x, \cdot) \text{ is bijective } \forall x \in \mathbb{X}$$

and compare modeling a counterfactual cocycle  $T$  versus a *bijective generating mechanism* (BGM)  $(f, \mathbb{P}_\xi)$ . While a full SCM specifies the entire joint system (cf. (11) and (17)), when the treatments  $X \in \mathbf{V}$  are pre-defined and  $X \prec Y$  one can always set  $X = \xi_X$  and estimate the noise distribution via the empirical distribution of  $X$ . In this case, the substantive difference between the two approaches lies in how the conditional law  $\mathbb{P}_{Y|X}$  is modeled. The same arguments apply with  $(X, Z)$  in place of  $X$  under the more general framework in Section 4, so we need not explicitly distinguish between treatments and covariates.

### 6.1 Noise Invariance of Counterfactual Cocycles

It is known that BGMs can only be identified up to noise automorphisms (Nasr-Esfahany et al., 2023). In particular, let  $Y =_{\text{a.s.}} f(X, \xi)$ , where  $\xi \in \mathcal{E}$ ,  $\xi \perp\!\!\!\perp X$ , and  $f_x := f(x, \bullet) : \mathcal{E} \rightarrow \mathbb{Y}$  is bijective. Then, letting  $\text{Aut}(\mathcal{E})$  be the group of bi-measurable bijections on  $\mathcal{E}$ , we have

$$f(x, \xi) = f(x, g \circ g^{-1}(\xi)) \quad \forall g \in \text{Aut}(\mathcal{E}).$$

Defining  $f^{(g)} : (x, \xi) \mapsto f(x, g(\xi))$ , it is clear that  $\{(f^{(g)}, g_\# \mathbb{P}_\xi)\}_{g \in \text{Aut}(\mathcal{E})}$  is an equivalence class of BGMs that all generate the same conditional distribution  $\mathbb{P}_{Y|X}$ —each one with a different noise parameterization  $\mathbb{P}_{g^{-1}(\xi)}$ . However, while the structural function  $f^{(g)}$  depends on the choice of  $g$ , the corresponding cocycle that satisfies (CC) does not:

$$T_{x,x'} := f_x \circ f_{x'}^{-1} = f_x \circ g \circ g^{-1} \circ f_{x'}^{-1}, \quad \text{for any } g \in \text{Aut}(\mathcal{E}), x, x' \in \mathbb{X}. \quad (31)$$

This has implications for whether a model is well-specified. To illustrate, fix an underlying generating mechanism  $(f^*, \mathbb{P}_\xi^*)$ . With a flow-based SCM, one specifies a fixed base distribution  $\widehat{\mathbb{P}}_\xi$  and a class of functions,  $\mathcal{F} := \{f : \mathbb{X} \times \mathcal{E} \rightarrow \mathbb{Y}\}$ . In this case, the model is well-specified if there is some  $f \in \mathcal{F}$  such that  $f_{x\#} \widehat{\mathbb{P}}_\xi = f_{x\#}^* \mathbb{P}_\xi^*$  for each  $x \in \mathbb{X}$ . Denoting  $\mathcal{T}(\mathbb{P}_\xi^*, \widehat{\mathbb{P}}_\xi)$  as the set of invertible transports from  $\mathbb{P}_\xi^*$  to  $\widehat{\mathbb{P}}_\xi$ , this is true if and only if:

- (I) There exists  $h \in \mathcal{T}(\mathbb{P}_\xi^*, \widehat{\mathbb{P}}_\xi)$  such that  $f^{*(h)} \in \mathcal{F}$ .

On the other hand, the cocycle approach specifies a function class of the same type,  $\mathcal{F} := \{f : \mathbb{X} \times \mathcal{E} \rightarrow \mathbb{Y}\}$ , whose elements will be used to construct candidate cocycles via  $T_{x,x'} = f_x \circ f_{x'}^{-1}$ . The model  $\mathcal{F}$  is well-specified if there is some  $f \in \mathcal{F}$  such that  $f_x \circ f_{x'}^{-1} = T_{x,x'}^* = f_x^* \circ f_{x'}^{*-1}$ . By (31), it is easy to see that this is true if and only if:

- (II) There exists  $h \in \text{Aut}(\mathcal{E})$  such that  $f^{*(h)} \in \mathcal{F}$ .

Since  $\mathcal{T}(\mathbb{P}_\xi^*, \widehat{\mathbb{P}}_\xi)$  is a strict subset of  $\text{Aut}(\mathcal{E})$ , flow-based models with fixed  $\widehat{\mathbb{P}}_\xi$  and function class  $\mathcal{F}$  are well-specified for a strictly smaller set of generating mechanisms than cocycle models that use the same function class  $\mathcal{F}$ . The following example demonstrates how cocycle modelling can avoid the tail and support mis-specification problems outlined in Section 2.1.

**Example 3.** Suppose that  $Y = f^*(X, \xi^*)$ , where  $\xi^* \in \mathcal{E} \subset \mathbb{R}^2$ ,  $X \in \mathbb{R}$  and

$$f^*(X, \xi^*) = AX + \xi^*, \quad \xi^* \sim t_3(0, 1) \otimes \left( \frac{1}{2} U\left(-\frac{3}{2}, -\frac{1}{2}\right) + \frac{1}{2} U\left(\frac{1}{2}, \frac{3}{2}\right) \right).$$

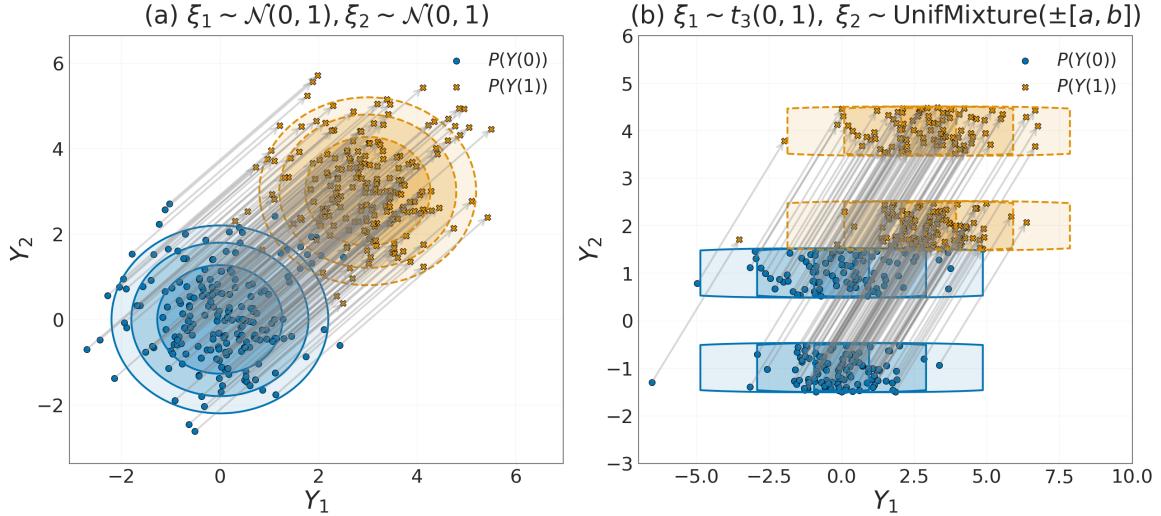


Figure 8: Illustration of invariance of cocycle complexity to noise distribution. Counterfactual transport (gray arrows)  $T_{1,0}$  from samples of  $\mathbb{P}_{Y(0)}$  (blue) to  $\mathbb{P}_{Y(1)}$  (orange) for structural model  $Y(x) = Ax + \xi$  where (i)  $\xi \sim \mathcal{N}(0, I)$  (left) and (ii)  $\xi \sim t_3(0, 1) \otimes (\frac{1}{2}U(-\frac{3}{2}, -\frac{1}{2}) + \frac{1}{2}U(\frac{3}{2}, \frac{1}{2}))$  (right). If  $\widehat{\mathbb{P}}_\xi = \mathcal{N}(0, I)$  is used as a base distribution to learn the BGM  $(f, \widehat{\mathbb{P}}_\xi)$ , there is no continuous bijection  $\widehat{\mathbb{P}}_\xi \mapsto \mathbb{P}_{Y(x)}$  in case (ii). In contrast, the direct transport  $T_{1,0}(y) = A + y$  is a simple linear map, regardless of the true noise distribution  $\mathbb{P}_\xi$ . Intuitively, one can view  $T_{x,0}$  as a structural map from ‘base’ distribution  $\mathbb{P}_{Y(0)}$  to  $\mathbb{P}_{Y(x)}$ .

Suppose we model this mechanism using base distribution  $\widehat{\mathbb{P}}_\xi = \mathcal{N}(0, I)$  and a class of autoregressive Lipschitz diffeomorphisms  $\mathcal{F} \subset \mathcal{F}_{\text{G}_{\text{TMI}}}$ , as in SOTA flow-based SCMs. In this case, both the tails and support of  $\mathbb{P}_\xi^*$  are mis-specified. By the theory of TMI maps (Bogachev et al., 2005), the only  $h \in \mathcal{T}(\mathbb{P}_\xi^*, \widehat{\mathbb{P}}_\xi)$  that results in  $f^{*(h)} \in \mathcal{F}_{\text{TMI}}$  is the map  $h(\hat{\xi}) = (h_1(\hat{\xi}_1), h_2(\hat{\xi}_2))$  where  $h_1$  and  $h_2$  are the quantile transforms from  $\widehat{\mathbb{P}}_{\xi_1}$  to  $\mathbb{P}_{\xi_1}^*$  and  $\widehat{\mathbb{P}}_{\xi_2}$  to  $\mathbb{P}_{\xi_2}^*$  respectively:

$$h_1(\hat{\xi}) = \text{sign}(\hat{\xi}_2) \sqrt{\frac{3(1 - I^{-1}(2\Phi(\hat{\xi}_2); \frac{3}{2}, \frac{1}{2}))}{I^{-1}(2\Phi(\hat{\xi}_2); \frac{3}{2}, \frac{1}{2})}}, \quad h_2(\hat{\xi}_2) = \begin{cases} -\frac{3}{2} + 2\Phi(\hat{\xi}_2), & \Phi(\hat{\xi}_2) < \frac{1}{2}, \\ 2\Phi(\hat{\xi}_2) - \frac{1}{2}, & \Phi(\hat{\xi}_2) \geq \frac{1}{2} \end{cases}$$

Here  $I$  is the regularized incomplete Beta function and  $\Phi$  is the CDF of  $\mathcal{N}(0, 1)$ .  $h_1$  is not Lipschitz and  $h_2$  has a discontinuous jump. Therefore,  $f^{*(h)}$  does not lie in  $\mathcal{F}$ . In contrast, the cocycle  $T_{x,x'}(y) = A(x - x') + y$  does not depend on  $\mathbb{P}_\xi^*$  or  $\widehat{\mathbb{P}}_\xi$  and can be modeled using the set of linear maps  $\mathcal{F}_{\text{LIN}} = \{f(x, \xi) = Ax + \xi \mid A \in \mathbb{R}^2\} \subset \mathcal{F}$ , since  $f^{*(h)} \in \mathcal{F}_{\text{LIN}}$  for  $h = \text{id} \in \text{Aut}(\mathcal{E})$ . This is illustrated on Fig. 8.

The same concept applies to Example 1. In that example, one has  $f^* \in \mathcal{F}_{\text{LIN-SCALE}} = \{f(x, \xi) = \sigma(x)\epsilon \mid \sigma(x) = \beta x + \alpha, (\alpha, \beta) \in [-1, 1]^2\}$ , a set of linearly-parameterized scale transforms, but using fixed  $\widehat{\mathbb{P}}_\xi \in \{\mathcal{N}(0, 1), \text{Lap}(0, 1)\}$  results in a non-affine, non-Lipschitz flow for the mixed-tailed noise design, and no well-defined flow for the binary noise design.

For our last example, we show the *dependence* structure between the coordinates of  $\xi$  can also induce mis-specification problems for SCMs, but again has no effect on the cocycle.

**Example 4.** Consider again the set-up of Example 3, but now where there is a single latent cause of both outcomes:

$$\xi_1^* =_{\text{a.s.}} \xi_2^* \sim N(0, 1).$$

In this case, using a factored Gaussian base distribution  $\hat{\mathbb{P}}_\xi = N(0, I_2)$  for the SCM admits the correct marginals for the noise, but since  $\text{supp}(\mathbb{P}_\xi^*)$  is a lower-dimensional set than  $\text{supp}(\hat{\mathbb{P}}_\xi)$  we have  $\mathcal{T}(\mathbb{P}_\xi^*, \hat{\mathbb{P}}_\xi) = \emptyset$ . In contrast, the true cocycle  $T$  is unaffected by the particular dependence structure in  $\xi$  and can still be well-specified using the class  $\mathcal{F}_{\text{LIN}} = \{f(x, \xi) = Ax + \xi \mid A \in \mathbb{R}^2\} \subset \mathcal{F}$ , since we still have  $f^{*(\text{id})} \in \mathcal{F}_{\text{LIN}}$ .

The above examples all serve to demonstrate an underlying point: any properties of the conditionals  $(\mathbb{P}_{Y|X=x})_{x \in \mathbb{X}}$  inherited from the (true) noise distribution  $\mathbb{P}_\xi$  need not be learned by the counterfactual cocycle, since it is invariant to this distribution. As a practical implication, we can employ popular Lipschitz flow classes (e.g., Table 1) to model counterfactual cocycles while often avoiding the mis-specification issues that arise in flow-based SCMs.

## 6.2 Counterfactual Cocycles as Minimum Complexity SCMs

Another way to view the mis-specification robustness of cocycle modeling is as a ‘minimal complexity’ property of counterfactual cocycles. In particular, the cocycle can be constructed using *any* coboundary map from the equivalence class of SCMs  $\{f^{(g)}\}_{g \in \text{Aut}(\mathcal{E})}$ . We are therefore free to choose a representative  $f^*$  that minimizes a given notion of functional complexity (e.g., Sobolev norm). Since each map  $f^{(g)}$  corresponds to a base distribution  $\mathbb{P}^{(g)}$  such that  $\mathbb{P}_{Y|X=x} = (f_x^{(g)})_\# \mathbb{P}^{(g)}$ , counterfactual cocycles are therefore no more complicated than the *minimal complexity* SCM, induced by an ‘optimal’ noise distribution  $\mathbb{P}^*$ .

This means we may be able to use simpler flow-based model classes  $\mathcal{F}$  than otherwise for the coboundary map of a cocycle, while remaining well-specified. Using a simpler class of models can improve finite sample performance via reduced estimation variance. In practice, we will cross-validate over a hierarchy of flows of increasing expressivity to adapt to the underlying complexity (see Section 7).

**Conditions for Stricter Cocycle Simplicity** It is natural to ask under what conditions a counterfactual cocycle can be constructed using a strictly simpler model class than for an SCM with a fixed base distribution  $\hat{\mathbb{P}}_\xi$ . Below we provide an exact characterization when the simplicity of a model class is measured by its size. In what follows we choose the noise space to be  $\mathcal{E} = \mathbb{Y}$  without loss of generality<sup>4</sup>, so that each  $f_x$  is bijective  $\mathbb{Y} \rightarrow \mathbb{Y}$ .

Recall from Section 3.2 that whenever the maps  $(f_x)_{x \in \mathbb{X}}$  are exactly invertible on  $\mathbb{Y}$ , they lie in a transformation group  $\mathbb{G}$  on  $\mathbb{Y}$ . In this case, we call the coboundary map (i.e.,  $f$ )  $\mathbb{G}$ -valued. The smallest group containing  $(f_x)_{x \in \mathbb{X}}$  is the subgroup generated by them, which we denote  $\mathbb{G}_f := \langle f_x : x \in \mathbb{X} \rangle$ . The set of all  $\mathbb{G}_f$ -valued coboundary maps is denoted  $\mathcal{F}_{\mathbb{G}_f}$ . This model class naturally reflects the smallest model class that is guaranteed to

---

4. Choosing  $\mathcal{E} = \mathbb{Y}$  just corresponds to a reparameterization of the BGM, since by construction there exists a bijection  $h : \mathbb{Y} \rightarrow \mathcal{E}$ . Moreover, the automorphism groups are conjugate:  $\text{Aut}(\mathcal{E}) \cong \text{Aut}(\mathbb{Y})$ , with  $g \in \text{Aut}(\mathcal{E})$  if and only if  $h \circ g \circ h^{-1} \in \text{Aut}(\mathbb{Y})$ . Thus, working with  $\mathcal{E} = \mathbb{Y}$  entails no loss of generality.

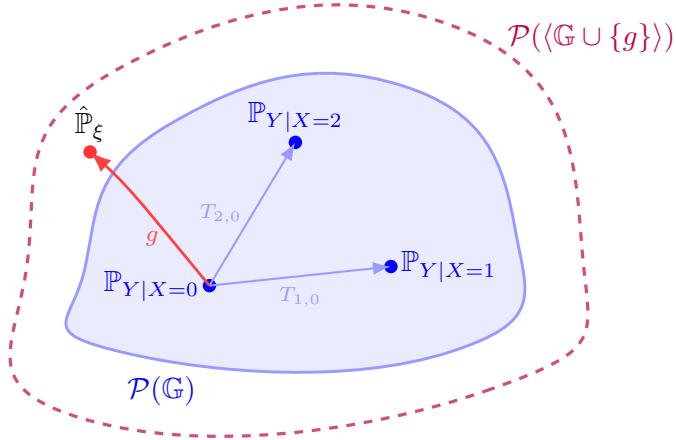


Figure 9: Illustration of Theorem 13 (minimum cocycle complexity). The blue region  $\mathcal{P}(G)$  shows the set of distributions reachable from the base distribution  $\mathbb{P}_{Y|X=0}$  under the group  $G = \langle T_{x,0} : x \in \{0, 1, 2\} \rangle$ . The family  $(T_{x,0})_{x \in \{0, 1, 2\}}$  defines a coboundary map  $f \in \mathcal{F}_G$  for the cocycle  $T$ , via  $f(x, y) = T_{x,0}(y)$ . If each  $T_{x,0}$  is composed with some  $g \notin G$ , the base distribution shifts to  $\hat{\mathbb{P}}_\xi = (g)_\# \mathbb{P}_{Y|X=0}$ , which lies outside  $\mathcal{P}(G)$ . The smallest group transporting to all  $(\mathbb{P}_{Y|X=x})_{x \in \{0, 1, 2\}}$  is then  $\langle G \cup \{g\} \rangle$ , which is strictly larger than  $G$  and has reachable set  $\mathcal{P}(\langle G \cup \{g\} \rangle)$ . Since enlarging the group enlarges the associated function class of coboundary maps (i.e.,  $\mathcal{F}_G \subsetneq \mathcal{F}_{\langle G \cup \{g\} \rangle}$ ), starting from  $\mathbb{P}_{Y|X=0}$  (or any  $\mathbb{P}_{Y|X=x}$ ) yields the smallest possible transformation group  $G$  and hence the smallest associated model class  $\mathcal{F}_G$  in which the cocycle can be represented.

be well-specified for  $f$ , when the dependence of the function  $f$  on  $x$  is not known. Note that size of this model class  $\mathcal{F}_{G_f}$  is controlled by the size of  $G_f$ : if two coboundary maps  $f_1$  and  $f_2$  generate groups  $G_{f_1}$  and  $G_{f_2}$  with  $G_{f_1} \subsetneq G_{f_2}$ , then  $\mathcal{F}_{G_{f_1}} \subsetneq \mathcal{F}_{G_{f_2}}$ . Altogether, this implies that if  $G_{f_1} \subsetneq G_{f_2}$ , then in a certain sense  $f_1$  can be modeled with a smaller (i.e., simpler) model class than  $f_2$ . For an intuitive example, let  $G_{f_1} := \text{Diff}^k(\mathbb{Y})$ , the set of diffeomorphisms on  $\mathbb{Y}$  with  $k$  continuous derivatives, and  $G_{f_2} := \text{Diff}^1(\mathbb{Y})$ . Then,  $\mathcal{F}_{G_{f_1}}$  contains only those functions in  $\mathcal{F}_{G_{f_2}}$  that are at least  $k$ -smooth on  $\mathbb{Y}$ .

Below we show that, out of all equivalence class members  $(f^{(g)})_{g \in \text{Aut}(\mathbb{Y})}$  that can be used to construct a given counterfactual cocycle, the construction  $f_x^* := T_{x,x_0}$  from Theorem 3 induces the smallest possible set  $\mathcal{F}_{G_{f^*}}$  and so can be modeled using the smallest class.

**Theorem 13** (Minimum Cocycle Complexity). *Let  $\{T_{x,x'}\}_{x,x' \in \mathbb{X}}$  satisfy (DA), (ID) and (PI) with respect to  $\mathbb{P}_{Y|X}$ . For each  $x \in \mathbb{X}$ , define  $f_x := T_{x,x_0}$  and suppose each  $f_x : \mathbb{Y} \rightarrow \mathbb{Y}$  is bijective. Let  $G_f = \langle f_x : x \in \mathbb{X} \rangle$  and  $G_{f(g)} = \langle f_x \circ g : x \in \mathbb{X} \rangle$ , for any  $g \in \text{Aut}(\mathbb{Y})$ . Then,*

$$(i) \mathcal{F}_{G_f} \subseteq \mathcal{F}_{G_{f(g)}} \quad \text{and} \quad (ii) \mathcal{F}_{G_f} \subsetneq \mathcal{F}_{G_{f(g)}} \quad \forall g \notin G_f .$$

Since choosing  $f_x^* := T_{x,x_0}$  corresponds to the base distribution  $\hat{\mathbb{P}}_\xi^* := \mathbb{P}_{Y|X=x_0}$  (i.e.,  $(f_x^*)_\# \mathbb{P}_{Y|X=x_0} = \mathbb{P}_{Y|X=x}$ ) and  $x_0$  is arbitrary above, Theorem 13 implies  $\mathbb{P}_{Y|X=x}$ , for any

$x \in \mathbb{X}$  is always an ‘optimal’ base distribution. Moreover, any choice of base distribution  $\tilde{\mathbb{P}}_\xi$  such that  $\tilde{\mathbb{P}}_\xi \neq g_{\#}\mathbb{P}_{Y|X=x}$  for all  $g \in \mathbb{G}_{f^*}$  must require a more complex coboundary map than can be used to construct the cocycle. This result may seem strange, since Fig. 5 illustrated an example where the valid set of noise distributions under a function class  $\mathcal{F}$  did not include either conditional. However, when the class  $\mathcal{F}$  is specified via a transformation group, the sets of reachable distributions from any two conditionals  $\mathbb{P}_{Y|X=x_0}, \mathbb{P}_{Y|X=x_1}$  either do not intersect (if the group is too small), or are identical (if the group is large enough to map between them). Fig. 9 illustrates the minimum complexity result in more detail.

As a concrete application of this result, in Example 3, we saw that  $f_x^* \in (\mathbb{R}^2, +)$ , the group of shifts on  $\mathbb{R}^2$ . Since there  $\mathbb{P}_\xi^* = \mathbb{P}_{Y|X=0}$ , we know that  $f^*(x, y) = Ax + y$  is a minimal complexity coboundary map for the cocycle. Moreover, if  $\hat{\mathbb{P}}_\xi$  is not a translation of  $\mathbb{P}_{Y|X=0}$  then the structural map from this base distribution will necessarily be more complex. Note that the transport in Fig. 8 corresponds to this map for  $x = 1$ , since  $T_{1,0}(y) = f^*(1, y)$ .

### 6.3 Robustness of Cocycle-based Approach to Causal Quantity Estimation

Since counterfactual cocycles are well-specified under strictly milder conditions than flow-based BCMs, any procedure for estimating causal quantities that depends only on the cocycle is naturally more robust. For instance, consider any estimand of the form

$$\gamma(x) = \mathbb{E}[\rho(Y(x), Y(0))], \quad \text{for some } \rho : \mathbb{R}^{2p} \rightarrow \mathbb{R},$$

(e.g., those in (20)-(22)). Suppose the true cocycle  $T$  can be constructed by a flow-based coboundary map parameterized in  $\mathbb{R}^d$ ,  $f_{\theta_0} \in \mathcal{F} = \{f_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ . Now, let  $\hat{\theta}$  be any estimator of  $\theta_0$  from  $\{X^{(i)}, Y^{(i)}\}_{i=1}^n$ . Our approach uses the cocycle to empirically estimate the causal estimand via the imputed counterfactuals

$$\hat{\gamma}(x) := \frac{1}{n} \sum_{i=1}^n \rho(T_{\hat{\theta}, x, X^{(i)}}(Y^{(i)}), T_{\hat{\theta}, x_0, X^{(i)}}(Y^{(i)})).$$

Under standard empirical process theory arguments on  $\theta \mapsto \mathbb{E}\rho(T_{\theta, x, X}(Y), T_{\theta, x_0, X}(Y))$  (e.g., see Kennedy (2016)), if  $\hat{\theta} \xrightarrow{p} \theta^*$ , then we have  $\hat{\gamma}(x) \xrightarrow{p} \gamma(x)$ . By contrast, the abduct–act–predict (AAP) estimator by Monte Carlo sampling from a flow-based BCM is

$$\hat{\gamma}^{\text{AAP}}(x) = \frac{1}{m} \sum_{i=1}^m \rho(f_{\hat{\theta}, x}(\xi^{(i)}), f_{\hat{\theta}, 0}(\xi^{(i)})), \quad \{\xi^{(i)}\}_{i=1}^m \sim_{iid} \hat{\mathbb{P}}_\xi.$$

Although one can appeal to equivalent empirical process conditions for this estimator, one can still have  $\hat{\gamma}^{\text{AAP}}(x) \not\rightarrow_p \gamma(x)$  even if the estimator for the cocycle is consistent,  $\hat{\theta} \rightarrow_p \theta_0$ . This is because, while  $f_{\theta_0} \in \mathcal{F}$ , we may have  $\hat{f} \notin \mathcal{F}$ , where  $\hat{f}$  is the conditional flow from  $\hat{\mathbb{P}}_\xi \mapsto \mathbb{P}_{Y|X}$  (e.g., if  $\hat{\mathbb{P}}_\xi$  has a mis-specified tail and support—as in Example 1 and 3).

As we showed in Section 5, our proposed cocycle estimator actually ensures  $\hat{\theta} \rightarrow_p \theta_0$  under *more general* conditions than existing estimators which use a base distribution to estimate the flow parameters, giving our approach an additional source of robustness.

## 6.4 Comparison to SCM-based Transport Methods via Conditional Quantiles

SCMs combining conditional-quantile estimation techniques with causal DAGs have recently been used to construct counterfactual transports in fairness applications (Plečko and Meinshausen, 2020; Machado et al., 2024). These are situated in a longer line of work in causality that uses the quantile transform (e.g., Chernozhukov and Hansen, 2005; Athey and Imbens, 2006; Vansteelandt and Joffe, 2014). The methods in Plečko and Meinshausen (2020) and Machado et al. (2024) start from an SCM with independent, uniform noise,

$$Y_j =_{\text{a.s.}} f_j(Y_{\text{pa}(j)}, \xi_j), \quad \xi_j \sim U[0, 1] \quad \forall j \in \{1, \dots, p\}.$$

In their context,  $X := Y_0$  is a “sensitive” source attribute (e.g., sex, race) to be manipulated, and  $Y_1, \dots, Y_p$  are downstream outcomes that follow a known causal DAG. The main idea is that, under the independent, uniform noise assumption,  $f_j$  coincides with the conditional quantile function,  $f_j(y_{\text{pa}(j)}, \xi_j) = Q_{y_{\text{pa}(j)}}(\xi_j)$ , of the conditional law  $\mathbb{P}_{Y_j|Y_{\text{pa}(j)}=y_{\text{pa}(j)}}$ . It can therefore be estimated using standard quantile regression techniques (Meinshausen and Ridgeway, 2006). To generate counterfactuals under the intervention  $\text{do}(X = x)$  one can recursively compute for each node in the DAG,

$$\hat{y}_j(x) = Q_{\hat{y}_{\text{pa}(j)}(x)} \circ F_{y_{\text{pa}(j)}}(y_j), \quad \forall j = 1, \dots, p$$

where  $F_{y_{\text{pa}(j)}}$  is the conditional CDF given the *factual* parent values  $y_{\text{pa}(j)}$ , and  $\hat{y}_{\text{pa}(j)}(x)$  are the counterfactual values of the parents obtained in previous steps. Thus, each node is updated by a map of the form  $T_{\hat{y}_{\text{pa}(j)}(x), y_{\text{pa}(j)}} = Q_{\hat{y}_{\text{pa}(j)}(x)} \circ F_{y_{\text{pa}(j)}}$ , which is precisely the conditional OT map between  $\mathbb{P}_{Y_j|Y_{\text{pa}(j)}=y_{\text{pa}(j)}}$  and  $\mathbb{P}_{Y_j|Y_{\text{pa}(j)}=\hat{y}_{\text{pa}(j)}(x)}$  (Carlier et al., 2016; Hosseini et al., 2025). The resulting procedure recently has been branded sequential-OT (Machado et al., 2024), since it decomposes the transport into a set of conditional OT maps.

When the true noise variables  $(\xi_j)_{j=1}^p$  are independent, the sequential-OT transports coincide with the (TMI) counterfactual cocycle. However, the conditional-quantile approach requires estimating the mapping  $(Q_x)_\# : U[0, 1]^p \mapsto \prod_{j=1}^p \mathbb{P}_{Y_j|Y_{\text{pa}(j)}}$ , and so still implicitly commits to a particular noise distribution for the purposes of estimation. Thus, in principle, this approach can suffer from similar mis-specification problems as flow-based SCMs, depending on the exact function class used to learn the quantile function. This is somewhat mitigated by using nonparametric CDF estimation techniques, but in either case fails to exploit the minimal complexity of counterfactual cocycles.

A potentially more serious limitation is that the resulting transports generally do *not* coincide with the true counterfactual cocycle when the noise variables are dependent, thus resulting in biased, inconsistent estimators. This is demonstrated in the following example.

**Example 5.** Consider the three-variable SCM with treatment variable  $X$ ,

$$X \sim N(0, 1), \quad Y_1 = X + \xi_1, \quad Y_2 = Y_1 + \xi_2,$$

We assume  $(\xi_1, \xi_2) \sim N(0, \Sigma_\rho)$ ,  $\text{Var}(\xi_j) = 1$ ,  $\text{Corr}(\xi_1, \xi_2) = \rho$ , and  $(\xi_1, \xi_2) \perp\!\!\!\perp X$ . In this case, the true counterfactual cocycle reduces to a joint shift

$$T_{x',x}^\star(y_1, y_2) = (y_1 + \Delta, y_2 + \Delta), \quad \Delta = x' - x.$$

By contrast, the sequential-OT procedure applies conditional quantile maps node by node. For  $Y_1$ , since  $(Y_1|X=x) \sim N(x, 1)$ , the quantile transform gives  $\hat{y}_1(x') = y_1 + \Delta$ . For  $Y_2$ , since  $(Y_2|Y_1=y_1) \sim N((1+\frac{\rho}{2})y_1, 1-\rho^2/2)$ , the conditional quantile transform gives

$$\hat{y}_2(x') = (1 + \frac{\rho}{2})\hat{y}_1(x') + \sqrt{1 - \frac{\rho^2}{2}} \Phi^{-1}(\hat{\xi}_2) = y_2 + (1 + \frac{\rho}{2})\Delta.$$

Combining these transports together gives the sequential-OT map

$$T_{x',x}^{\text{seq}}(y_1, y_2) = (y_1 + \Delta, y_2 + (1 + \frac{\rho}{2})\Delta),$$

which coincides with  $T_{x',x}^*$  only when  $\rho = 0$ . The discrepancy arises because the TMI map between joint laws  $\mathbb{P}_{Y_1, Y_2|X}$  does not factorize into the TMI maps of the conditionals  $\mathbb{P}_{Y_1|X} \otimes \mathbb{P}_{Y_2|Y_1}$  when  $\xi_1, \xi_2$  are dependent. By targeting the joint cocycle directly, our method avoids this issue and can recover the correct transport regardless of  $\rho$ , even when using the DAG structure to sparsify the cocycle architecture.

## 7 Implementation Details

In this section, we discuss implementation and optimization details for counterfactual cocycle models parameterized by autoregressive flows. Formal algorithms are in Section B.

### 7.1 Flow-based Cocycle Parameterizations

As discussed in Section 3, a natural modeling choice for the coboundary map  $(x, y) \mapsto f_x(y)$  of a cocycle is to use conditional normalizing flows (Kobyzev et al., 2020; Papamakarios et al., 2021). Each flow is the composition of (i) a conditioner  $\tau_\theta: \mathbb{X} \rightarrow \Lambda$ , mapping inputs  $x$  to a vector of flow parameters  $\lambda$ , and (ii) a bijector  $g_\lambda \in \mathbb{G}$  that transforms  $y$ :

$$\hat{f}_{\theta,x}(y) = g_{\tau_\theta(x)}(y) \Leftrightarrow T_{\theta,x,x'}(y) = g_{\tau_\theta(x)} \circ g_{\tau_\theta(x')}^{-1}(y)$$

The conditioner  $\tau_\theta$  can be *any* learnable function class—linear model, MLP, convolutional network, or transformer—so long as it maps  $x$  to valid flow parameters  $\lambda$ . The bijector  $g_\lambda$  can be a single transform or a multi-layer composition of such transforms that lie in  $\mathbb{G}_{\text{TMI}}$ . Popular examples of such transforms are given by the autoregressive flows in Table 1. Table 2 presents several cocycles constructed using simple transforms and existing autoregressive flows, together with the lower triangular restrictions under a known partial ordering of the variables in  $Y$  used to preserve identifiability. When further constraining these flows using a known causal DAG, we follow Javaloy et al. (2023) and specify the inverse map  $f_x^{-1}$  as the (forward) autoregressive flow with a mask on the adjacency matrix reflecting the sparsity of the DAG. When using a single layer this prevents spurious correlations from being induced by the architecture (see Javaloy et al. (2023) for details).

### 7.2 Optimization and Model Selection

**CMMMD Implementation** We optimize all flow-based cocycles using gradient descent on our empirical CMMMD losses (V/U-statistic) introduced in Section 3. Our default kernel choice for CMMMD is the Gaussian kernel  $k(y, y') = \exp(-\lambda \|y - y'\|^2)$ , where the bandwidth

Table 2: Example cocycle parameterizations with classes of TMI maps. Here  $\mathbb{G}$  denotes the transformation group of the cocycle. MAF = Masked Autoregressive Flow (Papamakarios et al., 2017); NSF = Neural Spline Flow (Durkan et al., 2019).

Transformation group $\mathbb{G}$	Conditioner output	Cocycle $T_{x',x}(y)$	Restriction for identifiability
Shifts $(\mathbb{R}^d, +)$	$a_x \in \mathbb{R}^d$	$y + a_{x'} - a_x$	None
$\text{GL}_+(\mathbb{R}^d)$	$A_x \in \text{GL}_+(\mathbb{R}^d)$	$A_{x'} A_x^{-1} y$	$A_x$ lower-triangular, $[A_x]_{jj} \geq 0$
$\text{GA}_+(\mathbb{R}^d)$	$(A_x, a_x)$	$a_{x'} + A_{x'} A_x^{-1} (y - a_x)$	$A_x$ lower-triangular, $[A_x]_{jj} \geq 0$
$\text{Diff}(\mathbb{R}^d) \cap \mathbb{G}_{\text{TMI}}$	MAF parameters $\theta_x$	$\text{MAF}^{-1}[\theta_{x'}] \circ \text{MAF}[\theta_x](y)$	None
$\text{Diff}(\mathbb{R}^d) \cap \mathbb{G}_{\text{TMI}}$	NSF parameters $\theta_x$	$\text{NSF}^{-1}[\theta_{x'}] \circ \text{NSF}[\theta_x](y)$	None

$\lambda$  is chosen using the median heuristic on the observations  $\{Y^{(i)}\}_{i=1}^n$  (Garreau et al., 2017). For gradient-based optimization, since both evaluation and gradients have a computational complexity of  $\mathcal{O}(n^3)$ , at each iteration we subsample  $B \ll n$  datapoints, and then approximate  $\nabla_\theta \ell_n^U$  (resp.  $\nabla_\theta \ell_n^V$ ) with  $\nabla_\theta \ell_B^U$  (resp.  $\nabla_\theta \ell_B^V$ ). This stochastic optimization approach has been used for kernel-based estimators in several works (Greenfeld and Shalit, 2020; Jankowiak and Pleiss, 2021; Dance and Paige, 2022) and, in our case, estimates  $\ell_n^U$  without bias and  $\ell_n^V$  with a bias of order  $1/B$ . We use the ADAM optimizer (Kingma and Ba, 2014) with a default batch size is  $B = \min(n, 128)$ . Algorithm 1 in Section B presents pseudo-code for this procedure, for the V-statistic.

**Model Selection** As established in earlier sections, a key advantage to modeling counterfactual cocycles rather than an SCM with a base distribution, is that we maximize the ‘chance’ of remaining well-specified using a simpler conditional flow (i.e., one contained in a smaller transformation group  $\mathbb{G}$ ). We therefore advocate training a *hierarchy* of flow-based cocycle classes with increasing transformation group expressivity, such as those presented in Table 2—coupled with conditioners of matched or increasing capacity. In practice we use  $K$ -fold cross-validation to do this. Algorithm 2 in Section B demonstrates the procedure.

## 8 Experiments

We now implement counterfactual cocycles in a range of simulations and a real application, and compare against state-of-the-art SCM and OT methods for counterfactuals. Code for our method and experiments can be found at <https://github.com/HWDance/Cocycles>.

### 8.1 Noise Ablation in Flow-based SCMs

We begin by comparing counterfactual cocycles to equivalent flow-based SCMs in a simple linear causal model. The goal of this experiment is to assess the robustness of our estimation approach to the underlying noise distribution  $\mathbb{P}_\xi$ , and to illustrate how our method can leverage the simplicity of the underlying cocycle, in contrast to the flow that arises from pushing forward a fixed base distribution  $\mathbb{P}_0$ .

**Experimental Set-up** We generate  $n = 1000$  points from a linear structural model  $Y = \beta X + \xi$  under different settings of  $\mathbb{P}_\xi$ . We implement cocycles trained using both CMMD-V and CMMD-U losses implemented as in Section 7.2 (see Algorithm 1), and benchmark

Method	$\mathbb{P}_\xi = \mathcal{N}(0, 1)$	$\mathbb{P}_\xi = \mathcal{G}(1, 1)$	$\mathbb{P}_\xi = t_1(0, 1)$	$\mathbb{P}_\xi = IG(1, 1)$	$\mathbb{P}_\xi = Rad(1/2)$
<b>Interventional KS</b>					
ML-G	<b>0.015 ± 0.007</b>	0.081 ± 0.041	0.121 ± 0.079	0.208 ± 0.162	0.405 ± 0.069
ML-L	0.055 ± 0.010	0.074 ± 0.036	0.110 ± 0.069	0.120 ± 0.090	0.415 ± 0.067
ML-T	0.018 ± 0.008	0.079 ± 0.040	0.029 ± 0.007	0.075 ± 0.031	0.412 ± 0.062
CMMMD-V	0.028 ± 0.009	<b>0.027 ± 0.008</b>	<b>0.027 ± 0.009</b>	<b>0.027 ± 0.008</b>	<b>0.271 ± 0.031</b>
CMMMD-U	<b>0.010 ± 0.004</b>	<b>0.012 ± 0.005</b>	<b>0.008 ± 0.003</b>	<b>0.009 ± 0.004</b>	<b>0.268 ± 0.008</b>
<b>Counterfactual RMSE</b>					
ML-G	<b>0.035 ± 0.035</b>	0.277 ± 0.118	113.667 ± 341.875	114.946 ± 147.841	0.326 ± 0.258
ML-L	0.036 ± 0.028	0.258 ± 0.073	97.745 ± 351.815	112.300 ± 165.014	0.480 ± 0.294
ML-T	<b>0.033 ± 0.031</b>	0.270 ± 0.097	0.044 ± 0.053	29.872 ± 41.628	0.391 ± 0.307
CMMMD-V	<b>0.035 ± 0.026</b>	<b>0.020 ± 0.015</b>	<b>0.040 ± 0.031</b>	<b>0.028 ± 0.024</b>	<b>0.017 ± 0.019</b>
CMMMD-U	0.040 ± 0.027	<b>0.022 ± 0.016</b>	<b>0.033 ± 0.027</b>	<b>0.027 ± 0.023</b>	<b>0.014 ± 0.011</b>
<b>True Architecture Selection %</b>					
ML-G	96%	14%	0%	2%	2%
ML-L	<b>100%</b>	2%	4%	0%	0%
ML-T	98%	8%	94%	0%	0%
CMMMD-V	<b>100%</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>	<b>98%</b>
CMMMD-U	<b>100%</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>

Table 3: Mean  $\pm$  SE of the interventional Kolmogorov—Smirnov distance (top block) and counterfactual RMSE (middle block), averaged over 50 trials, plus the percent of correct architecture selections (bottom block), for SCM  $Y = X + \xi$  under  $\text{do}(X = 0)$  across different noise laws. “ML-” denotes ML flows with Gaussian (G), Laplace (L) or Student- $t$  (T) bases; “CMMMD-” denotes our cocycle estimators (CMMMD-V/U). Boldface marks the top two performers per column.

them against flow-based SCMs which estimate the bijective generative mechanism  $(f, \mathbb{P}_\xi)$  via maximum likelihood, using different base distributions ( $\mathcal{N}(\mu, \sigma^2)$ ,  $\text{Lap}(\mu, \sigma^2)$ ,  $t_\nu(\mu, \sigma^2)$ ) with learnable parameters. For all methods, we perform 2-fold cross-validation across flow architectures of increasing complexity: (i) linear  $f_x(\xi) = \theta x + \xi$ ; (ii) additive  $f_x(\xi) = m_\theta(x) + \xi$  with  $m_\theta$  an MLP; (iii) a one-layer MAF (Papamakarios et al., 2017); (iv) a one-layer NSF (Durkan et al., 2019) with 8 spline bins, flanked by affine transforms. In (iii)–(iv), we use the standard autoregressive architectures implemented in the `zuko` library, extended to also condition on  $x$  via neural networks. All neural networks are set as MLPs with  $L = 2$  layers and  $W = 32$  nodes per layer. All models are trained in PyTorch using gradient descent on the respective objective functions (i.e., CMMMD, ML) with the Adam optimizer with learning rate=0.01 (except NSF with 0.001), batch size=128 and 1000 epochs.

**Results** Table 3 shows performance results for different true noise distributions under a hard intervention  $\text{do}(X = 0)$ , averaged over 50 seeds. The top block shows the Kolmogorov—Smirnov (KS) distance between the true and estimated marginal distribution of  $Y(0)$ ; the middle block shows the RMSE between the true and estimated counterfactuals  $Y(0)$  for the units in the dataset  $\{X^{(i)}, Y^{(i)}\}_{i=1}^n$ , and the bottom block shows the fraction of trials on which the true linear architecture was selected by each method. When the true noise is Gaussian, all methods performed similarly well for both metrics, and selected the true

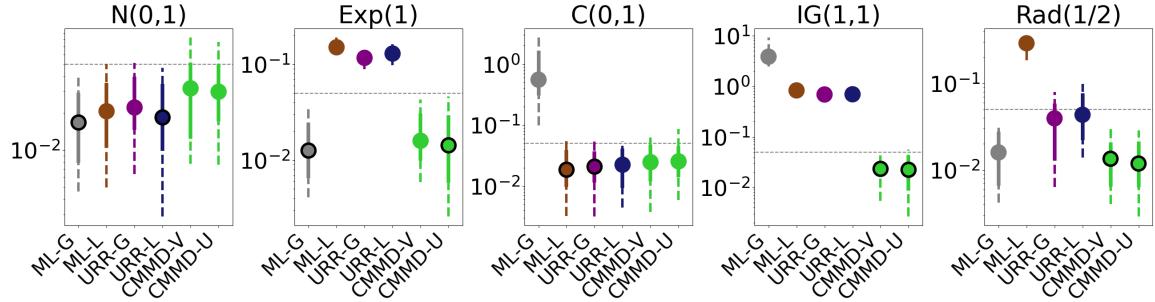


Figure 10: Mean (circle), (25-75) percentiles (solid line) and (10-90) percentiles (dashed line) of absolute error in treatment effect  $Y(X + 1) - Y(X) = 1$  for different estimation methods in a linear model  $Y = X + \xi$  under different noise designs, when fixing the flow to be the true architecture  $f_x(\xi) = \beta x + \xi$ . Dashed horizontal line = 5% error. Black edges = best two methods on average.

linear architecture in almost all cases. This reflects the fact that the base distributions are either well-specified (i.e., Normal and Student’s t) or close enough to well-specified (i.e., Laplace). The performance of the cocycle-based approach is roughly invariant to the true noise distribution, reflecting the robustness of our estimation approach to this aspect of the data generating process. The one exception is interventional KS for  $\mathbb{P}_\xi = \text{Rad}(1/2)$ , which is naturally higher since the true interventional distribution lies on two points and so is very sensitive to imperfect matching. In contrast, the performance of all other methods deteriorates drastically when the noise distribution does not match the specified base distribution, and a more complex flow is chosen in an attempt to correct for this mis-specification. As a result, our cocycle estimators performed best for all non-Gaussian noise distributions. For counterfactual RMSE, the performance gain is in some cases more than two orders of magnitude.

For the interventional KS, the performance gap was greatest when  $\mathbb{P}_\xi \in \{t_1(0, 1), \text{IG}(1, 1), \text{Rad}(1/2)\}$ , reflecting that the base distributions have mis-specified tails or support in these cases (except the Student’s-t base in the  $t_1(0, 1)$  noise case). For  $\mathbb{P}_\xi = \text{Ga}(1, 1)$ , there is still a substantial performance gap, since the flow-based methods need to use a very complex neural spline flow in order to effectively learn the base distribution, which results in worse finite sample performance. For counterfactual RMSE, all methods compute counterfactuals using the same formula:  $Y(0) = f_0 \circ f_X^{-1}(Y)$ . Thus, the performance gain using cocycle targeting here purely reflects the robustness of our CMMMD estimator and that we do not need a more complex flow to compensate for a poorly matched base distribution.

**Extension** To analyze how the CMMMD estimator performs when all architectures are fixed, in Figure 10 we also produce counterfactual RMSE results under a shift intervention  $X \mapsto X + 1$ , when restricting all flow architectures to be the true (linear) architecture  $f_x(\xi) = \beta x + \xi$ . Note that in this case,  $\hat{Y}(X + 1) - Y(X + 1) = \hat{\beta} - \beta$ , so counterfactual error is isometric to cocycle estimation error. For this we compare the CMMMD estimator against ML estimators with the Gaussian and Laplace base distributions (i.e.,  $\ell_2$  and  $\ell_1$  regression), as well as a recently proposed MMD-based estimator for conditional generative

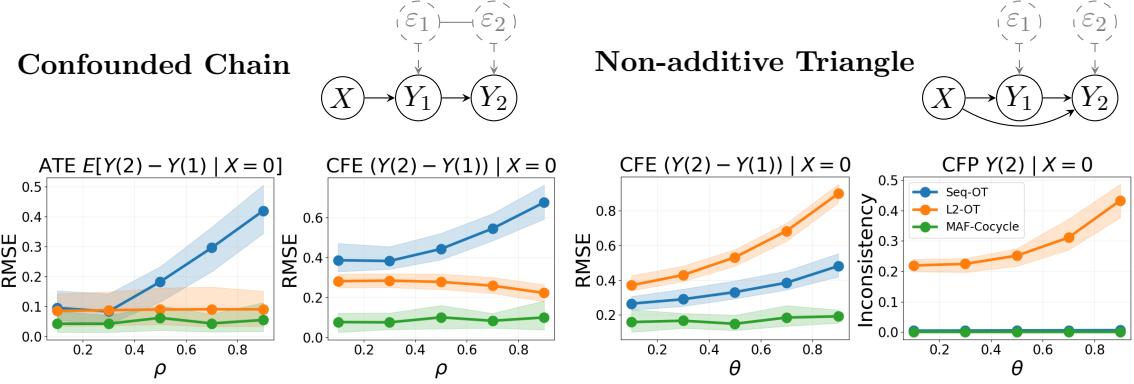


Figure 11: Left and middle left (Confounded Chain): Mean  $\pm$  SD (20 trials) of error in estimated average incremental effect (left) and incremental counterfactual effect (middle left) of second treatment, for observed units  $\{Y^{(i)}(0)\}_{i=1}^n$  in control group. blue = sequential optimal transport (Machado et al., 2024), orange = optimal transport (Brenier maps) (De Lara et al., 2024), green = Masked Autoregressive Flow cocycle. Under the chain DAG, Sequential-OT becomes biased as correlation strength  $\rho := \text{Corr}(\xi_1, \xi_2)$  increases, whereas the MAF-cocycle does not, due to the robustness of our estimation procedure. Right and middle right (Non-additive Triangle): Mean  $\pm$  SD (20 trials) error in incremental counterfactual effect (middle right) and inconsistency in counterfactual predictions  $Y(2)$  (right) when imputing for observed units  $\{Y^{(i)}(0)\}_{i=1}^n$  in control group, via (a) indirect  $T_{2,1} \circ T_{1,0}$  and (b) direct  $T_{2,0}$  transports.  $\theta$  controls degree of non-additivity in true transport maps. As  $\theta$  increases, OT performance degrades and transport inconsistency increases, reflecting underlying model incoherence.

models: Universal Robust Regression (URR) (Alquier and Gerber, 2023). The latter bears similarities with the CMMD estimator, with the exception that it requires specifying the full generative model (i.e., a base distribution). For URR we optimize the conditional estimator Eq. (5) from (Alquier and Gerber, 2023), using  $m = n$  Monte Carlo samples from  $\widehat{\mathbb{P}}_{Y|X}^\theta$ . The kernel is chosen identically to CMMD. Both CMMD estimators estimate  $\beta$  within 5% error on average across all noise designs, reflecting its noise-robustness. By contrast, all other estimators perform poorly (i.e.,  $\gg 50\%$  error) on at least two noise distributions.

## 8.2 Confounding and Path-Consistency Ablation in Transport-based Models

In this experiment, we compare counterfactual cocycles against OT-based approaches and assess how estimation accuracy and path-consistency varies under different assumptions.

**Experimental Set-up** Suppose we have collected data under a randomized controlled trial which tests two treatments and a control, i.e.,  $X \in \{0, 1, 2\}$ . We observe two outcomes  $Y := (Y_1, Y_2)$  for  $n = 500$  patients under control,  $\{Y^{(i)}(0)\}_{i=1}^n$ , under treatment  $X = 1$ ,

$\{Y^{(i)}(1)\}_{i=n}^{2n}$ , and under alternative treatment  $X = 2$ ,  $\{Y^{(i)}(2)\}_{i=2n}^{3n}$ ). We aim to estimate the incremental effectiveness of treatment  $X = 2$  by estimating transports  $T_{0,1}, T_{0,2}, T_{1,2}$  between each state, and using them to compute the contrast  $Y(2) - Y(1)$  for each unit. For ease of demonstrating path-consistency issues, we only do this for the control group units.

We consider two generating designs to isolate different weaknesses of competing methods: (i) a chain SCM  $Y_1 = X + \xi_1$ ,  $Y_2 = Y_1 + \xi_2$  with  $(\xi_1, \xi_2) \sim \text{Lap}(\mathbf{0}, \rho \mathbf{1}\mathbf{1}^\top + (1 - \rho)I)$ ; and (ii) a non-additive SCM  $\mathbf{Y} = \mathbf{1}X + L(X)\boldsymbol{\xi}$  with  $\boldsymbol{\xi} \sim \text{Lap}(0, I_2)$  and

$$L(0) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad L(1) = \begin{pmatrix} 1 & -\theta \\ -\theta & 1 \end{pmatrix}, \quad L(2) = \begin{pmatrix} \theta + 1 & 1 \\ 1 & (1 + \theta)^{-1} \end{pmatrix}$$

In design (i) we vary the noise correlation  $\rho \in [0, 1]$ ; in design (ii) we vary  $\theta \in [0, 1]$ , which intuitively controls the degree of non-additivity in the true transport maps. The former lets us analyze how noise dependence affects estimation performance of sequential-OT methods, while the latter lets us analyze how non-additivity affects path-consistency of OT methods (in general, OT maps are coherent between distributions which are shifts of one another).

For the cocycle, we use a single layer MAF (Papamakarios et al., 2017) per treatment level  $x \in \{0, 1, 2\}$ , i.e.,  $f_x := \text{MAF}[\theta_x]$  with  $\theta_x = (\theta_0, \theta_1, \theta_2)[x]$ . The network that parameterizes each conditional affine transform of  $y_2$  given  $y_1$  is set as a two-layer MLP with  $L = 2$  layers and  $W = 32$  nodes per layer. We optimize the cocycle to target (DA) w.r.t.  $\mathbb{P}_{Y_1, Y_2 | X}$  via CMMMD using Algorithm 1. We use the same optimization settings as Section 8.1.

We compare our flow-based cocycle against OT with quadratic cost (De Lara et al., 2024; Charpentier et al., 2023; Balakrishnan et al., 2025), implemented using the network-simplex solver and barycentric projection via the POT package, following De Lara et al. (2024). These methods directly estimate  $T_{x,x'}$  between each pair of treatment levels without enforcing path-consistency (PI). When the true transport maps are not additive (as in design (ii)), the resulting pairwise maps may be mutually inconsistent. We also compare against the sequential-OT approach discussed in Section 6.4 (Plečko and Meinshausen, 2020; Machado et al., 2024). This approach factorizes the joint transport into a set of (1D) conditional transports of each node given its parents in the DAG, with each transport estimated via the conditional quantile transform. We estimate the conditional CDFs using (Gaussian) kernel smoothing, with bandwidths chosen via the median heuristic (Garreau et al., 2017). Sequential-OT is expected to guarantee (PI) in the large sample limit, since each coordinate map satisfies (PI) in this limit. However, given our analysis in Section 6.4, we expect performance to degrade as the noise correlation increases (as in design (i)).

**Results** For design (i), Fig. 11 reports the RMSE in the estimated average contrast  $\mathbb{E}[Y(2) - Y(1)|X = 0]$  (ATE) (left), and the RMSE in the counterfactual effect  $Y(2) - Y(1)$  (CFE) for control group units (middle left), over different noise correlation levels (averaged over 20 trials). Increasing the noise correlation  $\rho$  leaves cocycle and OT performance essentially unchanged, while sequential OT shows steadily increasing bias in both ATE and CFE. This is to be expected given the analysis in Section 6.4. Note, by default we impute counterfactuals using  $T_{1,0}$  and  $T_{2,0}$ , i.e.,  $(\hat{Y}(1), \hat{Y}(2)) = (T_{1,0}(Y(0)), T_{2,0}(Y(0)))$ .

For design (ii), Fig. 11 reports the RMSE of the CFE again (middle right) as well as the estimated RMSE between the predicted counterfactuals  $\hat{Y}(2)$ , when imputing them from the control group units via the direct transport  $T_{2,0}$ , and the indirect composition

$T_{2,1} \circ T_{1,0}$  (right). Note that each approach is equally valid here, as to impute both  $Y(1)$  and  $Y(2)$  for control-group units two out of three maps are always needed. However, each pair may induce different couplings over all three states. As expected, increasing the ‘non-additivity’ parameter  $\theta$  leads to large increases in both counterfactual RMSE and path-inconsistency for global OT. This shows that the incoherence of OT transports and resultant non-identifiability problem generalizes beyond the 2D Gaussian example in Section 2.2. Sequential OT is less affected here, but its performance remains inferior to the MAF-cocycle.

### 8.3 Performance on SCM Benchmarks

We now assess how our method performs on causal benchmarks used in the SCM literature, against state-of-the-art flow-based SCMs.

**Experimental Setup** We consider linear and non-linear variants of the following benchmark SCMs used in previous work (Geffner et al., 2022; Javaloy et al., 2023): (i) **Triangle**, a 3-node SCM with a dense causal graph; (ii) **Fork** a 4-node SCM with a sparse causal graph; and (iii) **5-chain**, a 5-node SCM with a chain structure. The linear and non-linear mechanisms for **Triangle**, **Fork** and **5-chain** can be found in Javaloy et al. (2023). We also implement a two-variable SCM (2var (lin):  $Y = X + \xi$ , 2var (nonlin):  $Y = \sin(X) + \xi$ ). Unlike in previous implementations where all noise distributions were Gaussian, we set each node in the SCM with a different noise distribution, enabling us to assess performance in a more challenging and realistic setting. In particular, we set  $\xi_1 = N(0, 1)$  and  $\xi_{d:2} = \{\text{IG}(1, 1), \text{Rad}(1/2), \frac{1}{2}N(-\sqrt{3}/2, 1/2) + \frac{1}{2}N(\sqrt{3}/2, 1/2), \text{Ga}(1, 1)\}$ .

We implement flow-based cocycles on  $(X, Y) = (V_1, V_{>1})$  with the CMMDF-V loss, against several state-of-the-art flow-based SCMs: (i) CAREFL (Khemakhem et al., 2021), which uses affine autoregressive flows to learn  $\mathbb{P}_{X,Y}$  (ii) CAUSALNF (Javaloy et al., 2023), which extends CAREFL to arbitrary flows but enforces a single (abductive) flow layer to prevent the flow from enforcing spurious edges in the adjacency matrix, and (iii) BGM (Nasr-Esfahany et al., 2023), which trains a conditional flow to match  $\mathbb{P}_{Y|X}$  and uses the empirical  $\widehat{\mathbb{P}}_X$ . For all methods (including ours) we assume the causal ordering is known, but the DAG is unknown. Hence, all autoregressive network architectures are dense. We use the same architectures, training and cross-validation procedure for all methods as in Section 8.1. However, note CAREFL is restricted to affine architectures, and we additionally cross-validate over Gaussian and Laplace base distributions for the baselines.

Table 4: Mean  $\pm$  SD of  $\text{KS}_{\text{int}}$  and  $\text{KS}_{\text{CF}}$  on the *linear* SCMs.

Method	2var (lin)		triangle (lin)		fork (lin)		5chain (lin)	
	$\text{KS}_{\text{int}}$	$\text{KS}_{\text{CF}}$	$\text{KS}_{\text{int}}$	$\text{KS}_{\text{CF}}$	$\text{KS}_{\text{int}}$	$\text{KS}_{\text{CF}}$	$\text{KS}_{\text{int}}$	$\text{KS}_{\text{CF}}$
BGM	$0.13 \pm 0.07$	$0.19 \pm 0.12$	$0.37 \pm 0.05$	$0.06 \pm 0.01$	$0.05 \pm 0.07$	$0.61 \pm 0.07$	$0.14 \pm 0.05$	$0.06 \pm 0.01$
CausalNF	$0.31 \pm 0.11$	$0.24 \pm 0.10$	$0.44 \pm 0.05$	$0.06 \pm 0.02$	$0.04 \pm 0.02$	$0.66 \pm 0.09$	$0.14 \pm 0.02$	$0.06 \pm 0.01$
CAREFL	$0.40 \pm 0.04$	$0.15 \pm 0.14$	$0.43 \pm 0.05$	$0.06 \pm 0.01$	$0.19 \pm 0.01$	$0.58 \pm 0.10$	$0.13 \pm 0.02$	$0.06 \pm 0.01$
CocycleNF	<b><math>0.03 \pm 0.02</math></b>	<b><math>0.04 \pm 0.04</math></b>	<b><math>0.23 \pm 0.19</math></b>	<b><math>0.02 \pm 0.01</math></b>	<b><math>0.02 \pm 0.01</math></b>	<b><math>0.19 \pm 0.23</math></b>	<b><math>0.02 \pm 0.01</math></b>	<b><math>0.03 \pm 0.01</math></b>

**Results** For each method we evaluate the learned distribution of  $V_{>1}(0)$  and the distribution of the current policy-effect  $V_{>1} - V_{>1}(0)$  under the intervention  $\text{do}(V_1 = 0)$ , by

Table 5: Mean  $\pm$  SD of  $\text{KS}_{\text{int}}$  and  $\text{KS}_{\text{CF}}$  on the *nonlinear* SCMs.

Method	2var (nonlin)		triangle (nonlin)		fork (nonlin)		5chain (nonlin)	
	$\text{KS}_{\text{int}}$	$\text{KS}_{\text{CF}}$	$\text{KS}_{\text{int}}$	$\text{KS}_{\text{CF}}$	$\text{KS}_{\text{int}}$	$\text{KS}_{\text{CF}}$	$\text{KS}_{\text{int}}$	$\text{KS}_{\text{CF}}$
BGM	$0.12 \pm 0.07$	$0.27 \pm 0.13$	$0.41 \pm 0.07$	<b><math>0.09 \pm 0.05</math></b>	<b><math>0.04 \pm 0.01</math></b>	$0.59 \pm 0.08$	$0.07 \pm 0.03$	$0.41 \pm 0.12$
CausalNF	$0.30 \pm 0.11$	$0.26 \pm 0.08$	$0.47 \pm 0.04$	$0.23 \pm 0.08$	$0.07 \pm 0.06$	$0.61 \pm 0.09$	$0.06 \pm 0.06$	$0.24 \pm 0.16$
CAREFL	$0.44 \pm 0.04$	$0.22 \pm 0.15$	$0.47 \pm 0.06$	$0.22 \pm 0.09$	$0.19 \pm 0.01$	$0.51 \pm 0.21$	$0.17 \pm 0.04$	$0.60 \pm 0.25$
CocycleNF	<b><math>0.04 \pm 0.02</math></b>	<b><math>0.13 \pm 0.05</math></b>	<b><math>0.28 \pm 0.13</math></b>	$0.20 \pm 0.16$	$0.08 \pm 0.05$	<b><math>0.20 \pm 0.24</math></b>	<b><math>0.05 \pm 0.02</math></b>	<b><math>0.20 \pm 0.09</math></b>

computing the *average marginal Kolmogorov–Smirnov (KS) distance*:

$$\text{KS}_{\text{int}} = \frac{1}{d-1} \sum_{j=2}^d D_{\text{KS}}(\hat{F}_{V_j(0)}, F_{V_j(0)}), \quad \text{KS}_{\text{CF}} = \frac{1}{d-1} \sum_{j=2}^d D_{\text{KS}}(\hat{F}_{V_j-V_j(0)}, F_{V_j-V_j(0)}),$$

Here  $\hat{F}$  is the CDF, which for flow-based cocycles is estimated empirically using the imputed counterfactuals  $\hat{V}_{>1}^{(i)}(0) = \hat{T}_{0,V_1^{(i)}}(V_{>1}^{(i)})$ , and for baselines is estimated via the abduct-act-predict procedure in Section 2.1, with  $m = 10^5$  Monte Carlo samples. Tables 4 and 5 report the mean  $\pm$  SD results from 10 trials. Cocycles achieves the best  $\text{KS}_{\text{int}}$  and  $\text{KS}_{\text{CF}}$  in all SCMs except one (**Triangle-nonlin** for  $\text{KS}_{\text{CF}}$  and **Fork-nonlin** for  $\text{KS}_{\text{int}}$ ). The performance gap is generally greatest when the true DGP is linear, reflecting our method’s ability to be well-specified with simpler architectures. Out of the baselines, CAREFL performed worst, likely reflecting its restriction to affine flows, while BGM generally performed best. The latter is to be expected, as under the intervention  $\text{do}(V_1 = 0)$ , the only part of the flow used at test time is the component which is directly optimized by BGM,  $f : (\xi_1, \xi_{>1}) \mapsto f_{\xi_1}(\xi_{>1})$ .

#### 8.4 Application: Counterfactual Effects of 401(k) Pension Plan Eligibility

As an application to real data, we use counterfactual cocycles to estimate the impact of 401(k) eligibility on net financial assets, using the well-known economic dataset studied in Chernozhukov and Hansen (2004). The dataset contains  $n = 9915$  households with variables  $(Y^{(i)}, D^{(i)}, Z^{(i)})_{i=1}^n$ .  $Y^{(i)} \in \mathbb{R}_+$  is net financial assets,  $D^{(i)} \in \{0, 1\}$  is a binary indicator for eligibility to enroll in a 401(k) savings plan, and  $Z^{(i)} \in \mathbb{Z} \subseteq \mathbb{R}^9$  are covariates measuring demographics and earnings, as described in Chernozhukov and Hansen (2004).

We investigate the hypothesis that the effect of the 401(k) pension scheme on wealth accumulation follows a “rich get richer” phenomenon—i.e., whether those that benefited most from the 401(k) scheme are those who would otherwise have most wealth in the first place. To answer this question, we estimate how quantiles of the treatment effect  $Y(1) - Y(0)$  varies across quantiles of (i) income levels  $I \in Z$ , and (ii) net financial assets under the no-treatment scenario  $Y(0)$ . Estimating (i) will help us to determine whether individuals who are richer in income are able to take better advantage of the 401(k) pension scheme distributions, through their ability to save. Estimating (ii) will help us to determine the extent to which those who benefit most are those who were better off regardless.

We work under the assumptions laid out in Section 4.1 on counterfactuals  $\{Y(d, z) : (d, z) \in \{0, 1\} \times \mathbb{R}^9\}$ . We estimate flow-based counterfactual cocycles on these counterfactuals using the same architectures and procedure as in Section 8.1. We use the estimated cocycle to impute counterfactuals  $\hat{Y}(1) := \hat{Y}(1, Z)$  and  $\hat{Y}(0) := \hat{Y}(0, Z)$ , and estimate the

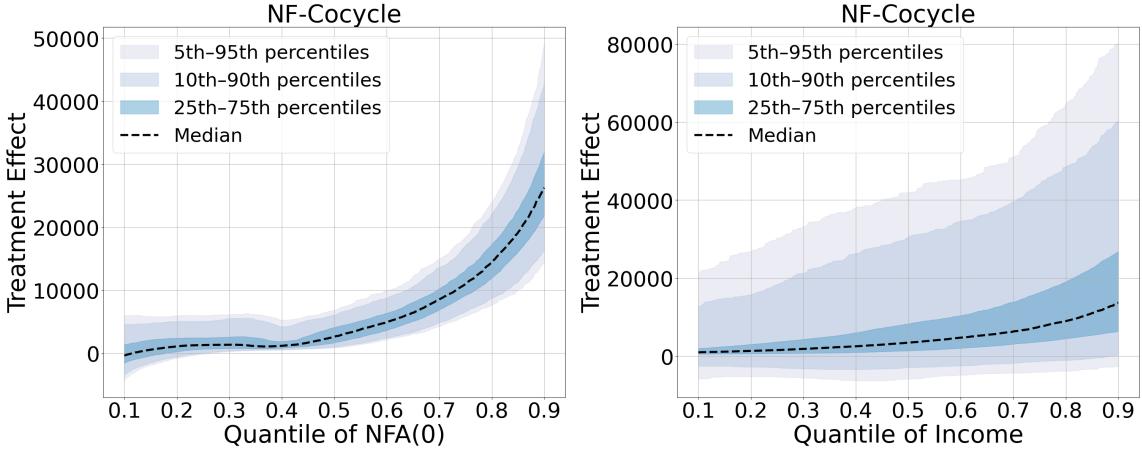


Figure 12: Estimated quantiles of the treatment effect  $Y(1) - Y(0)$  (i.e., change in net financial assets) given quantiles of (i)  $Y(0)$  (i.e., net financial assets under the no treatment scenario) and (ii) income, using normalizing flow-based cocycle.

quantile of the treatment effect  $\tau = Y(1) - Y(0)$  at level  $q$  given covariate  $V \in \{I, Y(0)\}$  by

$$\hat{Q}_{\tau|V=v}(q) := \inf \left\{ t \in \mathbb{R} : \sum_{i=1}^n \hat{w}_i(v) \mathbf{1}\{\hat{\tau}_i \leq t\} \geq q \right\},$$

i.e., the empirical quantile of  $\{\hat{\tau}_i\}_{i=1}^n$  reweighted via Nadaraya–Watson (Gaussian) kernel weights  $\hat{w}_i(v) \propto K_\lambda(v - V_i)$ . The bandwidth  $\lambda$  is selected by 5-fold cross-validation, minimizing the squared error of predicting  $\mathbf{1}\{\hat{\tau} \leq t_j\}$  from  $V$ , across different thresholds  $\{t_j\}_{j=1}^{100}$ .

The estimated conditional quantiles of the treatment effect are displayed in Fig. 12. Both distributions are consistent with the “rich get richer” phenomenon. In particular, the effects are generally largest for those who would have had greater net financial assets without access to a 401(k) pension scheme in the first place. While the majority of the 10th percentile of the distribution barely see a positive effect, the majority of the 90th percentile of the distribution see increases of  $\geq \$20,000$ . The story is somewhat similar when conditioning the effect on income levels (i.e., those with larger incomes see greater increases in net financial assets), albeit with larger treatment effect variance per quantile.

The fact that  $Y(1) - Y(0)$  typically increases with  $Y(0)$  implies that the effect of treatment on the quantile,  $ETQ(\tau) = Q_{Y(1)}(\tau) - Q_{Y(0)}(\tau)$  is also increasing over  $\tau$ , which has been reported in previous studies (Belloni et al., 2017; Chernozhukov and Hansen, 2004).

## 9 Conclusion

In this work, we introduced a general framework for modeling transport-based couplings over counterfactuals. Such couplings are essential for estimating measures of treatment risk and heterogeneity. To overcome the incoherence and identifiability problems of previous transport approaches, our key idea was to model transports between counterfactuals in a way that guarantees the necessary and sufficient algebraic properties to induce valid couplings. We called the resulting set of transports a *counterfactual cocycle*, given the connection to cocycles in dynamical systems (Arnold, 1998).

To achieve this, we characterized the structure of all counterfactual cocycles, and showed that each cocycle is equivalent to a class of injective SCMs. This equivalence enables parameterization via the same autoregressive flows used in flow-based SCMs, and identifiability under a known causal ordering. Crucially, however, cocycles are *noise invariant*: they depend only on the transports, not on the choice of latent noise distribution. This allows estimation to be centered directly on the transports, eliminating the need to model the noise law. Moreover, the flows required to represent a cocycle can be significantly simpler than those needed for the corresponding SCM, yielding models that are both well-specified under milder conditions and less prone to mis-specification.

To estimate cocycles efficiently, we proposed a new estimator based on minimizing the maximum mean discrepancy (MMD) between the true and predicted counterfactual marginals under the transports. In contrast to maximum likelihood approaches used in traditional flow-based SCMs, the estimator is *noise-robust*: its consistency does not rely on the properties of the underlying noise distribution. These advantages translate into strong empirical performance: across synthetic benchmarks and a 401(k) eligibility study, cocycle models outperformed both OT-based methods and flow-based SCMs.

One interesting direction for future research is how to construct identifiable classes of counterfactual cocycles without knowledge of the causal ordering. A promising avenue could be to combine the algebraic structure of cocycles with the optimal transport criteria, yielding valid transports that satisfy the causal principle of counterfactual similarity (Lewis, 1973). It also remains to be seen how one could extend the framework to settings with more severe forms of unobserved confounding, where counterfactual marginals are not conditionals.

**Acknowledgements** The authors are grateful to Peter Orbanz for introducing them to cocycles, and for helpful conversations about them. HD is supported by the Gatsby Charitable Foundation. BBR acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC): RGPIN2020-04995, RGPAS-2020-00095, DGECR2020-00343.

## Appendix A. Proofs

### A.1 Proofs for Section 3

**Proof** [Theorem 2] For the first direction, suppose  $\{T_{x,x'}\}_{x,x' \in \mathbb{X}}$  satisfies (ID), (PI) and (DA) for a given  $\mathbb{P}_{Y|X}$ . Now, fix the (standard Borel) probability space  $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}), \mathbb{P}_{Y|X=x_0})$ , where  $\mathcal{B}(\mathbb{Y})$  is the Borel  $\sigma$ -algebra on  $\mathbb{Y}$ . We can therefore define the random variable  $Y(x_0) : \mathbb{Y} \rightarrow \mathbb{Y}$  as the identity map, which is Borel measurable. By construction,  $Y(x_0) \sim \mathbb{P}_{Y|X=x_0}$ . Now, define the random variables

$$\tilde{Y}(x) := T_{x,x_0}(Y(x_0)), \quad \forall x \in \mathbb{X}$$

Note by measurability of each  $T_{x,x_0}$ , they are well-defined. By (DA) we have  $\tilde{Y}(x) \sim \mathbb{P}_{Y|X=x}$  for each  $x \in \mathbb{X}$ , which is the marginal distribution requirement of admissibility. To show (CC) holds almost surely, we note by (ID) and (PI) that  $T_{x_0,x} \circ T_{x,x_0} = T_{x_0,x_0} = \text{id}$ ,  $\mathbb{P}_{Y|X=x_0}$ -a.s. This means that

$$T_{x_0,x}(\tilde{Y}(x)) := T_{x_0,x} \circ T_{x,x_0}(Y(x_0)) =_{\text{a.s.}} Y(x_0), \quad \forall x \in \mathbb{X}.$$

Applying  $T_{x',x_0}$  to both sides and noting that  $x$  and  $x'$  are arbitrary gives

$$T_{x',x_0} \circ T_{x_0,x}(\tilde{Y}(x)) =_{\text{a.s.}} T_{x',x_0}(Y(x_0)) := \tilde{Y}(x') , \quad \forall x, x' \in \mathbb{X}.$$

By (PI), the LHS  $= T_{x',x}(\tilde{Y}(x))$  a.s. Thus, the transports are admissible.

For the reverse direction, let  $\{T_{x,x'}\}_{x,x' \in \mathbb{X}}$  be admissible w.r.t.  $\mathbb{P}_{Y|X}$ . In this case, there exist random variables  $\{\tilde{Y}(x)\}_{x \in \mathbb{X}}$  with  $\tilde{Y}(x) \sim \mathbb{P}_{Y|X=x}$  and  $\tilde{Y}(x) =_{\text{a.s.}} T_{x,x'}(\tilde{Y}(x'))$  for all  $x, x' \in \mathbb{X}$ . Below we show that this implies (DA), (ID), and (PI) hold:

(DA) For any  $x, x' \in \mathbb{X}$ , the fact that  $\tilde{Y}(x) =_{\text{a.s.}} T_{x,x'}(\tilde{Y}(x'))$  implies  $\tilde{Y}(x) =_{\text{d}} T_{x,x'}(\tilde{Y}(x'))$ . Since  $Y(x) \sim \mathbb{P}_{Y|X=x}$ , this means  $(T_{x,x'})_{\#} \mathbb{P}_{Y|X=x'} = \mathbb{P}_{Y|X=x}$ .

(NI) For any  $x \in \mathbb{X}$  we have  $\tilde{Y}(x) =_{\text{a.s.}} T_{x,x}(\tilde{Y}(x))$ , so  $T_{x,x} = \text{id}$   $\mathbb{P}_{Y|X=x}$ -a.s.

(CPI) For any  $x, x', x'' \in \mathbb{X}$ , we have

$$T_{x'',x'} \circ T_{x',x}(\tilde{Y}(x)) =_{\text{a.s.}} T_{x'',x'}(\tilde{Y}(x')) =_{\text{a.s.}} \tilde{Y}(x'') =_{\text{a.s.}} T_{x'',x}(\tilde{Y}(x))$$

So,  $T_{x'',x'} \circ T_{x',x} = T_{x'',x}$   $\mathbb{P}_{Y|X=x}$ -a.s. ■

**Proof** [Theorem 3] For each  $x \in \mathbb{X}$  set  $\tilde{f}_x := T_{x,x_0} : \mathbb{Y} \rightarrow \mathbb{Y}$ , where  $x_0 \in \mathbb{X}$  is arbitrary. Now, let  $\mathbb{Y}_{x_0}$  be the set of full  $\mathbb{P}_{Y|X=x_0}$ -measure on which  $T_{x_0,x_0} = \text{id}$ . Since  $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}))$  is a Polish space and  $\mathbb{P}_{Y|X=x_0}$  a probability measure on this space, it is a standard fact that any measurable subset  $A \subseteq \mathbb{Y}$  with  $\mathbb{P}_{Y|X=x_0}(A) = 1$  contains a Borel subset  $B \subseteq A$  also with  $\mathbb{P}_{Y|X=x_0}(B) = 1$ . Thus, we choose  $\mathbb{Y}_0$  as the Borel subset of  $\mathbb{Y}_{x_0}$  of full  $\mathbb{P}_{Y|X=x_0}$ -measure and note that  $\mathbb{Y}_0 \subseteq \mathbb{Y}$ , as required. Let  $f_x := \tilde{f}_x|_{\mathbb{Y}_0} : \mathbb{Y}_0 \rightarrow \mathbb{Y}$  be the restriction of  $\tilde{f}_x$  to  $\mathbb{Y}_0$ . It is Borel measurable since  $\mathbb{Y}_0$  is a Borel subset of  $\mathbb{Y}$ . Furthermore, by (ID) we have  $T_{x_0,x} \circ f_x = \text{id}_{\mathbb{Y}_0}$ . Therefore,  $f_x : \mathbb{Y}_0 \rightarrow \mathbb{Y}$  is a Borel *injection*. The Lusin-Souslin theorem then implies that the set  $f_x(\mathbb{Y}_0)$  is Borel and that the map  $T_{x_0,x}|_{f_x(\mathbb{Y}_0)} : f_x(\mathbb{Y}_0) \rightarrow \mathbb{Y}_0$  is also Borel measurable (Kechris, 2012, Corr. 15.2).

Now, we will extend  $T_{x_0,x}|_{f_x(\mathbb{Y}_0)}$  appropriately to define a measurable left inverse  $f_x^+ : \mathbb{Y} \rightarrow \mathbb{Y}_0$  of  $f_x$ . To that end, choose  $y_0 \in \mathbb{Y}_0$  and define a total map  $f_x^+ : \mathbb{Y} \rightarrow \mathbb{Y}_0$  by

$$f_x^+(y) := \begin{cases} T_{x_0,x}(y), & y \in f_x(\mathbb{Y}_0), \\ y_0, & y \notin f_x(\mathbb{Y}_0). \end{cases}$$

As  $f_x(\mathbb{Y}_0)$  is a Borel set,  $f_x^+$  is Borel measurable (with codomain  $\mathbb{Y}_0$  equipped with the trace  $\sigma$ -algebra  $\mathcal{B}(\mathbb{Y})|_{\mathbb{Y}_0} := \{\mathbb{Y}_0 \cap B : B \in \mathcal{B}(\mathbb{Y})\}$ ). Moreover  $f_x^+ \circ f_x = \text{id}_{\mathbb{Y}_0}$  by (ID). To show the cocycle identity, note for any  $x' \in \mathbb{X}$  and  $y \in f_x(\mathbb{Y}_0)$  we have  $f_x^+(y) = T_{x_0,x}(y)$ . Since  $\mathbb{P}_{Y|X=x_0}(\mathbb{Y}_0) = 1$ , (DA) implies  $\mathbb{P}_{Y|X=x'}(f_{x'}(\mathbb{Y}_0)) = 1$  and so we can write

$$f_x \circ f_{x'}^+ = T_{x,x_0} \circ T_{x_0,x'} = T_{x,x'} \quad \mathbb{P}_{Y|X=x'}\text{-a.s.}$$

where the last equality holds by (PI). Since  $x$  and  $x'$  are arbitrary this proves the result. ■

**Proof** [Theorem 5] We prove that conditions 1 and 2 of Theorem 4 hold. For 1 (existence) we construct the required cocycle using a coboundary map from  $\mathcal{F}_{\mathbb{G}}$ . To that end, fix  $x_0 \in \mathbb{X}$  and note by (ID) and (PI) that any  $\mathbb{P}_{Y|X}$ -adapted,  $\mathbb{G}$ -valued cocycle  $T : \mathbb{X}^2 \times \mathbb{Y} \rightarrow \mathbb{Y}$  satisfies

$$T_{x,x_0} \circ T_{x_0,x} = \text{id}, \quad \mathbb{P}_{Y|X=x}\text{-a.s.} \quad (32)$$

Since  $T_{x,x_0} \in \mathbb{G}$  it has an exact inverse  $T_{x,x_0}^{-1} : \mathbb{Y} \rightarrow \mathbb{Y}$ . Applying this inverse to both sides of (32) gives  $T_{x,x_0}^{-1} = T_{x_0,x}$   $\mathbb{P}_{Y|X=x}$ -a.s. Thus for every  $x, x' \in \mathbb{X}$  we can write

$$T_{x,x'} = f_x \circ f_{x'}^{-1} \quad \mathbb{P}_{Y|X=x}\text{-a.s.}, \quad \text{where } f_x = T_{x,x_0}. \quad (33)$$

Thus, 1 holds for  $f^\star \in \mathcal{F}_{\mathbb{G}}$  defined by  $f^\star(x, y) = T_{x,x_0}(y)$ . Now, for 2 (uniqueness), we first show that  $\tilde{T}$  is another  $\mathbb{P}_{Y|X}$ -adapted and  $\mathbb{G}$ -valued cocycle if and only if

$$\tilde{T}_{x,x'} = f_x \circ b_x^{-1} \circ b_{x'} \circ f_{x'}^{-1}, \quad \mathbb{P}_{Y|X=x'}\text{-a.s.}, \quad \forall x, x' \in \mathbb{X}, \quad (34)$$

for some  $\{b_x\}_{x \in \mathbb{X}} \subset \text{Aut}(\mathbb{P}_{Y|X=x_0})|_{\mathbb{G}}$ . For the first direction, we fix  $\tilde{T}$  and note that since (DA) holds, for any  $x \in \mathbb{X}$  we have

$$(\tilde{T}_{x,x_0})_\# \mathbb{P}_{Y|X=x_0} = \mathbb{P}_{Y|X=x} \implies (\tilde{T}_{x,x_0}^{-1})_\# \mathbb{P}_{Y|X=x} = \mathbb{P}_{Y|X=x_0}$$

and the same holds for  $T$ . Thus, setting  $\tilde{f}_x := \tilde{T}_{x,x_0}$  and recalling that  $f_x := T_{x,x_0}$ , we have

$$\begin{aligned} \mathbb{P}_{Y|X=x_0} &= (\tilde{f}_x^{-1})_\# \mathbb{P}_{Y|X=x} = (f_x^{-1})_\# \mathbb{P}_{Y|X=x}, \\ \implies \mathbb{P}_{Y|X=x_0} &= (\tilde{f}_x^{-1} \circ f_x)_\# \mathbb{P}_{Y|X=x_0} = (f_x^{-1} \circ \tilde{f}_x)_\# \mathbb{P}_{Y|X=x_0}. \end{aligned}$$

So,  $b_x := \tilde{f}_x^{-1} \circ f_x \in \text{Aut}(\mathbb{P}_{Y|X=x_0})|_{\mathbb{G}}$  and likewise for  $b_x^{-1} = f_x^{-1} \circ \tilde{f}_x$ . Now, note that

$$\tilde{f}_x \circ \tilde{f}_{x'}^{-1} = f_x \circ b_x^{-1} \circ b_{x'} \circ f_{x'}^{-1}, \quad (35)$$

for any  $x' \in \mathbb{X}$ . Moreover, following the same steps used to prove 1, we know  $\tilde{T}_{x,x'} = \tilde{f}_x \circ \tilde{f}_{x'}^{-1}$   $\mathbb{P}_{Y|X=x'}$ -a.s. Combining this with (35) and noting  $x, x' \in \mathbb{X}$  were arbitrary proves (34). Conversely, note that for any  $\{b_x\}_{x \in \mathbb{X}} \subset \text{Aut}(\mathbb{P}_{Y|X=x_0})|_{\mathbb{G}}$ , the construction  $\hat{T}_{x,x'} := f_x \circ b_x^{-1} \circ b_{x'} \circ f_{x'}^{-1}$  satisfies (34) and defines a valid  $\mathbb{P}_{Y|X}$ -adapted,  $\mathbb{G}$ -valued cocycle.

Thus, the set of  $\mathbb{G}$ -valued,  $\mathbb{P}_{Y|X}$ -adapted cocycles take the form (34) for any  $\{b_x\}_{x \in \mathbb{X}} \in \text{Aut}(\mathbb{P}_{Y|X=x_0})|_{\mathbb{G}}$ . The cocycle is therefore (a.s.) unique if and only if

$$f_x \circ f_{x'}^{-1} = f_x \circ b_x^{-1} \circ b_{x'} \circ f_{x'}^{-1} \quad \mathbb{P}_{Y|X=x'}\text{-a.s.} \quad (36)$$

$$\implies b_x^{-1} \circ b_{x'} = \text{id} \quad \mathbb{P}_{Y|X=x_0}\text{-a.s.} \quad (37)$$

for  $(\mathbb{P}_X \otimes \mathbb{P}_X)$ -every  $x, x' \in \mathbb{X}$  and for all possible collections  $\{b_x\}_{x \in \mathbb{X}} \in \text{Aut}(\mathbb{P}_{Y|X=x_0})|_{\mathbb{G}}$ . The latter holds trivially if  $\text{Aut}(\mathbb{P}_{Y|X=x_0})|_{\mathbb{G}} \subseteq [\text{id}]_{\mathbb{P}_{Y|X=x_0}}$ . To see that it holds only if  $\text{Aut}(\mathbb{P}_{Y|X=x_0})|_{\mathbb{G}} \subseteq [\text{id}]_{\mathbb{P}_{Y|X=x_0}}$ , suppose by contradiction that  $\text{Aut}(\mathbb{P}_{Y|X=x_0})|_{\mathbb{G}} \not\subseteq [\text{id}]_{\mathbb{P}_{Y|X=x_0}}$ . Take any Borel set  $A \in \mathcal{B}(\mathbb{X})$  such that  $\mathbb{P}_X(A) = 1/2$  and  $g \in \text{Aut}(\mathbb{P}_{Y|X=x_0})|_{\mathbb{G}} \cap [\text{id}]_{\mathbb{P}_{Y|X=x_0}}^c$ . Using these elements, we can set

$$\hat{b}_x := \begin{cases} g : x \in A, \\ \text{id} : x \in A^c. \end{cases} \quad \text{and} \quad \hat{T}_{x,x'} := f_x \circ \hat{b}_x^{-1} \circ \hat{b}_{x'} \circ f_{x'}^{-1}, \quad \forall x, x' \in \mathbb{X}$$

Then,  $\hat{T} : (x, x', y') \mapsto \hat{T}_{x,x'}(y')$  is a valid  $\mathbb{P}_{Y|X}$ -adapted,  $\mathbb{G}$ -valued cocycle, but (37) does not hold w.r.t.  $(\hat{b}_x, \hat{b}_{x'})$  for  $(\mathbb{P}_X \otimes \mathbb{P}_X)$ -every  $x, x' \in \mathbb{X}$ . Thus, uniqueness of the cocycle holds if and only if  $\text{Aut}(\mathbb{P}_{Y|X=x_0})|_{\mathbb{G}} \subseteq [\text{id}]_{\mathbb{P}_{Y|X=x_0}}$ . This completes the proof.  $\blacksquare$

**Proof** [Theorem 6] By Theorem 5, it suffices to prove that  $\text{Aut}(\mathbb{P}_{Y|X=x_0})|_{\mathbb{G}_{\text{TMI}}} \subseteq [\text{id}]_{\mathbb{P}_{Y|X=x_0}}$  for arbitrary  $x_0 \in \mathbb{X}$ . To that end, fix  $x_0 \in \mathbb{X}$  and for ease of notation put  $P := \mathbb{P}_{Y|X=x_0}$ . To prove the result, we first prove the following lemma.

**Lemma 14.** *Let  $\nu$  be a Borel probability measure on  $\mathbb{R}$ . If  $S : \mathbb{R} \rightarrow \mathbb{R}$  is measurable, non-decreasing and satisfies  $S_\# \nu = \nu$ , then  $S(y) = y$ ,  $\nu$ -almost surely.*

**Proof** [Theorem 14] Write  $F$  for the distribution function of  $\nu$ . Since  $S$  is monotone increasing and preserves  $\nu$ , we know by the Portmanteau Lemma (e.g., Van der Vaart, 2000, Lem. 2.2) that  $S_\# \nu(A) = \nu(A)$  for every closed  $A \in \mathcal{B}(\mathbb{R})$ . Since the family of sets  $\{(-\infty, t] : t \in \mathbb{R}\}$  are closed, we have the following equalities for every  $y \in \mathbb{R}$ :

$$F(S(y)) = \mathbb{P}(Y \leq S(y)) = \mathbb{P}(S(Y) \leq S(y)) = \mathbb{P}(Y \leq y) = F(y)$$

Where the second inequality follows from the Portmanteau Lemma under measure preserving  $S$  and the third inequality follows since  $S$  is monotone increasing. Now, let  $Q(\alpha) = \inf\{t \in \mathbb{R} : F(t) \geq \alpha\}$  be the generalized quantile function. It is a standard fact that  $Q \circ F = \text{id}$   $\nu$ -a.s. Since  $S$  preserves the null-sets of  $\nu$  by definition, we have  $Q \circ F \circ S = S$   $\nu$ -a.s. Since  $Q \circ F \circ S = Q \circ F = \text{id}$   $\nu$ -a.s., this immediately implies the result.  $\blacksquare$

Now we use the Lemma to prove the main result by induction. Let  $T = (T_1, \dots, T_p) \in \mathbb{G}_{\text{TMI}}$  be measurable and satisfy  $T_\# P = P$ . We show  $T_k(y) = y_k$   $P$ -a.s. for  $k = 1, \dots, p$ . For the base case  $k = 1$ , the first marginal  $P^{(1)}$  is a probability measure on  $\mathbb{R}$ . Because  $T_1$  is non-decreasing and  $(T_1)_\# P^{(1)} = P^{(1)}$ , Lemma 14 forces  $T_1(y_1) = y_1$   $P^{(1)}$ -a.s.

For the inductive part, assume  $T_j(y) = y_j$   $P$ -a.s. for each  $j \leq k$ , where  $k \in [1, p - 1]$  is arbitrary. Let  $P^{(k)} := P \circ \pi_{1:k}^{-1}$  be the projection of  $P$  onto the first  $k$  coordinates (i.e.,  $\pi_{1:k}(y) = (y_1, \dots, y_k)$ ). Fix  $y_{\leq k}$  and let  $P_{y_{\leq k}}$  be the regular conditional distribution of  $Y_{k+1}$  given  $Y_{\leq k} = y_{\leq k}$ . As the first  $k$  coordinates of  $Y$  are already the identity,  $T_{k+1}$  satisfies

$$(T_{k+1})_\# P_{y_{\leq k}} = P_{y_{\leq k}}, \quad x \mapsto T_{k+1}(y_{\leq k}, x) \text{ non-decreasing.}$$

Applying Lemma 14 to  $\nu = P_{y_{\leq k}}$  yields  $T_{k+1}(y) = y_{k+1}$   $P$ -a.s. By induction,  $T = \text{id}$   $P$ -a.s. Therefore, any  $T \in \mathbb{G}_{\text{TMI}}$  with  $T_\# P = P$  belongs to the  $P$ -a.s. equivalence class of the identity, so  $\text{Aut}_{\text{TMI}}(P) \subseteq [\text{id}]_P$ . Because  $x_0 \in \mathbb{X}$  was arbitrary, the statement holds for every conditional law  $\mathbb{P}_{Y|X=x_0}$ , completing the proof.  $\blacksquare$

**Proof** [Theorem 7] Assume the counterfactuals satisfy Assumption 1 and (CC) with cocycle  $T$ . Then by Theorem 3 we have  $Y(x) =_{\text{a.s.}} f_x \circ f_{x'}^+(Y(x'))$  with  $f_x := T_{x,x_0}|_{\mathbb{Y}_0}$  for any  $x, x', x_0 \in \mathbb{X}$ , and  $f_x^+$  the (measurable) left inverse of  $f_x$ . Now, fix  $x_0 \in \mathbb{X}$  and set  $\xi := f_{x_0}^+(Y(x_0)) \in \mathbb{Y}_0$ , so that  $Y(x) =_{\text{a.s.}} f_x(\xi)$  for every  $x \in \mathbb{X}$ . By Assumption 1.2 (exchangeability), we have  $X \perp\!\!\!\perp \xi$ . We next define  $f : (x, y) \mapsto f_x(y)$  and note since

$T : (x, x', y') \mapsto T_{x,x'}(y')$  is Borel measurable so is  $f$ . By Assumption 1.1 (consistency) we have  $Y =_{\text{a.s.}} f(X, \xi)$ . Lastly, by Theorem 3  $f_x := f(x, \bullet)$  is injective for every  $x \in \mathbb{X}$ .

Conversely, if  $Y =_{\text{a.s.}} f(X, \xi)$  with  $\xi \in \mathbb{Y}_0$ ,  $\xi \perp\!\!\!\perp X$  and  $f(x, \bullet) : \mathbb{Y}_0 \rightarrow \mathbb{Y}$  injective for every  $x \in \mathbb{X}$ , then by definition of counterfactuals in an SCM we have  $Y(x) =_{\text{a.s.}} f_x(\xi)$ , with  $f_x := f(x, \bullet)$  (Pearl, 2009a). Since  $f_x$  is injective on  $\mathbb{Y}_0$ , it admits a left inverse  $f_x^+$ . We construct this left inverse so that the map  $f^+ : (x, y) \mapsto f_x^+(y)$  is Borel measurable.

To that end, define the embedding  $F : \mathbb{X} \times \mathbb{Y}_0 \rightarrow \mathbb{X} \times \mathbb{Y}$ ,  $F(x, e) = (x, f(x, e))$ . Since  $F$  is Borel and injective, by the Lusin–Souslin theorem (Kechris, 2012, Corr. 15.2), the image  $G := F(\mathbb{X} \times \mathbb{Y}_0)$  is Borel in  $\mathbb{X} \times \mathbb{Y}$  and the inverse  $F^{-1} : G \rightarrow \mathbb{X} \times \mathbb{Y}_0$  is also Borel. Now, let  $\pi_{\mathbb{Y}_0}$  be the projection onto  $\mathbb{Y}_0$  and fix  $e^* \in \mathbb{Y}_0$ . We define  $f^+ : \mathbb{X} \times \mathbb{Y} \rightarrow \mathbb{Y}_0$ ,

$$f^+(x, y) := \begin{cases} \pi_{\mathbb{Y}_0}(F^{-1}(x, y)), & (x, y) \in G, \\ e^*, & (x, y) \notin G. \end{cases}$$

Note that  $f^+$  is Borel and  $\forall (x, e) \in \mathbb{X} \times \mathbb{Y}_0$  we have  $(x, f(x, e)) \in G$ , so  $f^+(x, f(x, e)) = e$ . Since  $\xi \in \mathbb{Y}_0$ , we have  $f_x^+(Y(x)) =_{\text{a.s.}} f_x^+ \circ f_x(\xi) =_{\text{a.s.}} \xi$ . Composing with  $f_{x'}$  for any  $x' \in \mathbb{X}$ ,

$$Y(x') =_{\text{a.s.}} f_{x'} \circ f_x^+(Y(x)) =: T_{x',x}(Y(x)) ,$$

which verifies (CC). Now, since  $X \perp\!\!\!\perp \xi$ , and independence is preserved under transformation, it holds that  $Y(x) =_{\text{a.s.}} f(x, \xi) \perp\!\!\!\perp X$ . Moreover,  $Y(X) =_{\text{a.s.}} f(X, \xi)$  which verifies consistency. The measurability of  $T(x', x, y) := f(x', f^+(x, y))$  follows from the fact that a composition of Borel maps is Borel.  $\blacksquare$

## A.2 Proofs for Section 5

**Proof** [Theorem 8] First note that by the definition of  $T$ , we have that

$$\mathbb{P}_{Y|X=x}(A) = \int_{\mathbb{X}} \mathbb{P}_{Y|X=x'}(T_{x,x'}^{-1}\{A\}) \mu(dx') = \int_{\mathbb{X} \times \mathbb{Y}} \mathbf{1}\{T_{x,x'}(y') \in A\} \mathbb{P}_{Y|X=x'}(dy') \mu(dx') ,$$

for every probability measure  $\mu \in \mathcal{P}(\mathbb{X})$ , every measurable set  $A \in \mathcal{B}(\mathbb{Y})$  (the Borel  $\sigma$ -algebra on  $\mathbb{Y}$ ) and  $x \in \mathbb{X}$ . By the identity  $\mathbb{E}\|H\|^2 = \|\mathbb{E}H\|^2 + \text{Tr}[\text{Cov}[H]]$  for any random  $H \in \mathcal{H}_k$ , we can write the following for some  $c \in \mathbb{R}$ ,

$$\begin{aligned} \ell(T) &= \mathbb{E}\|\psi(Y) - \mathbb{E}[\psi(T_{X,X'}(Y'))|X]\|_{\mathcal{H}_k}^2 = \mathbb{E}\|\mathbb{E}[\psi(Y)|X] - \mathbb{E}[\psi(T_{X,X'}(Y'))|X]\|_{\mathcal{H}_k}^2 + c \\ &= \mathbb{E}_{x \sim \mathbb{P}_X} D \left( \mathbb{P}_{Y|X=x}(\bullet), \int_{\mathbb{X}} \mathbb{P}_{Y|X=x'}(T_{x,x'}^{-1}\{\bullet\}) \mathbb{P}_X(dx') \right)^2 + c = c , \end{aligned}$$

and  $D = \text{MMD}$  is a metric on  $\mathcal{P}(\mathbb{Y})$  since  $k$  is taken as characteristic, we have  $T \in M := \text{arginf}_{\tilde{T} \in \mathcal{T}_{\mathbb{G}}} \ell(\tilde{T})$  and so  $M$  is non-empty. Now, take arbitrary  $T^* \in M$ . Since its coboundary map  $f^*$  lies in  $\mathcal{F}_{\mathbb{G}}$ , we have  $T_{x,x}^* = f_x^* \circ f_x^{*-1}$  (i.e., (ID)) and  $T_{x,x'}^* = f_x^* \circ f_{x''}^{*-1} \circ f_{x''}^* \circ f_{x'}^{*-1} = f_x^* \circ f_{x'}^{*-1}$  (i.e., (PI)) for every  $x, x', x'' \in \mathbb{X}$ . Therefore, all that remains is to show  $T^*$  satisfies (DA) ( $\mathbb{P}_X \otimes \mathbb{P}_X$ )-almost everywhere. The fact that  $D$  is a metric on  $\mathcal{P}(\mathbb{Y})$  and  $\ell(T^*) = 0$  implies the following inequalities for arbitrary  $A \in \mathcal{B}(\mathbb{Y})$  and  $\mathbb{P}_X$ -almost all  $x \in \mathbb{X}$ :

$$\mathbb{P}_{Y|X=x}(A) = \int_{\mathbb{X} \times \mathbb{Y}} \mathbf{1}\{T_{x,x'}^*(y') \in A\} \mathbb{P}_{Y|X=x'}(dy') \mathbb{P}_X(dx') = \mathbb{E}[\mathbf{1}\{T_{x,x}^*(Y) \in A\}] .$$

Therefore,  $T_{x,X}^*(Y) := f_x^* \circ f_X^{*-1}(Y) =_d \mathbb{P}_{Y|X=x}$ , for  $\mathbb{P}_X$ -almost all  $x \in \mathbb{X}$ . Defining  $\xi^* := f_X^{*-1}(Y) \sim \mathbb{P}_\xi^*$ , this implies  $\mathbb{P}_{Y|X=x} = (f_x^*)_\# \mathbb{P}_\xi^*$  for  $\mathbb{P}_X$ -almost all  $x \in \mathbb{X}$ , which immediately implies the result.  $\blacksquare$

**Proof** [Theorem 9] Note that for any cocycle  $T : \mathbb{X}^2 \times \mathbb{Y} \rightarrow \mathbb{Y}$ ,  $\ell_n^V(T)$  and  $\ell_n^U(T)$  are  $V$ -statistics and  $U$ -statistics of order three respectively. It is a standard fact (e.g., Serfling (2009) Sec 5) that one may replace the kernel of such statistics by its symmetrized version under  $\mathbb{S}_3$ , the group of permutations on  $\{1, 2, 3\}$ . That is, we may write

$$\ell_n^V(T) = \frac{1}{n^3} \sum_{i,j,k} h_T(Z^{(i)}, Z^{(j)}, Z^{(k)}), \quad \ell_n^U(T) = \frac{1}{\binom{n}{3}} \sum_{i < j < k} h_T(Z^{(i)}, Z^{(j)}, Z^{(k)}),$$

where  $Z^{(i)} := (X^{(i)}, Y^{(i)})$  and  $h_T(Z^{(i)}, Z^{(j)}, Z^{(k)}) = \frac{1}{6} \sum_{\sigma \in \mathbb{S}_3} \tilde{h}_T(Z^{(\sigma(i))}, Z^{(\sigma(j))}, Z^{(\sigma(k))})$  is the symmetrized version of the original kernel of the statistic:

$$\tilde{h}_T(Z^{(i)}, Z^{(j)}, Z^{(k)}) = -2k(Y^{(i)}, T_{X^{(i)}, X^{(j)}}(Y^{(j)})) + k(T_{X^{(i)}, X^{(j)}}(Y^{(j)}), T_{X^{(i)}, X^{(k)}}(Y^{(k)})).$$

Now, note that by the boundedness of the kernel,  $\tilde{h}_T$  is uniformly bounded and therefore so is  $h_T$ . Therefore, under the assumption that  $Z^{(1)}, \dots, Z^{(n)}$   $\stackrel{\text{iid}}{\sim} \mathbb{P}_Z$ , by Hoeffding's inequality for bounded U-statistics (e.g., see Sec 5.6.2. Theorem A in Serfling (2009)) we have  $\ell_n^U(T) - \mathbb{E}h_T(Z, Z', Z'') = \mathcal{O}_P(n^{-\frac{1}{2}})$ . Since  $f$  is bounded, it is known that  $|\ell_n^U(T) - \ell_n^V(T)| = \mathcal{O}_p(n^{-\frac{1}{2}})$  (e.g., see Lemma 5.7.3. in Serfling (2009)), which immediately implies  $\ell_n^V(T) - \mathbb{E}h_T(Z, Z', Z'') = \mathcal{O}_P(n^{-\frac{1}{2}})$  also. Note here  $Z, Z', Z'' \stackrel{\text{iid}}{\sim} \mathbb{P}_Z$  are independent copies.

All that remains is to show that  $\mathbb{E}h_T(Z, Z', Z'') = \ell(T) + \beta$ , where  $\beta$  is constant with respect to  $T$ . Since  $\mathbb{E}h_T(Z, Z', Z'') = \mathbb{E}\tilde{h}_T(Z, Z', Z'')$ , it suffices to show this for the latter (unsymmetrized) function). In what follows, we define  $\mu(\mathbb{P}_{Y|X}(\bullet|X)) := \mathbb{E}[\psi(Y)|X]$ ,  $\mu(\mathbb{P}_{Y_T|X}(\bullet|X)) := \mathbb{E}[\psi(T_{X,X'}(Y'))|X]$ , and  $Y_T := T_{X,X'}(Y')$ .

$$\begin{aligned} \mathbb{E}\tilde{h}_T(Z, Z', Z'') &= -2\mathbb{E}k(Y, T_{X,X'}(Y')) + \mathbb{E}k(T_{X,X'}(Y'), T_{X,X'}(Y')) \\ &= -2\mathbb{E}\langle \mu(\mathbb{P}_{Y|X}(\bullet|X)), \mu(\mathbb{P}_{Y_T|X}(\bullet|X)) \rangle_{\mathcal{H}_k} + \mathbb{E}\langle \mu(\mathbb{P}_{Y_T|X}(\bullet|X)), \mu(\mathbb{P}_{Y_T|X}(\bullet|X)) \rangle_{\mathcal{H}_k} \\ &= \mathbb{E}(\|\mu(\mathbb{P}_{Y_T|X}(\bullet|X))\|_{\mathcal{H}_k}^2 - 2\langle \mu(\mathbb{P}_{Y|X}(\bullet|X)), \mu(\mathbb{P}_{Y_T|X}(\bullet|X)) \rangle_{\mathcal{H}_k} + \|\mu(\mathbb{P}_{Y|X}(\bullet|X))\|_{\mathcal{H}_k}^2) \\ &= \mathbb{E}\|\mu(\mathbb{P}_{Y|X}(\bullet|X)) - \mu(\mathbb{P}_{Y_T|X}(\bullet|X))\|_{\mathcal{H}_k}^2 + \beta \\ &= \ell(T) + \beta \end{aligned}$$

Where  $\beta = -\mathbb{E}\|\mu(\mathbb{P}_{Y|X}(\bullet|X))\|_{\mathcal{H}_k}^2$ . This completes the proof.  $\blacksquare$

**Proof** [Theorem 10] For notational convenience let  $Z := (X, Y)$ ,  $Z^{(1)}, \dots, Z^{(n)} \stackrel{\text{iid}}{\sim} \mathbb{P}_Z$  and  $\ell_n := \ell_n^U$ . We prove the result by extending known results for convergence in probability (e.g., Theorem 5.7 in Van der Vaart (2000)) to the case of  $U$ -statistics and a minimizing set  $M$  rather than a unique minimizer  $\theta_0$ . Now, since  $\ell_n^U(\theta)$  is an order-3  $U$ -statistic, it is a standard fact (Serfling, 2009, Sec. 5) that we can express it as

$$\ell(\theta) = \mathbb{E}[h_\theta(Z, Z', Z'')], \quad \ell_n(\theta) = \frac{1}{\binom{n}{3}} \sum_{i < j < k} h_\theta(Z^{(i)}, Z^{(j)}, Z^{(k)}),$$

where  $h_\theta(Z^{(i)}, Z^{(j)}, Z^{(k)}) = \frac{1}{6} \sum_{\sigma \in \mathbb{S}_3} \tilde{h}_\theta(Z^{(\sigma(i))}, Z^{(\sigma(j))}, Z^{(\sigma(k))})$  is the symmetrized version of the original kernel of the statistic:

$$\tilde{h}_\theta(Z^{(i)}, Z^{(j)}, Z^{(k)}) = -2k(Y^{(i)}, T_{\theta, X^{(i)}, X^{(j)}}(Y^{(j)})) + k(T_{\theta, X^{(i)}, X^{(j)}}(Y^{(j)}), T_{\theta, X^{(i)}, X^{(k)}}(Y^{(k)})).$$

Define  $M = \arg \min_{\theta \in \Theta} \ell(\theta)$ . We start by showing  $\ell$  is continuous and  $M$  is compact. By Assumption 4.4 the kernel  $k$  is continuous and bounded, so the composite  $h_\theta$  is uniformly bounded by a constant  $C < \infty$  and, by Assumption 4.2, continuous in  $\theta$ . Therefore, for any sequence  $(\theta_n)_{n \geq 1} \in \Theta$  such that  $\theta_n \rightarrow \theta$ , we have point-wise convergence  $h_{\theta_n}(z, z', z'') \rightarrow h_\theta(z, z', z'')$  and the uniform bound  $|h_{\theta_n}(z, z', z'')| \leq C$ . Dominated Convergence then gives

$$\ell(\theta_n) = \mathbb{E}[h_{\theta_n}(Z, Z', Z'')] \rightarrow \mathbb{E}[h_\theta(Z, Z', Z'')] = \ell(\theta).$$

Hence  $\ell$  is continuous on  $\Theta$  and, since  $\Theta$  is compact, by the Weierstrass extreme-value theorem  $\ell$  attains its minimum. Therefore,  $M = \ell^{-1}(\{\min_{\vartheta \in \Theta} \ell(\vartheta)\})$  is the inverse image of a closed set under a continuous map, and so  $M$  is closed in  $\Theta$  and compact by the Heine–Borel Theorem. Now, for strong consistency of  $\hat{\theta}_n$ , we will show the following properties:

$$\text{Well-separatedness: } \inf_{\theta \in \Theta: \inf_{\theta' \in M} \|\theta - \theta'\|_2 \geq \varepsilon} [\ell(\theta) - \min_{\vartheta \in \Theta} \ell(\vartheta)] = \delta_\varepsilon > 0 \quad (38)$$

$$\text{Uniform convergence: } \sup_{\theta \in \Theta} |\ell_n(\theta) - \ell(\theta)| \rightarrow 0 \quad \text{a.s.} \quad (39)$$

For well-separatedness, fix  $\varepsilon > 0$  and set  $A_\varepsilon = \{\theta \in \Theta : d_M(\theta) \geq \varepsilon\}$ , where  $d_M(\theta) := \inf_{\theta' \in M} \|\theta - \theta'\|_2$ . By the triangle inequality, the map  $d_M : \mathbb{R}^d \rightarrow [0, \infty)$  is 1-Lipschitz, hence continuous. Therefore  $A_\varepsilon = d_M^{-1}([\varepsilon, \infty))$  is closed in  $\mathbb{R}^d$  and is compact since  $\Theta$  is compact. Now, define

$$\ell^\star := \min_{\vartheta \in \Theta} \ell(\vartheta), \quad \ell^\varepsilon := \min_{\vartheta \in A_\varepsilon} \ell(\vartheta).$$

Continuity of  $\ell$  on compact  $A_\varepsilon$  guarantees  $\ell^\varepsilon$  exists and, as  $A_\varepsilon$  is disjoint from  $M$ ,  $\ell^\varepsilon > \ell^\star$ . Define  $\delta_\varepsilon := \ell^\varepsilon - \ell^\star > 0$ . For any  $\theta \in \Theta$  with  $d_M(\theta) \geq \varepsilon$  we have  $\ell(\theta) \geq \ell^\varepsilon$ , hence (38) holds.

For uniform convergence of  $\ell_n(\theta)$  to  $\ell(\theta)$ , we note that  $\mathcal{F} := \{h_\theta : \theta \in \Theta\}$  is uniformly bounded, and continuous in  $\theta$ , and  $\Theta$  is compact. Such classes are known to be Glivenko–Cantelli and so have a finite bracketing number  $N_{[]}(\epsilon, \mathcal{F}, L^2(\mathbb{P}_Z^3) < \infty)$  (e.g., [Van der Vaart, 2000](#), Example 19.8). Since  $\mathcal{F}$  is continuous it is also Borel-measurable. This along with the finite bracketing number, means that the uniform strong law of large numbers (39) holds for the U-statistic  $\ell_n^U(\theta)$  ([De la Pena and Giné, 2012](#), Corr. 5.2.5).

Now we are ready to prove the consistency result. Since uniform convergence holds, for every  $\varepsilon > 0$  there exists an almost-surely finite random index  $N_\varepsilon$  such that, for all  $n \geq N_\varepsilon$ ,

$$\sup_{\theta \in \Theta} |\ell_n(\theta) - \ell(\theta)| < \delta_\varepsilon/2. \quad (40)$$

We work on the full-probability event  $\{N_\varepsilon < \infty\}$  and fix any  $n \geq N_\varepsilon$ . Suppose, for contradiction, that  $d_M(\hat{\theta}_n) \geq \varepsilon$ , so  $\hat{\theta}_n \in A_\varepsilon$ . Combining (40) with (38) gives

$$\ell_n(\hat{\theta}_n) \geq \ell(\hat{\theta}_n) - \delta_\varepsilon/2 \geq \ell^\star + \delta_\varepsilon/2,$$

whereas for any minimizer  $\theta^* \in M$  we have  $\ell_n(\theta^*) \leq \ell^* + \delta_\varepsilon/2$  by (40), contradicting the minimality of  $\hat{\theta}_n$ . Hence  $d_M(\hat{\theta}_n) < \varepsilon$  for all  $n \geq N_\varepsilon$ . Since  $\varepsilon > 0$  is arbitrary, we conclude that  $d_M(\hat{\theta}_n) \rightarrow 0$  a.s.. This proves the consistency result for  $\hat{\theta}_n$ .

Lastly, we transfer the result to the cocycle  $T_{\hat{\theta}_n}$ . Since  $M$  is compact and  $\theta \mapsto \|\hat{\theta}_n - \theta\|$  is continuous, by the Measurable Maximum Theorem (Theorem 18.19 in [Aliprantis and Border \(2006\)](#)) one can define the measurable function  $\eta_n := \arg \min_{\theta \in M} \|\hat{\theta}_n - \theta\|$  and a.s. consistency of  $\hat{\theta}_n$  implies  $\|\hat{\theta}_n - \eta_n\| \rightarrow 0$  almost surely. By continuity of  $T$  in  $\theta$ , this implies

$$T_{\hat{\theta}_n, x, x'}(y) \rightarrow T_{\eta_n, x, x'}(y) \quad \text{a.s.} \quad (41)$$

Since  $\mathbb{N}$  is countable, by Assumption 4.3 we have  $\{T_{\eta_n, x, x'}(y) = T_{\theta_0, x, x'}(y), \forall n \in \mathbb{N}\}$  ( $P_X \otimes P_{X,Y}$ )-a.s., for any  $\theta_0 \in M$ . Combining this with (41) yields point-wise convergence  $T_{\hat{\theta}_n, x, x'}(y) \rightarrow T_{\theta_0, x, x'}(y)$  on a full-measure set, which completes the proof. ■

**Proof** [Theorem 12] For notational convenience, let  $Z := (X, Y)$ ,  $Z^{(1)}, \dots, Z^{(n)} \stackrel{\text{iid}}{\sim} \mathbb{P}_Z$  and note we can express the gradient of the U-statistic  $\ell_n(\theta) := \ell_n^U(\theta)$  as

$$\nabla_\theta \ell_n(\theta) = \frac{1}{n(n-1)(n-2)} \sum_{i \neq j \neq k} \nabla_\theta f_\theta(Z^{(i)}, Z^{(j)}, Z^{(k)}).$$

Now, define the symmetrized and centered function

$$H_\theta(z^{(1)}, z^{(2)}, z^{(3)}) := \frac{1}{6} \sum_{\pi \in \mathbb{S}_3} \nabla_\theta f_\theta(z_{\pi(1)}, z_{\pi(2)}, z_{\pi(3)}) - \nabla_\theta \ell(\theta).$$

By standard theory of U-statistics (e.g., [Serfling, 2009](#), Sec. 5), we can express  $\nabla_\theta \ell_n(\theta) - \nabla_\theta \ell(\theta)$  as a centered U-statistic with symmetric kernel  $H_\theta$ :

$$\nabla_\theta \ell_n(\theta) - \nabla_\theta \ell(\theta) = \frac{1}{\binom{n}{3}} \sum_{i < j < k} H_\theta(Z^{(i)}, Z^{(j)}, Z^{(k)}). \quad (42)$$

Now, note that if  $\nabla_\theta \ell(\theta) = \mathbb{E}[\nabla_\theta f_\theta(Z^{(1)}, Z^{(2)}, Z^{(3)})]$ , by the Hoeffding decomposition of  $H_\theta$  ([Serfling, 2009](#), Sec 5.1.5., Lemma A), there exist zero-mean symmetric projections

$$\begin{aligned} h_{1,\theta}(z) &= \mathbb{E}[H_\theta(z, Z^{(2)}, Z^{(3)})], \\ h_{2,\theta}(z^{(1)}, z^{(2)}) &= \mathbb{E}[H_\theta(z^{(1)}, z^{(2)}, Z^{(3)})] - h_{1,\theta}(z^{(1)}) - h_{1,\theta}(z^{(2)}), \\ h_{3,\theta}(z^{(1)}, z^{(2)}, z^{(3)}) &= H_\theta(z^{(1)}, z^{(2)}, z^{(3)}) - \sum_{i=1}^3 h_{1,\theta}(z^{(i)}) - \sum_{1 \leq i < j \leq 3} h_{2,\theta}(z^{(i)}, z^{(j)}), \end{aligned}$$

which enables us to write

$$H_\theta(z^{(1)}, z^{(2)}, z^{(3)}) = \sum_{i=1}^3 h_{1,\theta}(z^{(i)}) + \sum_{1 \leq i < j \leq 3} h_{2,\theta}(z^{(i)}, z^{(j)}) + h_{3,\theta}(z^{(1)}, z^{(2)}, z^{(3)}).$$

We will make use of this representation of  $H_\theta$  to split  $\nabla_\theta \ell_n(\theta) - \nabla_\theta \ell(\theta)$  into a first-order term and remainder term. To do this, we first show that we can swap the expectation

and the gradient so that  $\nabla_\theta \ell(\theta) = \mathbb{E}[\nabla_\theta f_\theta(Z^{(1)}, Z^{(2)}, Z^{(3)})]$  via the dominated convergence theorem (DCT) (i.e.,  $\nabla_\theta f_\theta < \kappa_\theta : \mathbb{E}[|\kappa_\theta(Z^{(1)}, Z^{(2)}, Z^{(3)})|] < \infty$ ). In particular, note that by Assumption 5.2 we have  $\partial k < B$ , so by the chain rule

$$\begin{aligned}\nabla_\theta f_\theta(Z^{(1)}, Z^{(2)}, Z^{(3)}) &= \\ &\quad - 2\nabla_\theta k(Y^{(1)}, T_{\theta, X^{(1)}, X^{(2)}}(Y^{(2)})) + \nabla_\theta k(T_{\theta, X^{(1)}, X^{(2)}}(Y^{(2)}), T_{\theta, X^{(1)}, X^{(3)}}(Y^{(3)})) \\ &\leq 2B\nabla_\theta T_{\theta, X^{(1)}, X^{(2)}}(Y^{(2)}) + B\nabla_\theta(T_{\theta, X^{(1)}, X^{(2)}}(Y^{(2)}) + T_{\theta, X^{(1)}, X^{(3)}}(Y^{(3)}))\end{aligned}$$

Similarly, by Assumption 5.1, we have  $\nabla_\theta T_{\theta, X^{(1)}, X^{(3)}}(Y^{(3)}) \leq L_T(X^{(1)}, X^{(3)}, X^{(3)})$ . Thus, we can set

$$\kappa_\theta(Z^{(1)}, Z^{(2)}, Z^{(3)}) = B(3L_T(X^{(1)}, X^{(2)}, Y^{(2)}) + L_T(X^{(1)}, X^{(3)}, Y^{(3)})),$$

and it is clear by the integrability of  $L_T$  that  $\mathbb{E}[|\kappa_\theta(Z^{(1)}, Z^{(2)}, Z^{(3)})|] < \infty$ . Therefore, the DCT applies and  $\nabla_\theta \ell(\theta) = \mathbb{E}[\nabla_\theta f_\theta(Z^{(1)}, Z^{(2)}, Z^{(3)})]$ . Replacing  $H_\theta$  in (42) with its representation in terms of the projections, we get the standard ANOVA decomposition

$$\begin{aligned}\nabla_\theta \ell_n(\theta) - \nabla_\theta \ell(\theta) &= \\ &\quad \frac{3}{n} \sum_{i=1}^n h_{1,\theta}(Z^{(i)}) + \frac{6}{n(n-1)} \sum_{i < j} h_{2,\theta}(Z^{(i)}, Z^{(j)}) + \frac{1}{\binom{n}{3}} \sum_{i < j < k} h_{3,\theta}(Z^{(i)}, Z^{(j)}, Z^{(k)}).\end{aligned}$$

Since  $\mathbb{E}[h_{1,\theta}(Z)] = 0$ , multiply by  $\sqrt{n}$  and set

$$g_{1,\theta}(z) = 3h_{1,\theta}(z), \quad R_{n,\theta} = \sqrt{n} \left\{ \frac{6}{n(n-1)} \sum_{i < j} h_{2,\theta}(Z^{(i)}, Z^{(j)}) + \frac{1}{\binom{n}{3}} \sum_{i < j < k} h_{3,\theta}(Z^{(i)}, Z^{(j)}, Z^{(k)}) \right\}.$$

This allows us to obtain the desired split into leading and remainder terms

$$\sqrt{n}(\nabla \ell_n(\theta) - \nabla \ell(\theta)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n g_{1,\theta}(Z^{(i)}) + R_{n,\theta}. \quad (\text{HD})$$

We now bound each term accordingly. For the remainder term, it is known that, by construction,  $h_{2,\theta}, h_{3,\theta}$  are  $\mathbb{P}$ -degenerate (Serfling, 2009, pp. 178). Additionally, since  $\Theta \subset \mathbb{R}^d$  is compact, the packing number is  $\mathcal{M}(\epsilon, \Theta, \|\cdot\|_2) \leq D/\epsilon^d$ , where  $D = \text{Diam}(\Theta)$ . This polynomial dependence on  $\epsilon$  lets us apply Sherman's maximal inequality for degenerate U-statistics (Sherman, 1994, Corr. 4):

$$\sup_{\theta \in U_\delta} \|R_{n,\theta}\| = O_p(n^{-3/2}) = o_p(1). \quad (\text{U})$$

For the first order term, by Assumptions 5(i)–(ii),  $\{g_{1,\theta} : \theta \in U_\delta\}$  is a parametric Lipschitz class and so is known to be  $\mathbb{P}$ -Donsker (Van der Vaart, 2000, Example 19.7). Therefore

$$\sup_{\theta \in U_\delta} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n (g_{1,\theta}(Z^{(i)}) - \mathbb{E}[g_{1,\theta}(Z)]) \right\| = O_p(1). \quad (\text{D})$$

Combining (HD), (U) and (D) yields  $\sup_{\theta \in U_\delta} \|\nabla \ell_n(\theta) - \nabla \ell(\theta)\| = O_p(n^{-1/2})$ . Since  $\hat{\theta}_n$  minimizes  $\ell_n$ ,  $\nabla \ell_n(\hat{\theta}_n) = 0$  implies  $\|\nabla \ell(\hat{\theta}_n)\| = \|\nabla \ell_n - \nabla \ell\|(\hat{\theta}_n) = O_p(n^{-1/2})$ . Since

Assumption 4 holds, so does Theorem 10 and so we know that  $\hat{\theta}_n \in U_\delta$  w.p. 1. Local strong convexity (Assumption 5(iii)) then yields  $\|\nabla \ell(\hat{\theta}_n)\| \geq c d_{\hat{\theta}_n}(M)$  for all  $n \geq N(\delta)$ , where  $N(\delta) \in \mathbb{N}$ . Hence, for all such  $n$ , we have

$$d_{\hat{\theta}_n}(M) \leq c^{-1} \|\nabla \ell(\hat{\theta}_n)\| = O_p(n^{-1/2}).$$

■

### A.3 Proofs for Section 6

**Proof** [Theorem 13] By definition of  $\mathbb{G}_{f(g)}$ , one of its generators is  $f_{x_0} \circ g = \text{id}_{\mathbb{Y}} \circ g = g$ . Hence  $g, g^{-1} \in \mathbb{G}_{f(g)}$ . Since for each  $x \in \mathbb{X}$ , we have  $f_x = (f_x \circ g) \circ g^{-1}$ , and  $f_x \circ g$  is a generator of  $\mathbb{G}_{f(g)}$ , it follows that  $f_x \in \mathbb{G}_{f(g)}$  for all  $x \in \mathbb{X}$ . Hence  $\mathbb{G}_f = \langle f_x : x \in \mathbb{X} \rangle \subseteq \mathbb{G}_{f(g)}$ . This means that any  $\mathbb{G}_f$ -valued coboundary map is also  $\mathbb{G}_{f(g)}$ -valued, and so  $\mathcal{F}_{\mathbb{G}_f} \subseteq \mathcal{F}_{\mathbb{G}_{f(g)}}$ . This proves (i). To prove (ii), take by hypothesis  $g \notin \mathbb{G}_f$ . In this case,  $\mathbb{G}_{f(g)} \not\subseteq \mathbb{G}_f$ . Define  $\tilde{f} : (x, y) \mapsto g(y)$  and note since  $g \in \mathbb{G}_{f(g)}$  it is a  $\mathbb{G}_{f(g)}$ -valued coboundary map. Since  $g \notin \mathbb{G}_f$  we have  $\tilde{f} \notin \mathcal{F}_{\mathbb{G}_f}$  and so  $\mathcal{F}_{\mathbb{G}_{f(g)}} \not\subseteq \mathcal{F}_{\mathbb{G}_f}$ . Combining with (i) completes the proof. ■

## Appendix B. Algorithms

**Data:** samples  $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ ,  
cocycle  $T_\theta$ , kernel  $k$ , batch size  
 $B$ , epochs  $E$ , step-size  $\eta$   
**Result:** Optimized cocycle  $T_\theta^*$

```

for epoch = 1, ..., E do
    s ← 0 ;           // samples counter
    while s < n do
        Sample without replacement a
        set  $\mathcal{B} \subset \{1, \dots, n\}$ ,  $|\mathcal{B}| = B$ ;
        forall  $i, j \in \mathcal{B}$  do
             $y_\theta^{(i,j)} \leftarrow T_{\theta, x^{(i)}, x^{(j)}}(y^{(j)})$ 
        end
         $\ell_b^V(\theta) \leftarrow$ 
         $\frac{1}{B^3} \sum_{i,j,k \in \mathcal{B}} k(y_\theta^{(i,j)}, y_\theta^{(i,k)}) -$ 
         $\frac{2}{B^2} \sum_{i,j \in \mathcal{B}} k(y^{(i)}, y_\theta^{(i,j)})$ ;
         $\theta \leftarrow \theta - \eta \nabla_\theta \ell_b^V(\theta)$ ;
        s ← s + B;
    end
end
return  $T_\theta^*$ 

```

**Algorithm 1:** Minibatch CMMD-V optimization.

**Data:** samples  $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ ,  
cocycles  $\{T_m\}_{m=1}^M$ , folds  $K$

**Result:** Selected cocycle  $T^*$

1. Partition  $\{1, \dots, n\}$  into  $K$  disjoint folds  $\{I_{\text{tr}}^{(k)}, I_{\text{val}}^{(k)}\}_{k=1}^K$ ;
2. **for**  $m = 1$  **to**  $M$  **do**
  - for**  $k = 1$  **to**  $K$  **do**
    - 2.1. Train  $T_m$  on  $\{i \in I_{\text{tr}}^{(k)}\}$  by minimizing CMMD;
    - 2.2. Evaluate  $\ell_{m,k} \leftarrow \text{CMMD}(T_m, \{i \in I_{\text{val}}^{(k)}\})$ ;
  - 2.3. Compute  $\bar{\ell}_m \leftarrow \frac{1}{K} \sum_{k=1}^K \ell_{m,k}$ ;
3.  $m^* \leftarrow \arg \min_m \bar{\ell}_m$ ;
4. Retrain  $T_{m^*}$  on all data by minimizing CMMD;
5. **return**  $T^*$

**Algorithm 2:** Cocycle model selection via K-fold cross-validation on CMMD.

## References

- C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer-Verlag, 3rd edition, 2006.
- P. Alquier and M. Gerber. Universal robust regression via maximum mean discrepancy. *Biometrika*, 111(1):71–92, 05 2023.
- S. Amiri, E. Nalisnick, A. Belloum, S. Klous, and L. Gommans. Practical synthesis of mixed-tailed data with normalizing flows. *Transactions on Machine Learning Research*, 2024.
- L. Arnold. *Random Dynamical Systems*. Springer, 1998.
- S. Athey and G. W. Imbens. Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497, 2006.
- R. Bai and M. Ghosh. On the beta prime prior for scale parameters in high-dimensional bayesian regression models. *Statistica Sinica*, 31(2):843–865, 2021.
- S. Balakrishnan, E. Kennedy, and L. Wasserman. Conservative inference for counterfactuals. *Journal of Causal Inference*, 13(1):20230071, 2025.
- A. Belloni, V. Chernozhukov, I. Fernandez-Val, and C. Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.
- A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.
- V. I. Bogachev, A. V. Kolesnikov, and K. V. Medvedev. Triangular transformations of measures. *Sbornik: Mathematics*, 196(3):309, 2005.
- G. Carlier, V. Chernozhukov, and A. Galichon. Vector quantile regression: An optimal transport approach. *The Annals of Statistics*, 44(3), 2016.
- A. Charpentier, E. Flachaire, and E. Galic. Optimal transport for counterfactual estimation: A method for causal inference. In *Optimal Transport Statistics for Economics and Related Topics*, pages 45–89. Springer, 2023.
- V. Chernozhukov and C. Hansen. The effects of 401(k) participation on the wealth distribution: an instrumental quantile regression analysis. *Review of Economics and Statistics*, 86(3):735–751, 2004.
- V. Chernozhukov and C. Hansen. An IV model of quantile treatment effects. *Econometrica*, 73(1):245–261, 2005.
- V. Chernozhukov, C. Hansen, N. Kallus, M. Spindler, and V. Syrgkanis. *Applied Causal Inference Powered by ML and AI*. 2025.
- R. Cornish, A. Caterini, G. Deligiannidis, and A. Doucet. Relaxing bijectivity constraints with continuously indexed normalising flows. In *International conference on machine learning*, pages 2133–2143. PMLR, 2020.

- H. Dance and B. Paige. Fast and scalable spike and slab variable selection in high-dimensional Gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pages 7976–8002, 2022.
- V. De la Pena and E. Giné. *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.
- L. De Lara, A. González-Sanz, N. Asher, L. Risser, and J.-M. Loubes. Transport-based counterfactual models. *Journal of Machine Learning Research*, 25(136):1–59, 2024.
- C. Deidda, S. Engelke, and C. De Michele. Asymmetric dependence in hydrological extremes. *Water Resources Research*, 59(12):e2023WR034512, 2023.
- L. Dinh, D. Krueger, and Y. Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.
- L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.
- C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios. Neural Spline Flows. *Advances in Neural Information Processing Systems*, 32, 2019.
- D. Garreau, W. Jitkrittum, and M. Kanagawa. Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*, 2017.
- T. Geffner, J. Antoran, A. Foster, W. Gong, C. Ma, E. Kiciman, A. Sharma, A. Lamb, M. Kukla, A. Hilmkil, et al. Deep end-to-end causal inference. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*, 2022.
- D. Greenfeld and U. Shalit. Robust learning with the Hilbert–Schmidt independence criterion. In *International Conference on Machine Learning*, pages 3759–3768, 2020.
- J. A. Gregory and R. Delbourgo. Piecewise rational quadratic interpolation to monotonic data. *IMA Journal of Numerical Analysis*, 2(2):123–130, 1982.
- A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.
- L. Henckel, E. Perković, and M. H. Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):579–599, 2022.
- B. Hosseini, A. W. Hsu, and A. Taghvaei. Conditional optimal transport on function spaces. *SIAM/ASA Journal on Uncertainty Quantification*, 13(1):304–338, 2025.
- C.-W. Huang, D. Krueger, A. Lacoste, and A. Courville. Neural autoregressive flows. In *International Conference on Machine Learning*, pages 2078–2087, 2018.
- K. Imai, L. Keele, and D. Tingley. A general approach to causal mediation analysis. *Psychological methods*, 15(4):309, 2010.

- I. Ishikawa, T. Teshima, K. Tojo, K. Oono, M. Ikeda, and M. Sugiyama. Universal approximation property of invertible neural networks. *Journal of Machine Learning Research*, 24(287):1–68, 2023.
- P. Jaini, I. Kobyzhev, Y. Yu, and M. Brubaker. Tails of lipschitz triangular flows. In *International Conference on Machine Learning*, pages 4673–4681, 2020.
- M. Jankowiak and G. Pleiss. Scalable Cross Validation Losses for Gaussian Process Models. *arXiv preprint arXiv:2105.11535*, 2021.
- A. Javaloy, P. Sánchez-Martín, and I. Valera. Causal normalizing flows: from theory to practice. *Advances in Neural Information Processing Systems*, 36, 2023.
- N. Kallus. Treatment effect risk: Bounds and inference. *Management Science*, 69(8):4579–4590, 2023.
- A. Kechris. *Classical descriptive set theory*, volume 156. Springer Science & Business Media, 2012.
- E. H. Kennedy. Semiparametric theory and empirical processes in causal inference. In H. He, P. Wu, and D.-G. D. Chen, editors, *Statistical Causal Inferences and Their Applications in Public Health Research*, pages 141–167. Springer International Publishing, 2016.
- I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 2020.
- I. Khemakhem, R. Monti, R. Leech, and A. Hyvarinen. Causal autoregressive flows. In *International Conference on Artificial Intelligence and Statistics*, 2021.
- D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in Neural Information Processing Systems*, 31, 2018.
- D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improved variational inference with inverse autoregressive flow. *Advances in Neural Information Processing Systems*, 29, 2016.
- I. Kobyzhev, S. J. Prince, and M. A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, 2020.
- M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.
- M. Laszkiewicz, J. Lederer, and A. Fischer. Marginal tail-adaptive normalizing flows. In *International Conference on Machine Learning*, pages 12020–12048. PMLR, 2022.
- D. Lewis. Causation. *The journal of philosophy*, 70(17):556–567, 1973.

- Z. Li, D. Meunier, M. Mollenhauer, and A. Gretton. Optimal rates for regularized conditional mean embedding learning. *Advances in Neural Information Processing Systems*, 35:4433–4445, 2022.
- F. Liang, M. Mahoney, and L. Hodgkinson. Fat-tailed variational inference with anisotropic tail adaptive flows. In *International Conference on Machine Learning*, pages 13257–13270. PMLR, 2022.
- W. W. Loh and J.-S. Kim. Evaluating sensitivity to classification uncertainty in latent subgroup effect analyses. *BMC Medical Research Methodology*, 22(1):247, 2022.
- A. F. Machado, A. Charpentier, and E. Gallic. Sequential conditional transport on probabilistic graphs for interpretable counterfactual fairness. *arXiv preprint arXiv:2408.03425*, 2024.
- D. Malinsky, I. Shpitser, and T. Richardson. A potential outcomes calculus for identifying conditional path-specific effects. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3080–3088. PMLR, 2019.
- N. Meinshausen and G. Ridgeway. Quantile regression forests. *Journal of machine learning research*, 7(6), 2006.
- A. Nasr-Esfahany, M. Alizadeh, and D. Shah. Counterfactual identifiability of bijective causal models. In *International Conference on Machine Learning*, 2023.
- G. Papamakarios, T. Pavlakou, and I. Murray. Masked autoregressive flow for density estimation. *Advances in Neural Information Processing Systems*, 30, 2017.
- G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.
- N. Pawłowski, D. Coelho de Castro, and B. Glocke. Deep structural causal models for tractable counterfactual inference. *Advances in Neural Information Processing Systems*, 33:857–869, 2020.
- J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009a.
- J. Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009b.
- J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.
- D. Plečko and N. Meinshausen. Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, 21(242):1–44, 2020.
- T. S. Richardson and J. M. Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Technical Report 128, Center for the Statistics and the Social Sciences, University of Washington, 2013.

- R. T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of banking & finance*, 26(7):1443–1471, 2002.
- D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.
- P. Sanchez-Martin, M. Rateike, and I. Valera. Vaca: Design of variational graph autoencoders for interventional and counterfactual queries. *arXiv preprint arXiv:2110.14690*, 2021.
- F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Springer International Publishing, 2015.
- R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 2009.
- C. Shen, J. Jeong, X. Li, P.-S. Chen, and A. Buxton. Treatment benefit and treatment harm rate to characterize heterogeneity in treatment effect. *Biometrics*, 69(3):724–731, 2013.
- R. P. Sherman. Maximal inequalities for degenerate  $u$ -processes with applications to optimization estimators. *The Annals of Statistics*, 22(1):439–459, 1994.
- I. Shpitser and E. Tchetgen Tchetgen. Causal inference with a graphical hierarchy of interventions. *Annals of Statistics*, 44(6):2433, 2016.
- H. A. Simon. Causal ordering and identifiability. *Studies in Econometric Methods*, pages 49–74, 1953.
- P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.
- B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(70):2389–2410, 2011.
- S. Syrota, Y. Zainchikovskyy, J. Xi, B. Bloem-Reddy, and S. Hauberg. Identifying metric structures of deep latent variable models. In *Forty-second International Conference on Machine Learning*, 2025.
- U. Tanielian, T. Issenhuth, E. Dohmatob, and J. Mary. Learning disconnected manifolds: a no GAN’s land. In *International Conference on Machine Learning*, pages 9418–9427. PMLR, 2020.
- W. Torous, F. Gunsilius, and P. Rigollet. An optimal transport approach to estimating causal effects via nonlinear difference-in-differences. *Journal of Causal Inference*, 12(1):20230004, 2024.
- A. W. Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

- S. Vansteelandt and M. Joffe. Structural nested models and g-estimation: The partially realized promise. *Statistical Science*, 29(4), 2014.
- V. S. Varadarajan. *Geometry of Quantum Theory*. Springer, 2nd edition, 1968.
- P. Verhoeven and M. McAleer. Fat tails and asymmetry in financial volatility models. *Mathematics and Computers in Simulation*, 64(3-4):351–361, 2004.
- C. Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.
- J. Xi and B. Bloem-Reddy. Indeterminacy in generative models: Characterization and strong identifiability. In *International Conference on Artificial Intelligence and Statistics*, 2023.
- Y. Yin, L. Liu, and Z. Geng. Assessing the treatment effect heterogeneity with a latent variable. *Statistica Sinica*, pages 115–135, 2018.
- S. L. Young, M. Taylor, and S. M. Lawrie. “first do no harm.” a systematic review of the prevalence and management of antipsychotic adverse effects. *Journal of psychopharmacology*, 29(4):353–362, 2015.