Counterfactual Cocycles: Noise-Invariant and Coherent Transport-based Couplings

Hugh Dance HUGHWDANCE@GMAIL.COM
Gatsby Computational Neuroscience Unit
University College London
London, United Kingdom

Benjamin Bloem-Reddy BENBR@STAT.UBC.CA
Department of Statistics
University of British Columbia
Vancouver, Canada

Abstract

Estimating joint distributions (a.k.a. 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 use 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 *average treatment effect* (ATE) of a binary treatment $X \in \{0, 1\}$ on a set of outcomes $Y := (Y_1, \ldots, Y_p) \in \mathbb{R}^p$,

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

The ATE factorizes into the treated and control means. In a randomized control trial (RCT), these can be estimated by simple averages (e.g., $\hat{\mathbb{E}}[Y(1)] = \frac{1}{n_1}\sum_{i=1}^{n_1} Y^{(i)}(1)$ for treatment units $\{Y^{(i)}(1)\}_{i=1}^{n_1} \sim_{iid} \mathbb{P}_{Y(1)}$). Now, consider instead the *treatment harm rate* (THR) (Shen et al., 2013), which measures the probability that treatment worsens outcomes,

$$\text{THR} := \mathbb{P}\big(Y(1) - Y(0) \preceq 0\big) \;:=\; \iint \mathbf{1}\big\{\exists j : [y_1]_j - [y_0]_j \leq 0\big\}\, d\mathbb{P}_{Y(1),Y(0)}(y_1, y_0) \qquad (1)$$

The THR is crucial to determine 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 marginals $\mathbb{P}_{Y(1)}$ and $\mathbb{P}_{Y(0)}$ alone, instead requiring the joint distribution $\mathbb{P}_{Y(1),Y(0)}$, which is the coupling of the marginal counterfactual distributions. Unfortunately, one cannot identify the coupling using the marginals alone, as there may be infinitely many couplings that admit these marginals (Villani, 2021). It also cannot be directly estimated from data, since at most one of $Y(1)$, $Y(0)$ are ever observed per 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 forms of treatment risk assessment (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

$$Y \;=\; f\big(X, \xi\big), \quad \xi \sim \mathbb{P}_\xi, \quad \xi \perp\!\!\!\perp X,$$

where ξ 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

$$\big(Y(1)\,,\, Y(0)\big) \;=\; \big(f(1,\xi)\,,\, f(0,\xi)\big)\,, \quad \xi \sim \mathbb{P}_\xi\,.$$

To ensure the pair $(f, \mathbb{P}_\xi)$ can be identified up to model-specific automorphisms, present theory requires $f(x, \bullet)$ 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., $\hat{\mathbb{P}}_\xi = \mathsf{N}(0, I)$) 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' whilst satisfying the desired change (Lewis, 1973). The basic idea is to specify transports between counterfactuals $Y(1) = T_{0,1}(Y(0))$ and estimate them by solving a (quadratic cost) OT problem,

$$T^*_{0,1} = \underset{T_{0,1}:\mathbb{P}_{Y(0)}\mapsto\mathbb{P}_{Y(1)}}{\arg\min} \mathbb{E}\left[\left\|Y(1) - T_{0,1}(Y(0))\right\|_2^2\right] .$$

OT avoids specifying any noise distributions and guarantees transport identifiability for continuous distributions (Villani, 2021). However, as we show in Section 2.2, when treatments take more than two values (e.g., $X \in \{0, 1, 2\}$) and outcomes are multivariate, OT can fail to identify the counterfactual coupling. This is because OT maps are generally not closed under composition: $T_{0,2} \neq T_{0,1} \circ T_{1,2}$ (see Example 2 in Section 2.2) and so there may not exist a coupling over $\{Y(0), Y(1), Y(2)\}$ that is consistent with the pairwise transports.

1.1 Our Contributions

We develop a modeling framework for counterfactual couplings that is free of the noise misspecification problems of SCMs and the incoherence of OT methods. Here we summarize the main contributions and paper plan.

In Section 3, we focus on a simple set-up with counterfactuals $\{Y(x)\}_{x\in\mathbb{X}}$ under a randomized treatment $X \in \mathbb{R}$, and ask what properties a set of transports $\{T_{x,x'}\}_{x,x'\in\mathbb{X}}$ must satisfy to couple them:

$$Y(x) = T_{x,x'}(Y(x')) . \tag{2}$$

It turns out that the *necessary* properties are precisely that

$$\underbrace{T_{x,x} = \mathrm{id}}_{\text{Identity}} \quad \text{and} \quad \underbrace{T_{x'',x'} \circ T_{x',x} = T_{x'',x}}_{\text{Path Independence}} , \tag{3}$$

up to a $\mathbb{P}_{Y|X=x}$-null set, along with a marginal-matching property. Moreover, these properties are *sufficient* to induce an admissible coupling over *some* set of counterfactuals (Theorem 2). The path-independence property is the key ingredient missing from OT methods. The properties in (3) make the function $T : (x, x', y) \mapsto T_{x,x'}(y)$ a *cocycle* (Arnold, 1998) from dynamical systems theory. We hence call (2) a *counterfactual cocycle* model. In Theorem 3 we show that any counterfactual cocycle T can be written as

$$T_{x,x'} = f_x \circ f^+_{x'} , \tag{4}$$

for some injective f_x with left inverse f^+_x. This gives a natural route to parameterizing valid classes of counterfactual transports: one can use any class of conditional bijectors

$\mathcal{F} := \{f_\theta : \mathbb{X} \times \mathbb{Y} \to \mathbb{Y} \mid \theta \in \Theta\}$, such as normalizing flows (Papamakarios et al., 2021) or invertible neural networks (Ishikawa et al., 2023). In Theorem 5 we provide general conditions for counterfactual cocycle identifiability, and show that one can achieve these conditions using knowledge of the causal ordering (Theorem 6). We conclude the section by showing that, under standard counterfactual assumptions, the model (2) corresponds to an equivalence class of SCMs: $Y(x) = f_x(\xi) \implies Y = f(X, \xi)$ (Theorem 7)—each one with a different noise distribution $\mathbb{P}_\xi$. Our key insight is that, since the cocycle is invariant to the noise distribution, directly modeling it may sidestep any noise mis-specification problems.

In Section 4, we present counterfactual cocycle models under a more general setup that can handle confounding, and cover our high-level approach to estimating causal quantities with cocycles. In the unconfounded case where $Y(x) \sim \mathbb{P}_{Y|X=x}$, the basic idea is as follows. In contrast to flow-based SCMs, which estimate a map from a fixed noise law $\hat{\mathbb{P}}_\xi$ to each $\mathbb{P}_{Y|X=x}$, we aim to directly target the cocycle between conditionals $T_{x,x'} = f_x \circ f_{x'}^+ \colon \mathbb{P}_{Y|X=x'} \mapsto \mathbb{P}_{Y|X=x}$, by minimizing a distributional discrepancy of the form

$$\ell(T) = \mathbb{E}_{x,x' \sim \mathbb{P}_X} D\big(\mathbb{P}_{Y|X=x'},\ (T_{x,x'})_{\#}\, \mathbb{P}_{Y|X=x}\big)^2 . \tag{5}$$

Once T is estimated, we can impute any counterfactuals and use them to empirically estimate causal quantities, rather than sampling from a possibly mis-specified noise law. In Section 5, we derive a tractable optimization procedure and show that it is asymptotically equivalent to minimizing (5) (Theorem 8 and Proposition 9). The resulting cocycle estimators are consistent under general conditions (Theorem 10) and satisfy a noise-robustness property under appropriate function class restrictions: consistency for one underlying SCM implies consistency for all SCMs that share the same cocycle (see Remark 11 and Fig. 6). We also show $\sqrt{n}$-consistency for parametric classes under regularity conditions (Theorem 12).

In Section 6 we analyze the mis-specification robustness of modeling counterfactual cocycles versus SCMs. A central advantage is that the cocycle is *noise–invariant*. Specifically, let $Y(x) = f(x, \xi)$ be an SCM with latent $\xi \sim \mathbb{P}_\xi$, so that $T_{x,x'} := f_x \circ f_{x'}^+$ with $f_x := f(x, \bullet)$. Any bijective reparameterization of the SCM $f_x(\xi) = f_x \circ g \circ g^{-1}(\xi) = f_x^{(g)}(\xi^{(g)})$ affects the structural map and noise distribution, but leaves the cocycle itself unchanged:

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

This invariance means that, when modeling a cocycle T via a class of functions $\mathcal{F}$ for f, we are not tied to a single latent distribution. It is enough that there exists *some* choice of base distribution $\mathbb{P}_\xi^\star$ and $f \in \mathcal{F}$ such that $(f_x)_{\#}\, \mathbb{P}_\xi^\star = \mathbb{P}_{Y|X=x}$ for all x. By contrast, a flow-based SCM commits to a particular base $\hat{\mathbb{P}}_\xi$. If that choice is ill-matched (e.g. in tails or support), the true maps f_x may be discontinuous or non-bijective, making the class $\mathcal{F}$ mis-specified. Viewed differently, since every reparameterization $f^{(g)}$ induces the same cocycle, one can always model the cocycle using the reparameterization $f^{(g^*)}$ that has the lowest functional complexity (e.g., Sobolev norm). As we prove in Theorem 13 and show in Example 3, this 'minimal complexity' property permits using smaller model classes $\mathcal{F}$ for the cocycle than for an SCM with a fixed base distribution, while remaining well-specified.

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 $\boldsymbol{V} := (V_1, \ldots, V_d) \in \prod_j^d \mathbb{V}_j \subset \mathbb{R}^d$, be observed random variables with distribution $\mathbb{P}_{\boldsymbol{V}}$. For most of this work we assume $V = (Z, X, Y)$, where $X \in \mathbb{R}^l$ are a set of (treatment) variables we wish to manipulate (e.g., doses of different medications), $Y \in \mathbb{R}^q$ are outcomes of interest (e.g., patient blood pressure, resting heart rate etc.), and $Z \in \mathbb{R}^l$ are pre-treatment covariates (e.g., age, gender). We denote $Y(x)$ as the counterfactual outcome to Y under a 'do' intervention that fixes treatment levels to x (often denoted $\text{do}(X = x)$) (Pearl, 2009a; Chernozhukov et al., 2025). Note, when dealing with an i.i.d. dataset of observations, we index the samples as $(\boldsymbol{V}^{(i)})_{i=1}^n \sim_{iid} \mathbb{P}_{\boldsymbol{V}}$.

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 π over *collections* of counterfactuals, e.g.,

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

Here $\mathcal{L}(\{Y(x)\}_{x \in I \subset \mathbb{X}})$ 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_α is the α-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:

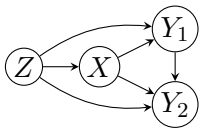
$$\mu(\alpha) := \mathbb{E}[Y(1) - Y(0) | Y(0) = q_\alpha] \,.$$

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 π 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.

$$
\begin{aligned}
Z &= f_1(\xi_1)\\
X &= f_2(Z, \xi_2)\\
Y_1 &= f_3(Z, X, \xi_3)\\
Y_2 &= f_4(Z, X, Y_1, \xi_4)
\end{aligned}
$$



$$
\begin{aligned}
Z &= f_1(\xi_1)\\
X(x) &= x\\
Y_1(x) &= f_3\big(Z, X(x), \xi_3\big)\\
Y_2(x) &= f_4\big(Z, X(x), Y_1(x), \xi_4\big)
\end{aligned}
$$

Figure 1: Left: SCM over (Z, X, Y_1, Y_2). Middle: corresponding maximal causal DAG. Right: modified SCM after the hard intervention $\mathrm{do}(X = x)$.

2.1 Structural Causal Models

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

$$\boldsymbol{V} = F(\boldsymbol{V}, \boldsymbol{\xi}) \,. \tag{6}$$

For identifiability, one usually assumes the variables $\boldsymbol{V}$ admit a known *causal ordering* $V_1 \prec \cdots \prec V_d$. 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*:

$$V_j = f_j(V_{<j}, \xi_j), \quad \forall j \in \{1, \ldots, d\} \,. \tag{7}$$

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

Every acyclic SCM can be associated with a causal directed acyclic graph (DAG) $\mathcal{G} = (\boldsymbol{V}, E)$, with nodes $\{V_1, \ldots, V_d\}$ and edges $E = \{V_i \to V_j : i < j\}$ 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 $V_1, \ldots, V_{j-1}$, 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

$$V_j = f_j(V_{\mathrm{pa}(j)}, \xi_j) \,,$$

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

The counterfactuals under a do-intervention $\mathrm{do}(X = x)$ for any subset $X \subset \boldsymbol{V}$ are determined by a modified SCM $\boldsymbol{V}(x) = F_x(\boldsymbol{\xi})$, 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 $\mathrm{do}(X = x)$.

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

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

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

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

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)	$\xi_j + \mu_j(v_{<j})1_{k\notin[j]}$	μ_j Lipschitz
MAF (Papamakarios et al., 2017)	$\xi_j \mapsto \sigma_j(v_{<j}) \cdot \xi_j + (1-\sigma_j(v_{<j}))\mu_j(v_{<j})$	σ_j bounded
IAF (Kingma et al., 2016)	$\xi_j \mapsto \exp\big(\lambda_j(v_{<j})\big)\,\xi_j + \mu_j(v_{<j})$	λ_j bounded, μ_j Lipschitz
Real-NVP (Dinh et al., 2016)	$\xi_j \mapsto \exp\big(\lambda_j(v_{<j})1_{k\notin[j]}\big)\,\xi_j + \mu_j(v_{<j})1_{k\notin[j]}$	λ_j bounded, μ_j Lipschitz
Glow (Kingma and Dhariwal, 2018)	$\xi_j \mapsto \sigma_j(v_{<j})\xi_j + \mu_j(v_{<j})1_{k\notin[j]}$	σ_j bounded, μ_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 of those models that can identify counterfactual couplings are *bijective causal models (BCMs)*, which constrain each f_j to be bijective on the noise ξ_j. In particular, if two BCMs $(F, \mathbb{P}_{\boldsymbol{\xi}})$, $(\tilde{F}, \tilde{\mathbb{P}}_{\boldsymbol{\xi}})$ produce the same distribution $\mathbb{P}_{\boldsymbol{V}}$, and each structural function f_j and $\tilde{f}_j$ is monotone increasing on ξ_j (note for maps $\mathbb{R} \to \mathbb{R}$, bijectivity and strict monotonicity are equivalent), they produce the same counterfactual couplings (Nasr-Esfahany et al., 2023). BCMs are 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., $\hat{\mathbb{P}}_{\boldsymbol{\xi}} = \mathsf{N}(0, I)$) (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 $\mathbb{P}_{\boldsymbol{V}}$, 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 BCMs 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 BCMs: tail and support mis-specification.

Tail Mis-specification It is known that if the base distribution $\hat{\mathbb{P}}_{\boldsymbol{\xi}}$ lies in a different tail class to the observational distribution $\mathbb{P}_{\boldsymbol{V}}$ (e.g., exponential vs. logarithmic tail decay), then no regular class of bi-Lipschitz flows can match the tails of $\mathbb{P}_{\boldsymbol{V}}$ (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

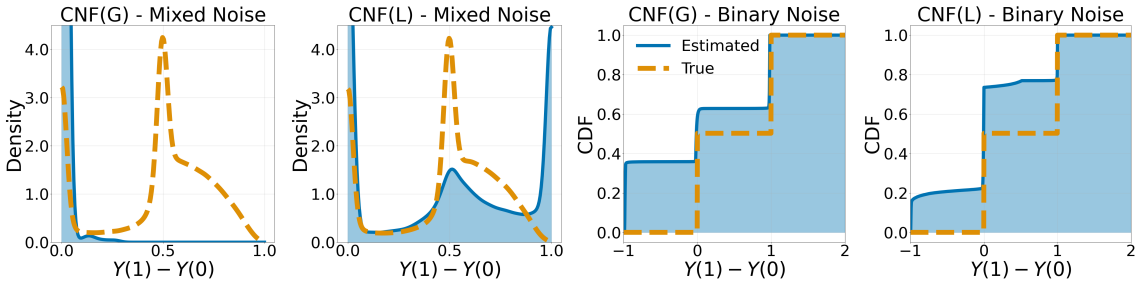


Figure 2: True (orange) vs. estimated (blue) distribution of the treatment effect $Y(1)-Y(0)$ for the SCM $Y(x) = (x+1)\xi$, under two different noise laws: (LHS) mixed-tailed ξ (density plots) and (RHS) binary ξ (CDF plots). Estimates are obtained with flow-based SCMs using Gaussian (CNF(G)) and Laplace (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 $\mathbb{E}\|\boldsymbol{V}\|_2 = \infty$ and $\hat{\mathbb{P}}_\xi = \mathsf{N}(\mathbf{0}, I)$ is chosen, the gradient may even be infinite and so the minimizer $F_{\theta_n^*}$ can diverge as $n \to \infty$.

Support Mis-specification If there is no differentiable function that transports $\hat{\mathbb{P}}_\xi$ to $\mathbb{P}_{\boldsymbol{V}}$, 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 ξ_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 $\mathbb{P}_{\boldsymbol{V}}$. As a result, if a minimizer F_{θ^*} satisfies $\text{supp}((F_{\theta^*}^{-1})_\# \mathbb{P}_{\boldsymbol{V}}) \not\supseteq \text{supp}(\hat{\mathbb{P}}_{\boldsymbol{\xi}})$, it may not be unique on $\text{supp}(\hat{\mathbb{P}}_{\boldsymbol{\xi}})$, and so each minimizer may result in a different distribution.

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

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

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 $\xi \sim \frac{1}{2}|\mathsf{N}(0,1)| - \frac{1}{2}|\text{NBP}(0.1, 0.1)|$.[1]

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 was trained by maximum likelihood on $n = 2000$ samples $\{X^{(i)}, Y^{(i)}\}_{i=1}^n$, using 2-fold cross-validation over Neural Spline Flow (NSF) (Durkan et al., 2019) and Masked Autoregressive Flow (MAF) architectures (Papamakarios et al., 2017). The full architecture and optimization settings are described 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 σ does not indicate a learned heavy left tail.

We next re-implement the same flow-based models with binary noise, $\xi \sim \mathsf{Rad}(1/2)$, 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 $\mathbb{P}_{\xi_Y}$. Since each model only evaluates the likelihood on a set of measure zero under the base distribution, there are potentially infinitely many transports achieving the same likelihood. Thus, both flows fail to learn the size and placement of the modes of the distribution (as indicated by the jumps in the learned vs. true CDF). Note that if X was continuous then Y would also be continuous in this setting, so such mis-specification can arise in practice.

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,

$$Y(x) = T_{x,x'}(Y(x')) \,, \quad x, x' \in \mathbb{X} \,.$$

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 $\mathbb{P}_{Y(x)}$,

$$T^*_{x,x'} = \arg\min_{T_{x,x'}:T_{x,x'}(Y(x'))=_d Y(x)} \mathbb{E}\left[\|Y(x) - T_{x,x'}(Y(x'))\|_2^2\right] \tag{8}$$

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 $Y := (Y_1, \ldots, Y_p)$. 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 $\hat{\mathbb{P}}_{Y(x)} = \frac{1}{n_x}\sum_{i=1}^{n_x} \delta_{Y^{(i)}(x)}$ 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).

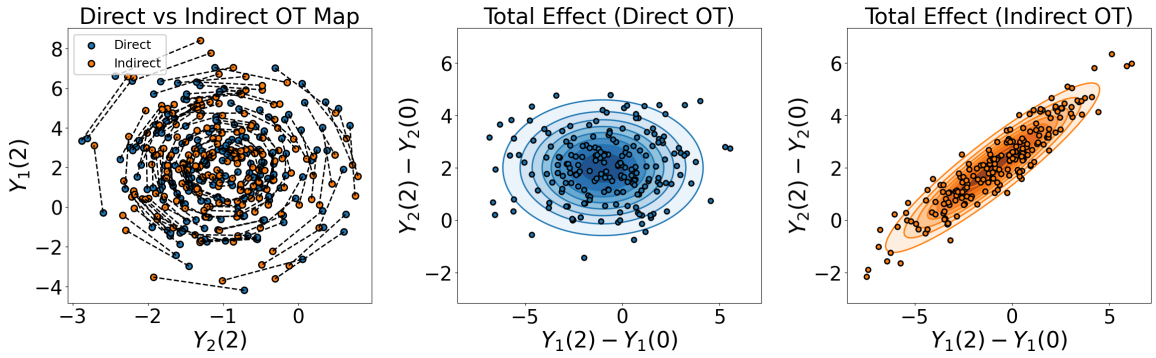


Figure 3: Transport of 200 samples $Y(0) \sim \mathbb{P}_0$ to $\mathbb{P}_1$ and $\mathbb{P}_2$ using OT maps. Blue: direct route $(Y(1), Y(2)) = (T_{1,0}(Y(0)), T_{2,0}(Y(0))$. Orange: indirect route $(T_{1,0}(Y(0)), T_{2,1} \circ T_{1,0}(Y(0))$. Left: Imputed $Y(2)$ under each method; dashed lines 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 certain 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) and the composition of two such functions will generally not yield another when $\dim(\mathbb{Y}) > 1$ (Santambrogio, 2015). As a result, counterfactual predictions depend on the path taken through treatment space, and so there is generally no coupling consistent with all pairwise transports. As the following example illustrates, this results in an identifiability problem.

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

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

When the covariance matrices do not commute (as is the case for (Σ_1, Σ_2)), the maps are not closed under composition, in the sense that $T_{x,x''} \neq T_{x,x'} \circ T_{x',x''}$ for $x \neq x' \neq x''$. As a result, there is no admissible coupling over $\{Y(0), Y(1), Y(2)\}$ that corresponds to these maps. 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 valid 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, a unique coupling cannot be identified 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 coupling identifiability under arbitrary treatment conditions (as in SCMs). To that end, in this section we analyze the algebraic properties that a set of transports $\{T_{x,x'}\}_{x,x' \in \mathbb{X}}$ must satisfy to induce a valid coupling between counterfactual outcomes $\{Y(x)\}_{x \in \mathbb{X}}$, in a simplified confounding-free setting. We show that these properties are precisely those of a *cocycle*. The cocycle properties induce a factorization of the transports so that they can always be constructed by a family of injective functions. We use this structure to prove general conditions under which such cocycles are identifiable. We then show that every counterfactual cocycle model corresponds to an equivalence class of injective SCMs. This connection gives a practical route for parameterizing identifiable model classes, and suggests that directly modeling counterfactual cocycles can avoid the mis-specification problems of SCMs.

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, $Y := (Y_1, \ldots, Y_p) \in \mathbb{Y} \subset \mathbb{R}^p$ are a set of outcomes of interest, and $\{Y(x)\}_{x \in \mathbb{X}} \in \mathbb{Y}$ are the counterfactuals under different treatment levels. Throughout, we assume all spaces (here $\mathbb{X}, \mathbb{Y}$) are standard Borel. To formalize counterfactuals under 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 block causal ordering of the form $X \prec [Y_1, \ldots, Y_p]$ or "coarse-grained" causal DAG $X \to Y$ (Richardson and Robins, 2013). In either formulation, the marginal counterfactual distribution is identified 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* $\{T_{x,x'} : \mathbb{Y} \to \mathbb{Y} \mid x, x' \in \mathbb{X}\}$ *that satisfy the* ***Counterfactual Coupling*** (CC) *property:*

$$Y(x) = T_{x,x'}(Y(x')) \,, \qquad \textit{for all } x, x' \in \mathbb{X} \,. \tag{CC}$$

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 $x \in \mathbb{X}$, $\exists\, \mathbb{Y}_x \subseteq \mathbb{Y}$ such that $\mathbb{P}_{Y|X=x}(\mathbb{Y}_x) = 1$ and
$$T_{x,x}(y) = y, \quad \forall y \in \mathbb{Y}_x \tag{ID}$$
2. **Path Independence**: For each $x \in \mathbb{X}$, $\exists\, \mathbb{Y}_x \subseteq \mathbb{Y}$ such that $\mathbb{P}_{Y|X=x}(\mathbb{Y}_x) = 1$ and
$$T_{x'',x'} \circ T_{x',x}(y) = T_{x'',x}(y), \quad \forall y \in \mathbb{Y}_x, \quad \forall x', x'' \in \mathbb{X} \tag{PI}$$
3. **Distribution Adaptedness**: For each $x, x' \in \mathbb{X}$,
$$(T_{x',x})_{\#}\mathbb{P}_{Y|X=x} = \mathbb{P}_{Y|X=x'} \tag{DA}$$

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 $x \to x'$. If a cocycle also satisfies (DA) we call it a $\mathbb{P}_{Y|X}$-*adapted cocycle*. Lastly, if a cocycle T 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. If (PI) fails, for example, the composition $T_{x_2,x_1} \circ T_{x_1,x_0}$ may differ from T_{x_2,x_0}, leading to logical impossibilities in the counterfactual outcomes:
$$Y(x_2) = T_{x_2,x_0}(Y(x_0)) \neq T_{x_2,x_1} \circ T_{x_1,x_0}(Y(x_0)) = Y(x_2).$$
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 invertibility ($T_{x,x'} = T_{x',x}^{-1}$ on $\mathbb{Y}_x$), 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* sufficient to ensure that there exists *some* set of variables almost-surely linked by the transports as in (CC) and thus guarantee the transports can construct an admissible coupling over the marginals $(\mathbb{P}_{Y|X=x})_{x\in\mathbb{X}}$. This is formalized below.

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

Theorem 2 (Cocycle Sufficiency for Admissibility)**.** *Let Assumption 1 hold. If* $\{T_{x,x'}\}_{x,x'\in\mathbb{X}}$ *satisfy* (ID), (PI), *and* (DA) *w.r.t.* $\mathbb{P}_{Y|X}$, *then they are admissible w.r.t.* $\mathbb{P}_{Y|X}$.

Given these results, our aim is to develop a framework for modeling and estimating counterfactual cocycles. In the rest of this section we focus on the modeling aspects.

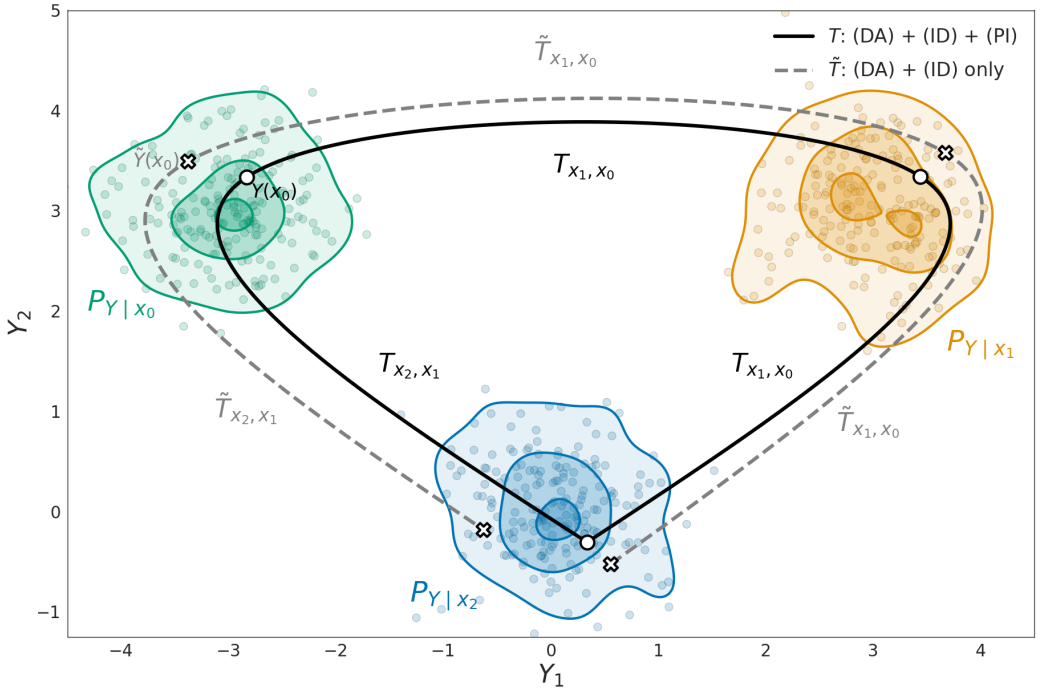


Figure 4: Counterfactual trajectories of two units: (a) $Y(x_0)$ (solid, black) and (b) $\tilde{Y}(x_0)$ (dashed, grey) under two respective transport collections $T, \tilde{T}$. $\tilde{T} = \{\tilde{T}_{x,x'}\}$ satisfies (DA) and (ID) only, while $T = \{T_{x,x'}\}$ also enforces (PI). Because $\tilde{T}$ fails (PI), the endpoint $\tilde{Y}(x_2)$ depends on the chosen path ($x_0 \to x_2$ vs. $x_0 \to x_1 \to x_2$) and so it cannot be a counterfactual cocycle (CC).

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 $f_x : \mathbb{Y}_0 \to \mathbb{Y}$ with left-inverse $f_x^+ : \mathbb{Y} \to \mathbb{Y}_0$, such that for every $x' \in \mathbb{X}$:*

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

The proof is simple: one defines $f_x := T_{x,x_0}$ and $f_x^+ := T_{x_0,x}$ for any fixed $x_0 \in \mathbb{X}$, and applies (ID) and (PI) to verify that $T_{x,x'} = f_x \circ f_{x'}^+$. When f_x^+ is an exact inverse, $f_x \circ f_{x'}^+$ is a special kind of cocycle known as a coboundary (Varadarajan, 1968). Hence, we call a function $f : (x, y) \mapsto f_x(y)$ that satisfies $T_{x,x'} = f_x \circ f_{x'}^+$ a *coboundary map*.

Theorem 3 shows that the construction of a parameterized family of cocycles $\{T_\theta\}_{\theta\in\Theta}$ reduces to specifying a parameterized family of coboundary maps—i.e., functions $f : \mathbb{X} \times \mathbb{Y}_0 \to \mathbb{Y}$ that are injective in their last argument (a.k.a. conditionally injective),

$$\mathcal{F} \subseteq \left\{ f : \mathbb{X} \times \mathbb{Y}_0 \to \mathbb{Y} \;\middle|\; f(x, \bullet) \text{ injective } \forall x \in \mathbb{X} \right\}.$$

Parameterized classes of conditionally *bijective* functions, such as normalizing flows and invertible neural networks (Papamakarios et al., 2021; Teshima et al., 2020; Ishikawa et al., 2023), are commonplace in statistics and machine learning and so can be used for this task. We note that specializing to bijections is not overly restrictive. The true f_x in Theorem 3 is already bijective on $\mathbb{Y}_0 \to \mathbb{Y}_x$, where $\mathbb{Y}_x$ is the full-measure set used in (ID), (PI). 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 $\tilde{f}_x : \mathbb{Y} \to \mathbb{Y}$ of f_x (Kechris, 2012, Thm. 15.6). For instance, $f_x(y) = \beta x + y$ with $\mathbb{Y}_0 = [0, 1]$ and $\mathbb{Y}_x = [x, x + 1]$ extends naturally to $\mathbb{Y} = \mathbb{R}$. Although the extension may not always be continuous, parameterized classes of continuous bijections (e.g., flows, invertible NNs), remain highly expressive.

General Conditions for Identifiability Theorem 3 provides a general route for parameterizing flexible classes of counterfactual cocycles, but does not provide any guidance on how to constrain $\mathcal{F}$ so that at most one cocycle T constructed by functions in $\mathcal{F}$ satisfies (ID), (PI) and (DA). Indeed, many sets of transports may satisfy these properties, but at most one satisfies (CC) for a particular set of counterfactual variables.

To demonstrate this point, a collection of OT maps $\{T^{(OT)}_{x,x'}\}_{x\in\mathbb{X}}$ can in principle be used as coboundary maps to construct a cocycle that satisfies (ID), (PI) and (DA). One can simply define $f_x := T^{(OT)}_{x,x_0}$ as above for some x_0, and define a $\mathbb{P}_{Y|X}$-adapted cocycle $\tilde{T}_{x,x'} := f_x \circ f^+_{x'}$. 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. 9: the middle plot shows the distribution of the treatment effect when the reference point $x_0 = 0$ is implicitly used, while the RHS plot shows the same distribution using the reference $x_0 = 1$.

The factorization structure of cocycles enables us to provide general conditions for cocycle identifiability. First, we formally define what it means for a cocycle to be identifiable.

Definition 4 (Identifiability)**.** *Let $\mathcal{F}$ be a set of coboundary maps $f : \mathbb{X} \times \mathbb{Y}_0 \to \mathbb{Y}$ injective in their second argument. For each $f \in \mathcal{F}$ define the associated cocycle $T_f(x, x') := f_x \circ f^+_{x'}$. A $\mathbb{P}_{Y|X}$-adapted cocycle T is **identifiable in $\mathcal{F}$** if, (i) $\exists f^\star \in \mathcal{F}$ with $T =_{\text{a.s.}} T_{f^\star}$; (ii) for any $f \in \mathcal{F}$, if T_f is a $\mathbb{P}_{Y|X}$-adapted cocycle satisfying (ID), (PI) and (DA), then $T_f =_{\text{a.s.}} T$.*

A convenient way to state general identifiability results is in terms of transformation groups.[3] A *transformation group* $\mathbb{G}$ on $\mathbb{Y}$ is a set of bijective transformations $g : \mathbb{Y} \to \mathbb{Y}$ that contains the identity and is closed under compositions and inverses. We say that a cocycle T is **$\mathbb{G}$-valued** if its coboundary map f satisfies $f(x, \bullet) := f_x \in \mathbb{G}$ for every $x \in \mathbb{X}$,

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 coboundary maps are bijections (and hence group-valued), but a nearly identical version holds when they are injections, in terms of a more complicated algebraic construction involving monoids instead of groups.

and write $\mathcal{F}_{\mathbb{G}}$ as the set of $\mathbb{G}$-valued coboundary maps,

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

For example, whenever $f(x, \bullet) : \mathbb{Y} \to \mathbb{Y}$ is bijective, it belongs to the automorphism group of $\mathbb{Y}$, denoted $\mathrm{Aut}(\mathbb{Y})$. If f_x is also bi-continuous, it belongs to the group of self-homeomorphisms, $\mathrm{Homeo}(\mathbb{Y})$. As discussed above, these are practically relevant in many situations. We denote $\mathrm{Aut}(P) := \{g \in \mathrm{Aut}(\mathbb{Y}) : g_{\#}P = P\}$ as the set of transformations that leave distribution $P \in \mathcal{P}(\mathbb{Y})$ unchanged, and $\mathrm{Aut}(P)|_{\mathbb{G}} := \mathrm{Aut}(P) \cap \mathbb{G}$ its restriction to $\mathbb{G}$. Lastly, we denote $[f]_P$ as the set of functions equal to f, P-almost surely.

Theorem 5 (Identifiability of Counterfactual Cocycle)**.** *Fix arbitrary $x_0 \in \mathbb{X}$, and suppose that T is a $\mathbb{P}_{Y|X}$-adapted, $\mathbb{G}$-valued cocycle with coboundary map f. Then $\tilde{T}$ is another $\mathbb{G}$-valued, $\mathbb{P}_{Y|X}$-adapted cocycle with coboundary map $\tilde{f}$, if and only if there exist functions $\{b_x\}_{x\in\mathbb{X}}$ with $b_x \in \mathrm{Aut}(\mathbb{P}_{Y|X=x_0})|_{\mathbb{G}}$ for each $x \in \mathbb{X}$, such that for all $x, x' \in \mathbb{X}$,*

$$\tilde{f}_x \circ \tilde{f}_{x'}^{-1} = f_x \circ b_x^{-1} \circ b_{x'} \circ f_{x'}^{-1} , \quad \mathbb{P}_{Y|X=x'}\text{-a.s.} \tag{9}$$

Therefore, T is identifiable in $\mathcal{F}_{\mathbb{G}}$ if and only if $\mathrm{Aut}(\mathbb{P}_{Y|X=x_0})|_{\mathbb{G}} \subseteq [\mathrm{id}]_{\mathbb{P}_{Y|X=x_0}}$.

Theorem 5 states that as long as we choose our model class $\mathcal{F}$ to contain only $\mathbb{G}$-valued cocycles (i.e., $\mathcal{F} \subseteq \mathcal{F}_{\mathbb{G}}$), and $\mathbb{P}_{Y|X=x_0}$ is only invariant under transformations in $\mathbb{G}$ that are equivalent to the identity map, then the cocycle is identifiable from $\mathcal{F}$. Note that the choice of x_0 here is arbitrary. This is because for a $\mathbb{G}$-valued, $\mathbb{P}_{Y|X}$-adapted cocycle to exist, for any pair $x, x' \in \mathbb{X}$, $\mathbb{G}$ must contain a transformation $g_{x,x'}$ such that $(g_{x,x'})_{\#} : \mathbb{P}_{Y|X=x} \mapsto \mathbb{P}_{Y|X=x'}$. Thus, if $g \neq_{\text{a.s.}} \mathrm{id}$ is an automorphism of $\mathbb{P}_{Y|X=x_0}$, then $g_{x,x_0} \circ g \circ g_{x,x_0}^{-1} \neq_{\text{a.s.}} \mathrm{id}$ is an automorphism of $\mathbb{P}_{Y|X=x}$.

Identifiable Parameterizations via TMI Maps In general, a smaller group $\mathbb{G}$ preserves identifiability with respect to more 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 the outcomes. In particular, suppose $Y = (Y_1, \ldots, Y_p) \in \mathbb{R}^p$ admit a known causal order $Y_1 \prec \cdots \prec Y_p$ (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 f_x:

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

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 counterfactual cocycles whilst guaranteeing identifiability. Although such architectures lead to mis-specification problems in SCMs in Section 2.1, we will later see that using them to model counterfactual cocycles avoids those problems.

3.3 Connection to SCMs

The factorization structure of cocycles and their viable parameterization using the same function classes as BCMs 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 if and only if there is a function $f : \mathbb{X} \times \mathbb{Y}_0 \to \mathbb{Y}$ injective on its second argument, such that*

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

The proof is straightforward: using f_x, f_x^+ as defined in Theorem 3, define $\xi_Y := f_{x_0}^+(Y(x_0))$ and apply (CC) to get $Y(x) = f(x, \xi_Y)$, where $f(x, \xi_Y) := f_x(\xi_Y)$. Assumption 1 then gives $Y = f(X, \xi_Y)$ by consistency and $\xi_Y \perp\!\!\!\perp X$ by exchangeability.

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

$$\boldsymbol{V} := \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \xi_X \\ f(\xi_X, \xi_Y) \end{pmatrix} =: F(\boldsymbol{\xi}) \ , \tag{10}$$

where $\xi_X \perp\!\!\!\perp \xi_Y$ and $F : \mathbb{V} \to \mathbb{V}$ is injective by the injectivity of $f(x, \bullet)$. 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 $f(x, \bullet)$ is bijective (or has a bijective extension), the equivalence is to a class of BCMs.

Note that when Y is multivariate, (10) shows only a partial factorization of F. However, under the TMI restriction in Section 3.2, f further decomposes into coordinate maps,

$$Y_j \ = \ f_j\big(X, Y_{<j}, \xi_{Y,j}\big), \quad \xi_{Y,j} \perp\!\!\!\perp X,$$

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.

Noise Invariance of Counterfactual Cocycles A key difference between counterfactual cocycles and SCMs is that cocycles are *noise-invariant*. In particular, the transport

$$T_{x,x'} = f_x \circ f_{x'}^+ \quad , \quad (T_{x,x'})_\# : \mathbb{P}_{Y|X=x} \mapsto \mathbb{P}_{Y|X=x'} \ , \tag{11}$$

does not change if we: (a) modify the SCM by changing the noise law $\mathbb{P}_\xi$, or (b) reparameterize the same SCM using an automorphism $g \in \mathrm{Aut}(\mathcal{E})$ via $f_x(\xi) = f_x \circ g \circ g^{-1}(\xi) =: f_x^{(g)}(\xi^{(g)})$. In both cases, we can still construct the cocycle with the original coboundary map f. This invariance suggests a practical robustness to directly targeting a cocycle that solves (11). In particular, the SCM approach to modeling the system in (10) is to fix some base distribution $\hat{\mathbb{P}}_{\xi_Y}$ and class of functions 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.,

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

depends on whether the properties (e.g. support and tails) of the *chosen* base distribution $\hat{\mathbb{P}}_{\xi_Y}$ matches $\mathbb{P}_{Y|X}$. In contrast, for there to exist a function $\hat{f} \in \mathcal{F}$ that solves (11) for every $x, x' \in \mathbb{X}$, we just require that the above holds w.r.t. *some* noise distribution $\mathbb{P}^*_\xi \in \mathcal{P}(\mathbb{Y})$:

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

This is a much weaker requirement and is indeed satisfied in Example 1 when choosing $\mathcal{F}$ according to the flows in Table 1. These properties underpin the estimation approach that we develop in the following sections, and are analyzed in greater detail in Section 6.

Relation to Conditional Transport Methods SCMs estimated with TMI maps have recently been used to construct counterfactual transports in fairness applications (Plečko and Meinshausen, 2020; Machado et al., 2024). The main idea is that, by choosing $\xi_{Y,j} \sim \mathsf{U}[0,1]$, the implied transport $T_{x,x'}$ decomposes into a set of one-dimensional conditional transports that can be estimated using quantile transforms. 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). Although our cocycle-based framework is much broader, when restricting to TMI maps it bears close connections to these methods. As we discuss in detail in Section 6.4, the main differences are: (i) by fixing each noise variable to an independent uniform distribution $\xi_{Y,j} \sim \mathsf{U}[0,1]$ during estimation, they fail to exploit the noise-invariance of counterfactual cocycles; and (ii) when the true SCM noises $(\xi_{Y,i})_{j=1}^p$ are dependent, the induced transports can be biased.

4 Counterfactual Cocycles: General Framework and 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 $\boldsymbol{V} := (V_1, \ldots, V_d) \in \prod_j^d \mathbb{V}_j \subset \mathbb{R}^d$ now denote the full set of observed variables. We retain focus on a structured setting where $X := (X_1, \ldots, X_q) \subset \boldsymbol{V}$ are a set of ('treatment') variables we wish to manipulate, $Y := (Y_1, \ldots, Y_p) \subset \boldsymbol{V}$ are a set of outcomes of interest that may be affected by the treatments, and $Z := (Z_1, \ldots, Z_l) \subset \boldsymbol{V}$ 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,

$$Z \prec X \prec Y\,, \tag{12}$$

where $Z \prec X \implies Z_i \prec X_j$ for every $(i,j) \in \{1,\ldots,l\} \times \{1,\ldots,q\}$. Fig. 5 shows several causal DAGs consistent with this ordering. Although we are yet to discuss unobserved confounding, the examples in Fig. 5 are compatible with the framework we develop here. Note that, when $X \in \boldsymbol{V}$ is a single treatment, one can trivially find sets $Z, Y \subset \boldsymbol{V}$ that satisfy (12) for *any* causal DAG over $\boldsymbol{V}$. The only notable exclusion is the dynamic treatment regime, where outcomes affect subsequent treatments (e.g., $X_j \to Y_i \to X_k$).

The partial ordering (12) 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

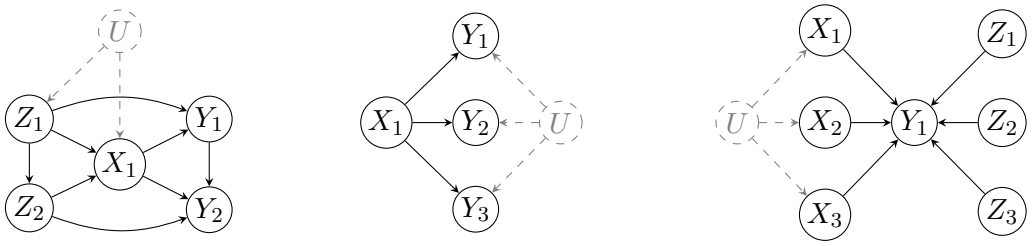


Figure 5: DAGs consistent with ordering in (12) and Assumption 3. U is unobserved.

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

We stress that these assumptions do not further restrict the causal structure implied by (12). They simply provide a compatible potential outcomes representation which lets us 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,

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

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. 5). 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 $\{Y(x)\}_{x\in\mathbb{X}}$:

$$Y(x)=T_{(x,Z),(x',Z)}(Y(x'))\,. \tag{13}$$

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 $T_{(x,z),(x',z')} = f_{x,z} \circ f^{+}_{x,z'}$ for some injective $f_{x,z} : \mathbb{Y}_0 \to \mathbb{Y}$ and that $Y = f(X, Z, \xi_Y)$, $\xi_Y \perp\!\!\!\perp (X, Z)$. Therefore, any joint BCM over the variable blocks,

$$\boldsymbol{V} = \begin{pmatrix} Z \\ X \\ Y \end{pmatrix} = \begin{pmatrix} \xi_Z \\ h(\xi_Z, \xi_X) \\ f\big(h(\xi_Z, \xi_X),\, \xi_Z,\, \xi_Y\big) \end{pmatrix} := F(\boldsymbol{\xi}) \,, \tag{14}$$

is again consistent with the cocycle T. Different choices of h and reparameterizations of the noise (ξ_Z, ξ_X, ξ_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. 5, 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 $\dim(\mathrm{supp}(\mathbb{P}_\xi)) \leq \dim(\mathrm{supp}(\mathbb{P}_Y))$ 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 identifiability we restrict our parameterizations of cocycles to TMI maps. In the present context, this essentially imposes a known partial causal ordering between the outcomes, i.e., $Y_1 \prec \dots Y_p$, since TMI maps are only unique up to an ordering of the variables. 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. In practice, we will parameterize these maps over the outcomes using autoregressive flows (Table 1), with flow parameters conditioned on X, Z. Implementation details are in Section 7.

Incorporating Causal DAGs If a specific causal DAG is known, then one can additionally restrict the transports to reflect the sparsity of the direct effects. In particular, under the TMI restriction, we know that Assumption 3 implies for each of $Y_1, \dots Y_p$,

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

If $\mathcal{G}$ is a known causal DAG over $\boldsymbol{V}$, then we can replace $(Y_{<j}, X, Z)$ with $\mathrm{pa}(Y_j)$ in $\mathcal{G}$,[4]

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

In Section 7 we discuss how to practically constrain flow architectures to reflect the DAG.

4.3 Estimating Causal Quantities with Cocycles

Any cocycle parameterized via autoregressive flows is $\mathbb{G}$-valued for some $\mathbb{G} \subset \mathrm{Aut}(\mathbb{Y})$, and so trivially satisfies (ID) and (PI) on all of $\mathbb{Y}$. Thus, all that remains is to enforce

4. Note here we abuse notation and assume $\mathrm{pa}(\bullet)$ returns the parents according to the indexes of $Y_1, \dots, Y_p$, instead of the indexes of the full variable set $V_1, \dots, V_d$.

(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, while our parameterizations mirror that of flow-based SCMs, our estimation procedure is centered entirely on the cocycle and avoids specifying a latent base distribution. This lets us take advantage of the invariances of counterfactual cocycles. 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 $T_{\tilde{x},\tilde{x}'} = f_{\tilde{x}} \circ f^+_{\tilde{x}'} : \mathbb{P}_{Y|\tilde{X}=\tilde{x}} \mapsto \mathbb{P}_{Y|\tilde{X}=\tilde{x}'}$. 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 ξ (see Remark 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 $x_1, \ldots, x_m$ using (13) and the consistency property:

$$\{\hat{Y}^{(i)}(x_1), \ldots, \hat{Y}^{(i)}(x_m)\} \;=\; \{\hat{T}_{(x_1,Z^{(i)}),(X^{(i)},Z^{(i)})}\big(Y^{(i)}\big), .., \hat{T}_{(x_m,Z^{(i)}),(X^{(i)},Z^{(i)})}\big(Y^{(i)}\big)\}$$

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_λ:

$$\text{Average Effect:} \qquad \widehat{\mathbb{E}}\big[Y(x) - Y(0)\big] = \frac{1}{n}\sum_{i=1}^n \big(\hat{Y}^{(i)}(x) - \hat{Y}^{(i)}(0)\big), \tag{15}$$

$$\text{True Harm Rate:} \qquad \widehat{\mathbb{P}}\big(Y(x) - Y(0) \preceq 0\big) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\hat{Y}^{(i)}(x) - \hat{Y}^{(i)}(0) \preceq 0\} \tag{16}$$

$$\text{Density of Effect:} \qquad \widehat{p}_{Y(x)-Y(0)}\big(y\big) = \frac{1}{n}\sum_{i=1}^n K_\lambda(\hat{Y}^{(i)}(x) - \hat{Y}^{(i)}(0) - y) \tag{17}$$

One can also condition such quantities on covariates $W \subset Z$ to examine effect heterogeneity. For this, one can replace the empirical averages in (15)-(17), with *weighted* averages estimated nonparametrically. For example, the conditional THR can be estimated as

$$\hat{\mathbb{P}}(Y(x) - Y(0) \preceq 0 \mid W = w) = \sum_{i=1}^n \hat{\alpha}_i(w)\mathbf{1}\{\hat{Y}^{(i)}(x) - \hat{Y}^{(i)}(0) \preceq 0\} \tag{18}$$

where $\hat{\alpha}_i(\cdot)$ are smoothing weights estimated via nonparametrically regressing $\mathbf{1}\{\hat{Y}^{(i)}(x) - \hat{Y}^{(i)}(0) \preceq 0\}$ 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 the 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

$$\ell_0(f) = \mathbb{E}_{X,X' \sim \mathbb{P}_X} D\big(\mathbb{P}_{Y|X}, (f_X \circ f_{X'}^{+})_{\#}\mathbb{P}_{Y|X'}\big)^2 . \tag{19}$$

where $\mathbb{P}_{Y|X} := \mathbb{P}(Y \in \bullet\,|X)$ is treated as a random probability measure on $\mathbb{Y}$, and f_X a random function $\mathbb{Y} \to \mathbb{Y}$. When the cocycle is identifiable from a model $\mathcal{F}$ (e.g., $\mathcal{F} \subset \mathcal{F}_{\mathrm{TMI}}$—see Section 3.2), then by definition ℓ_0 admits a minimizer that is almost-everywhere unique. Unfortunately, evaluating ℓ_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

$$\ell_0(T) = \mathbb{E}_{X,X' \sim \mathbb{P}_X} \|\mathbb{E}[\psi(Y)|X] - \mathbb{E}[\psi(T_{X,X'}(Y')|X, X']\|_{\mathcal{H}_k}^2 . \tag{20}$$

Here $(X, Y) \perp\!\!\!\perp (X', Y')$ are independent copies, $\psi : \mathbb{Y} \to \mathcal{H}_k$ is a feature map to a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ associated with a positive-definite kernel $k : \mathbb{Y}^2 \to \mathbb{R}$, in the sense that $\psi(y) = k(y, \cdot)$. As long as the kernel k is *characteristic* (i.e., the mapping $\mathbb{P}_Y \mapsto \mathbb{E}[\psi(Y)]$ is injective), this is 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 ℓ_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 $\mathbb{E}[\|W\|_{\mathcal{H}_k}^2] = \|\mathbb{E}[W]\|_{\mathcal{H}_k}^2 + \mathrm{Tr}[\mathrm{Cov}_{\mathcal{H}_k}[W]]$ for any RKHS-valued random variable $W \in \mathcal{H}_k$ with $\mathbb{E}\|W\|^2 < \infty$ (Berlinet and Thomas-Agnan, 2011), where

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

is the covariance operator, $B_1(\mathcal{H}_k)$ is the set of trace-class operators on $\mathcal{H}_k$, and $\mathrm{Tr}\colon B_1(\mathcal{H}_k) \to \mathbb{R}$ is the standard trace functional. Applying this identity to the norm inside the outer ex-

pectation in (20) with $W = \psi(Y) - \mathbb{E}[\psi(T_{X,X'}(Y')|X = x, X' = x']$, we get,

$$\tilde{\ell}(T) := \ell_0(T) - \mathrm{Tr}[\mathrm{Cov}[\psi(Y)]] = \mathbb{E}_{(X',X,Y)\sim\mathbb{P}_X\otimes\mathbb{P}_{X,Y}} \|\psi(Y) - \mathbb{E}[\psi(T_{X,X'}(Y'))|X, X']\|^2_{\mathcal{H}} \quad (21)$$

Note that minimizing $\tilde{\ell}$ is equivalent to minimizing ℓ_0, so we can simply work with $\tilde{\ell}$ instead. This removes one of the two conditional expectations in (20). Now, to remove the other conditional expectation, we pass the expectation over X' *inside* the norm, yielding

$$\ell(T) = \mathbb{E}_{(X,Y)\sim\mathbb{P}_{X,Y}} \|\psi(Y) - \mathbb{E}[\psi(T_{X,X'}(Y'))|X]\|^2_{\mathcal{H}_k} . \quad (22)$$

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,

$$\ell(T) = \mathbb{E}_{(X,Y),(X',Y'),(X'',Y'')\sim\mathbb{P}_{X,Y}}\Big(k(Y,Y) + k(T_{X,X'}(Y'), T_{X,X''}(Y'')) - 2k(Y, T_{X,X'}(Y'))\Big) \quad (23)$$

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 ℓ 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 $\mathcal{T}_{\mathbb{G}} := \{T_f : f \in \mathcal{F}_{\mathbb{G}}\}$ be the set of cocycles constructed from coboundary maps in $\mathcal{F}_{\mathbb{G}}$. Then, any $T^* \in \mathrm{arginf}_{T\in\mathcal{T}_{\mathbb{G}}}\ell(T)$ satisfies (ID), (PI) and (DA) w.r.t. $\mathbb{P}_{Y|X}$, $(\mathbb{P}_X \otimes \mathbb{P}_X)$-a.s.*

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 $\mathcal{F} \subset \mathcal{F}_{\mathbb{G}_{\mathrm{TMI}}}$ (or in any other situation in which $\mathcal{F}$ is identifiable), the minimizer of ℓ identifies the counterfactual cocycle T^*. Going forward, we refer to the objective (23) 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 $\widehat{\psi(Y(x^{(i)}))} := \frac{1}{n}\sum_{j\neq i}\psi(\hat{T}_{x^{(i)},x^{(j)}}(Y(x^{(j)}))$.

Tractable Empirical Analogues The only expectations in (23) 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 ℓ (dropping all terms independent of T):

$$\ell_n^V(T) = -\frac{2}{n^2}\sum_{i,j}^n k(Y^{(i)}, Y_T^{(i,j)}) + \frac{1}{n^3}\sum_{i,j,k}^n k(Y_T^{(i,j)}, Y_T^{(i,k)}) \quad (24)$$

$$\ell_n^U(T) = -\frac{2}{n(n-1)}\sum_{i\neq j}^n k(Y^{(i)}, Y_T^{(i,j)}) + \frac{1}{n(n-1)(n-2)}\sum_{i\neq j\neq k}^n k(Y_T^{(i,j)}, Y_T^{(i,k)}) , \quad (25)$$

where $Y_T^{(i,j)} := T_{X^{(i)},X^{(j)}}(Y^{(j)})$. Both loss functions ℓ_n^V and ℓ_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 ℓ_n^V, ℓ_n^U converge to ℓ 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 in set $\mathcal{T}$, we have $\ell_n^V(T) = \ell(T) + c + \mathcal{O}_P(n^{-\frac{1}{2}})$, where $c \in \mathbb{R}$ does not depend on T. The same holds for ℓ_n^U.*

We now 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: Θ 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 (25). By standard theory (e.g., Theorem 5.2.9 in De la Pena and Giné (2012)) analogous results hold for the V-statistic (24).

Theorem 10 (CMMD Strong Consistency)**.** *Let $\{(X^{(i)}, Y^{(i)})\}_{i=1}^n \sim_{iid} \mathbb{P}_{X,Y}$ and $\hat{\theta}_n \in \operatorname{argmin}_{\theta\in\Theta} \ell_n^U(\theta)$. If Assumption 4 holds, then $\inf_{\theta\in M} \|\hat{\theta}_n - \theta\|_2 \xrightarrow{n\to\infty} 0$ a.s. Moreover, $\exists\, \theta_0 \in M$ such that $T_{\theta_n} \overset{n\to\infty}{\to} T_{\theta_0}$ 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 be 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 *θ_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 $L_T : \mathbb{X}^2 \times \mathbb{T} \to \mathbb{R}_{>0}$ with $\mathbb{E}[L_T(X, X', Y')^2] < \infty$, such that for all $\theta, \theta' \in \Theta$,*

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

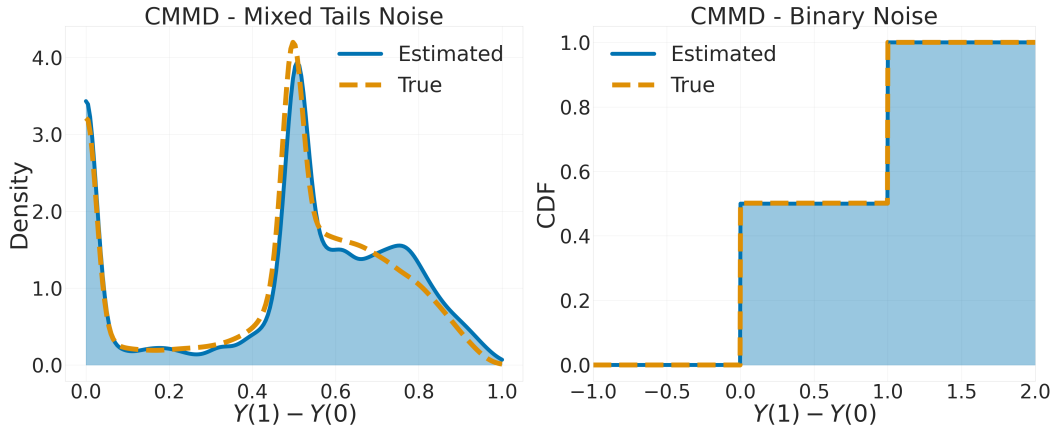


Figure 6: Estimated treatment effect distributions for Example 1 using flow-based cocycles with CMMD. Left: density under mixed-tailed noise. Right: CDF under discrete noise. Cocycles accurately recover both, unlike flow-based BCMs in Fig. 2.

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* $\hat{\theta}_n \in \arg\min_{\theta \in \Theta} \ell_n^U(\theta)$. *Then, under Assumption 4 and Assumption 5, we have* $\inf_{\theta \in M} \|\hat{\theta}_n - \theta\|_2 = O_p\big(n^{-1/2}\big)$.

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 θ_0 yields the gradient bound $\|\nabla \ell(\theta)\|_2 \geq \lambda_{\min} \|\theta - \theta_0\|_2$ for θ near θ_0. We assume the bound directly as it requires only first-order derivatives and is more natural when there are multiple minimizers.

Demonstration We conclude by demonstrating the performance of our CMMD estimator 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 architectures as the flow-based BCMs in that example. We train each model using the CMMD loss, choosing the best architecture by 2-fold cross-validation. We 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 (see (16) and (17)). CMMD implementation details are in Section 7 and the optimization routine in Section 8.1. Fig. 6 shows the estimated treatment effect density for the mixed-tails noise case (left) and CDF for the binary noise case (right). In contrast to the poor performance of flow-based BCMs (see Fig. 2), our approach accurately recovers both distributions. This reflects that (a) our procedure does not require specifying a latent noise distribution, (b) the cocycle is well-specified by a simpler (MAF) flow architecture, and (c) the CMMD estimator is robust to the (true) noise distribution.

6 Robustness and Simplicity of Counterfactual Cocycles Versus SCMs

In this section we analyze the advantages of modeling counterfactual cocycles in contrast to SCMs. In short, because a single cocycle corresponds to an entire equivalence class of SCMs, centering the modeling and estimation process around it makes the resulting procedure more robust to mis-specification, and can greatly simplify estimation. In many cases, this lets us sidestep the problems with flow-based BCMs 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

$$Y = f(X, \xi), \quad \xi \perp\!\!\!\perp X,$$

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. (10) and (14)), when the treatments $X \in \boldsymbol{V}$ 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 automorphisms of the noise (Nasr-Esfahany et al., 2023). In particular, let $Y = f(X, \xi)$, where $\xi \in \mathcal{E}$ and $f_x := f(x, \bullet) : \mathcal{E} \to \mathbb{Y}$ is bijective. Then, for any $g \in \mathrm{Aut}(\mathcal{E})$ (i.e., the group of bijections on $\mathcal{E}$),

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

Letting $f^{(g)}(x, \xi) := f(x, g(\xi))$, it is clear that $\{(f^{(g)}, g^{-1}_{\#}\mathbb{P}_\xi)\}_{g \in \mathrm{Aut}(\mathcal{E})}$ 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 does not, since

$$T_{x,x'} = f_x \circ f^+_{x'} = f_x \circ g \circ g^{-1} \circ f^+_{x'} \,, \qquad \text{for any } g \in \mathrm{Aut}(\mathcal{E}),\ x, x' \in \mathbb{X}. \tag{26}$$

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

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

On the other hand, the cocycle approach specifies a function class of the same type, $\mathcal{F} := \{f : \mathbb{X} \times \mathcal{E} \to \mathbb{Y}\}$, whose elements will be used to construct candidate cocycles as $T_{x,x'} = f_x \circ f^+_{x'}$. The model $\mathcal{F}$ is well-specified if there is some $f \in \mathcal{F}$ such that $f_x \circ f^+_{x'} = T^*_{x,x'} = f^*_x \circ f^{*+}_{x'}$. By (26), it is easy to see that this is true if and only if:

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

Since $\mathcal{T}(\mathbb{P}^*_\xi, \hat{\mathbb{P}}_\xi)$ is a strict subset of $\mathrm{Aut}(\mathcal{E})$, flow-based models with fixed $\hat{\mathbb{P}}_\xi$ 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

$$f^*(X, \xi^*) = AX + \xi^*, \quad \xi^* \sim \mathsf{t}_3(0,1) \otimes \left(\frac{1}{2}\mathsf{U}\left(-\tfrac{3}{2}, -\tfrac{1}{2}\right) + \frac{1}{2}\mathsf{U}\left(\tfrac{1}{2}, \tfrac{3}{2}\right) \right) .$$

Suppose we model this mechanism using base distribution $\hat{\mathbb{P}}_\xi = \mathsf{N}(0, I)$ and a class of autoregressive Lipschitz diffeomorphisms $\mathcal{F} \subset \mathcal{F}_{\mathbb{G}_{\mathrm{TMI}}}$, as in SOTA flow-based BCMs. 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 $h \in \mathcal{T}(\mathbb{P}^*_\xi, \hat{\mathbb{P}}_\xi))$ that results in $f^{*(h)} \in \mathcal{F}_{TMI}$ 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 $\hat{\mathbb{P}}_{\xi_1}$ to $\mathbb{P}^*_{\xi_1}$ and $\hat{\mathbb{P}}_{\xi_2}$ to $\mathbb{P}^*_{\xi_2}$ respectively:

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

Here I is the regularized incomplete Beta function and Φ is the CDF of $\mathsf{N}(0,1)$. 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 $\hat{\mathbb{P}}_\xi$ and can be modeled using the set of linear maps $\mathcal{F}_{\mathrm{LIN}} = \{f(x, \xi) = Ax + \xi \mid A \in \mathbb{R}^2\} \subset \mathcal{F}$, since $f^{*(h)} \in \mathcal{F}_{\mathrm{LIN}}$ for $h = \mathrm{id} \in \mathrm{Aut}(\mathcal{E})$. This is illustrated on Fig. 7.

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 $\hat{\mathbb{P}}_\xi \in \{\mathsf{N}(0,1), \mathit{Lap}(0,1)\}$ 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 ξ 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:

$$\xi^*_1 =_{a.s.} \xi^*_2 \sim \mathsf{N}(0,1) .$$

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

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.

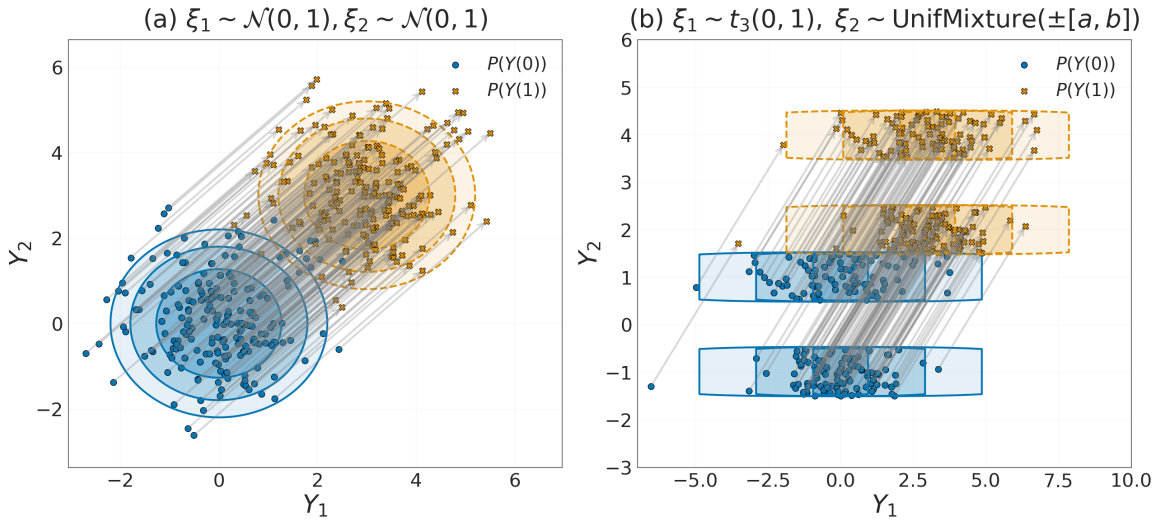


Figure 7: Counterfactual transport (gray arrows) $T_{1,0}$ from samples of $\mathbb{P}_{Y(0)}$ (blue) to $\mathbb{P}_{Y(1)}$ (orange) for structural model $Y(x) = Ax + \xi$ where (i) $\xi \sim \mathsf{N}(0, I)$ (left) and (ii) $(\xi) \sim \mathsf{t}_3(0,1) \otimes \left(\frac{1}{2}\mathsf{U}\left(-\frac{3}{2}, -\frac{1}{2}\right) + \frac{1}{2}\mathsf{U}\left(\frac{3}{2}, \frac{1}{2}\right)\right)$ (right). If $\hat{\mathbb{P}}_\xi = \mathsf{N}(0, I)$ is used as a base distribution to learn the BGM $(f, \hat{\mathbb{P}}_\xi)$, there is no continuous bijection $\hat{\mathbb{P}}_\xi \mapsto \mathbb{P}_{Y(x)}$ in case (ii). In contrast, the transport $T_{1,0}(y) = A + y$ is a simple linear map and remains invariant to the true noise distribution $\mathbb{P}_\xi$.

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 $\{f^{(g)}\}_{g \in \mathrm{Aut}(\mathcal{E})}$. We are therefore free to choose a representative $f^\star$ that minimizes a given notion of functional complexity (e.g., smoothness). 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 $\mathbb{P}^\star$.

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[5], so that each f_x is bijective $\mathbb{Y} \to \mathbb{Y}$.

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 be well-specified for f, when the dependence of the function f on x is not known. Note that size of this model class $\mathcal{F}_{\mathbb{G}_f}$ is controlled by the size of $\mathbb{G}_f$: if two coboundary maps f_1 and f_2 generate groups $\mathbb{G}_{f_1}$ and $\mathbb{G}_{f_2}$ with $\mathbb{G}_{f_1} \subsetneq \mathbb{G}_{f_2}$, then $\mathcal{F}_{\mathbb{G}_{f_1}} \subsetneq \mathcal{F}_{\mathbb{G}_{f_2}}$. Altogether, this implies that if $\mathbb{G}_{f_1} \subsetneq \mathbb{G}_{f_2}$, 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 $\mathbb{G}_{f_1} := \mathrm{Diff}^k(\mathbb{Y})$, the set of diffeomorphisms on $\mathbb{Y}$ with k^{th} derivatives continuous, and $\mathbb{G}_{f_2} := \mathrm{Diff}^1(\mathbb{Y})$. Then, $\mathcal{F}_{\mathbb{G}_{f_1}}$ contains only those functions in $\mathcal{F}_{\mathbb{G}_{f_2}}$ that are at least k-smooth.

Below we show that, out of all equivalence class members $(f^{(g)})_{g\in\mathrm{Aut}(\mathbb{Y})}$ that can be used to construct a given counterfactual cocycle, the construction $f^\star_x := T_{x,x_0}$ from Theorem 3 induces the smallest possible set $\mathcal{F}_{\mathbb{G}_{f^\star}}$ and so can be modeled using a simpler model 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 $f_x : \mathbb{Y} \to \mathbb{Y}$ is bijective. Let $\mathbb{G}_f = \langle f_x : x \in \mathbb{X}\rangle$ and $\mathbb{G}_{f^{(g)}} = \langle f_x \circ g : x \in \mathbb{X}\rangle$, for any $g \in \mathrm{Aut}(\mathbb{Y})$. Then,*

$$(i)\ \mathcal{F}_{\mathbb{G}_f} \subseteq \mathcal{F}_{\mathbb{G}_{f^{(g)}}} \quad \text{and} \quad (ii)\ \mathcal{F}_{\mathbb{G}_f} \subsetneq \mathcal{F}_{\mathbb{G}_{f^{(g)}}}\ \forall g \notin \mathbb{G}_f\ .$$

Since choosing $f^\star_x := T_{x,x_0}$ corresponds to the base distribution $\hat{\mathbb{P}}^\star_\xi := \mathbb{P}_{Y|X=x_0}$ (i.e., $(f^\star_x)_\# \mathbb{P}_{Y|X=x_0} = \mathbb{P}_{Y|X=x}$) 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 $g \in \mathbb{G}_{f^\star}$ must require a more complex coboundary map than can be used to construct the cocycle.

As an illustration of this result, in Example 3, we saw that $f^*_x \in (\mathbb{R}^2, +)$, the group of shifts on $\mathbb{R}^2$. Since in that case, $\mathbb{P}^*_\xi = \mathbb{P}_{Y|X=0}$, 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 shift of $\mathbb{P}_{Y|X=0}$ then the structural map from this base distribution will necessarily be more complex.

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

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

(e.g., those in (15)-(17)). Suppose the true cocycle T can be constructed by a flow parameterized on $\mathbb{R}^d$, $f_{\theta_0} \in \mathcal{F} = \{f_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$. Now, let $\hat{\theta}$ be *any* estimator of θ_0 from

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

$\{X^{(i)}, Y^{(i)}\}_{i=1}^n$. Our approach uses the cocycle to empirically estimate the causal estimand via the imputed counterfactuals

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

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

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

Although one can appeal to equivalent empirical process conditions for this estimator, one can still have $\hat{\gamma}^{AAP}(x) \not\to_p \gamma(x)$ even if the estimator for the cocycle is consistent, $\hat{\theta} \to_p \theta_0$. 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 $\hat{\theta} \to_p \theta_0$ 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

As mentioned in Section 4, alternative SCM-based approaches have been proposed in counterfactual fairness applications, combining conditional-quantile estimation techniques with causal DAGs (Plečko and Meinshausen, 2020; Machado et al., 2024). These methods start from an SCM with independent, uniform noise,

$$Y_j = f_j(X, Y_{\text{pa}(j)}, \xi_j), \quad \xi_j \sim \mathsf{U}[0,1] \quad \forall j \in \{1, \ldots, p\}\,.$$

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

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

where $F_{x, y_{\text{pa}(j)}}$ is the conditional CDF given $X = x$ and 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. In this way, each node is updated by a map of the form $T_{(x', \hat{y}_{\text{pa}(j)}),(x, y_{\text{pa}(j)})} = Q_{x', \hat{y}_{\text{pa}(j)}} \circ F_{x, y_{\text{pa}(j)}}$, which is precisely the conditional OT map between $\mathbb{P}_{Y_j|Y_{\text{pa}(j)}=y_{\text{pa}(j)},\, X=x}$ and $\mathbb{P}_{Y_j|Y_{\text{pa}(j)}=\hat{y}_{\text{pa}(j)}(x'),\, X=x'}$ (Carlier et al., 2016; Hosseini et al., 2025). The resulting procedure has been therefore recently branded sequential-OT (Machado et al., 2024), since it decomposes the transport into a collection 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 $(Q_x)_\# : \mathsf{U}[0,1]^p \mapsto \prod_{j=1}^p \mathbb{P}_{Y_j|Y_{\mathrm{pa}(j)},X=x}$, 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,

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

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

$$T^\star_{x',x}(y_1, y_2) = (y_1 + \Delta,\ y_2 + \Delta), \qquad \Delta = x' - x.$$

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

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

Combining these transports together gives the sequential-OT map

$$T^{\mathrm{seq}}_{x',x}(y_1, y_2) = \big(y_1 + \Delta,\ y_2 + (1 + \tfrac{\rho}{2})\Delta\big),$$

which coincides with $T^\star_{x',x}$ 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 ξ_1, ξ_2 are dependent. By targeting the joint cocycle directly, our method avoids this issue and can recover the correct transport regardless of ρ, 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 Appendix 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* $\tau_\theta \colon \mathbb{X} \to \Lambda$, mapping inputs x to a vector of flow parameters λ, and *(ii) a* bijector $g_\lambda \in \mathbb{G}$ that transforms y:

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

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).

Image group $\mathbb{G}$	Conditioner output	Cocycle $T_{x',x}(y)$	Identifiability restriction
Shifts $(\mathbb{R}^d, +)$	$a_x \in \mathbb{R}^d$	$y + a_{x'} - a_x$	None
$\mathrm{GL}_+(\mathbb{R}^d) \cap \mathbb{G}_{\mathrm{TMI}}$	$A_x \in \mathrm{GL}_+(\mathbb{R}^d)$	$A_{x'} A_x^{-1} y$	A_x lower-triangular
$\mathrm{GA}_+(\mathbb{R}^d) \cap \mathbb{G}_{\mathrm{TMI}}$	(A_x, a_x)	$a_{x'} + A_{x'} A_x^{-1}(y - a_x)$	A_x lower-triangular
$\mathrm{Diffeo}(\mathbb{R}^d) \cap \mathbb{G}_{\mathrm{TMI}}$	MAF parameters θ_x	$\mathrm{MAF}^{-1}[\theta_{x'}] \circ \mathrm{MAF}[\theta_x](y)$	None
$\mathrm{Diffeo}(\mathbb{R}^d) \cap \mathbb{G}_{\mathrm{TMI}}$	NSF parameters θ_x	$\mathrm{NSF}^{-1}[\theta_{x'}] \circ \mathrm{NSF}[\theta_x](y)$	None

The conditioner τ_θ can be *any* learnable function class—linear model, MLP, convolutional network, or transformer—so long as it maps x to valid flow parameters λ. The bijector g_λ can be a single transform or a multi-layer composition of such transforms that lie in $\mathbb{G}_{\mathrm{TMI}}$. 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 specify the inverse map f_x^{-1} as the (forward) autoregressive flow with a masked adjacency matrix. This prevents spurious correlations from being induced by the architecture (see Javaloy et al. (2023) for additional details).

7.2 Optimization and Model Selection

CMMD Implementation We optimize all flow-based cocycles using gradient descent on our empirical CMMD losses (V/U-statistic) introduced in Section 3. Our default kernel choice for CMMD is the Gaussian kernel $k(y, y') = \exp(-\lambda \|y - y'\|^2)$, where the bandwidth λ 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 ℓ_n^U without bias and ℓ_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 Appendix 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 Appendix 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 recovering counterfactual couplings. Code 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 equation model $Y = \beta X + \xi$ under different settings of $\mathbb{P}_\xi$. We implement cocycles trained using both CMMD-V and CMMD-U losses, and benchmark them against flow-based SCMs which estimate the bijective generative mechanism $(f, \mathbb{P}_\xi)$ via maximum likelihood, using different base distributions ($\mathsf{N}(\mu, \sigma^2)$, $\mathsf{Lap}(\mu, \sigma^2)$, $\mathsf{t}_\nu(\mu, \sigma^2)$) with learnable parameters. For all methods, we perform 2-fold cross-validation across a range of flow architectures of increasing complexity: (i) a linear flow $f_x(\xi) = \theta x + \xi$, (ii) an additive flow $f_x(\xi) = m_\theta(x) + \xi$, (iii) masked autoregressive flow (Papamakarios et al., 2017) and (iv) a neural spline flow (Durkan et al., 2019). All flows except (i) are parameterized using MLP neural networks with 2 layers and 32 hidden nodes per layer. Each model is trained for 1000 epochs using the ADAM optimizer in PyTorch with default hyperparameters and a learning rate 0.01. All results are averaged over 50 random seeds.

Results Table 3 shows performance results for different true noise distributions under a hard intervention $do(X = 0)$. 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 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 $\mathbb{P}_\xi = \mathsf{Rad}(1/2)$, 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 $\mathbb{P}_\xi \in \{\mathsf{t}_1(0,1), \mathsf{IG}(1,1), \mathsf{Rad}(1/2)\}$, reflecting the fact that the base distributions either have mis-specified

Method	$\mathbb{P}_\xi = \mathsf{N}(0,1)$	$\mathbb{P}_\xi = \mathsf{Ga}(1,1)$	$\mathbb{P}_\xi = \mathsf{t}_1(0,1)$	$\mathbb{P}_\xi = \mathsf{IG}(1,1)$	$\mathbb{P}_\xi = \mathsf{Rad}(1/2)$
			Interventional KS		
ML-G	$\mathbf{0.015 \pm 0.007}$	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
CMMD-V	0.028 ± 0.009	$\mathbf{0.027 \pm 0.008}$	$\mathbf{0.027 \pm 0.009}$	$\mathbf{0.027 \pm 0.008}$	$\mathbf{0.271 \pm 0.031}$
CMMD-U	$\mathbf{0.010 \pm 0.004}$	$\mathbf{0.012 \pm 0.005}$	$\mathbf{0.008 \pm 0.003}$	$\mathbf{0.009 \pm 0.004}$	$\mathbf{0.268 \pm 0.008}$
			Counterfactual RMSE		
ML-G	$\mathbf{0.035 \pm 0.035}$	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	$\mathbf{0.033 \pm 0.031}$	0.270 ± 0.097	0.044 ± 0.053	29.872 ± 41.628	0.391 ± 0.307
CMMD-V	$\mathbf{0.035 \pm 0.026}$	$\mathbf{0.020 \pm 0.015}$	$\mathbf{0.040 \pm 0.031}$	$\mathbf{0.028 \pm 0.024}$	$\mathbf{0.017 \pm 0.019}$
CMMD-U	0.040 ± 0.027	$\mathbf{0.022 \pm 0.016}$	$\mathbf{0.033 \pm 0.027}$	$\mathbf{0.027 \pm 0.023}$	$\mathbf{0.014 \pm 0.011}$
			True Architecture Selection %		
ML-G	96%	14%	0%	2%	2%
ML-L	**100%**	2%	4%	0%	0%
ML-T	98%	8%	94%	0%	0%
CMMD-V	**100%**	**100%**	**100%**	**100%**	**98%**
CMMD-U	**100%**	**100%**	**100%**	**100%**	**100%**

Table 3: Mean ± 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 $do(X = 0)$ across different noise laws. "ML-" denotes ML flows with Gaussian (G), Laplace (L) or Student-t (T) bases; "CMMD-" denotes our cocycle estimators (CMMD-V/U). Boldface marks the top two performers per column.

tails or support in these cases (except the Student's-t base in the $\mathsf{t}_1(0,1)$ noise case). For $\mathbb{P}_\xi = \mathsf{Ga}(1,1)$, 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 CMMD estimator and that we do not need a more complex flow to compensate for a poorly matched base distribution.

Extension To analyze how the CMMD estimator performs when all architectures are fixed, in Figure 8 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 CMMD estimator against ML estimators with the Gaussian and Laplace base distributions (i.e. ℓ_2 and ℓ_1 regression), as well as a recently proposed MMD-based estimator for conditional generative 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

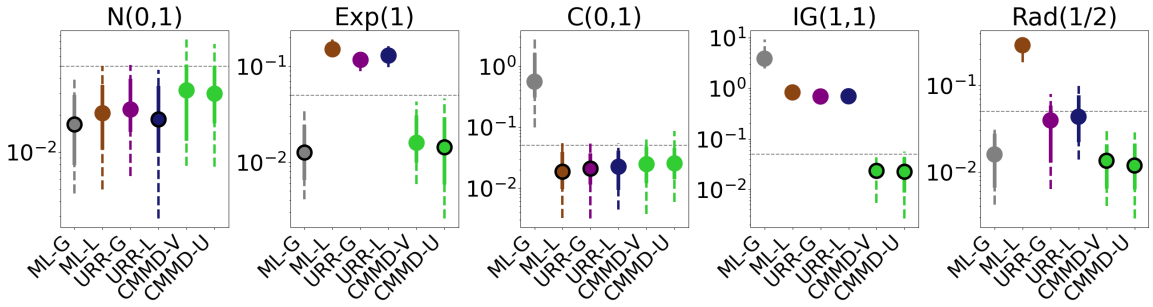


Figure 8: Mean (circle), (25-75) percentiles (solid line) and (10-90) percentiles (dashed line) of absolute error in treatment effect $Y(X+1)-Y(X)=1$ for different estimation methods in a linear model $Y=X+\xi$ under different noise designs, when fixing the flow to be the true architecture $f_x(\xi)=\beta x+\xi$. Dashed horizontal line = 5% error. Black edges = best two methods on average.

Eq. (5) from (Alquier and Gerber, 2023), using $m=n$ Monte Carlo samples from $\hat{\mathbb{P}}^{\theta}_{Y|X}$. The kernel is chosen identically to CMMD. Both CMMD estimators estimate β 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 $(\xi_1,\xi_2)\sim\mathsf{Lap}(\mathbf{0},\rho\,\mathbf{1}\mathbf{1}^{\mathsf{T}}+(1-\rho)I)$; and (ii) a non-additive SCM $\boldsymbol{Y}=\mathbf{1}X+L(X)\boldsymbol{\xi}$ with $\boldsymbol{\xi}\sim\mathsf{Lap}(0,I_2)$ 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 masked autoregressive flow (Papamakarios et al., 2017), with a two-layer MLP conditioner (32 nodes per layer) for the dependence $Y_1\to Y_2$. We use a

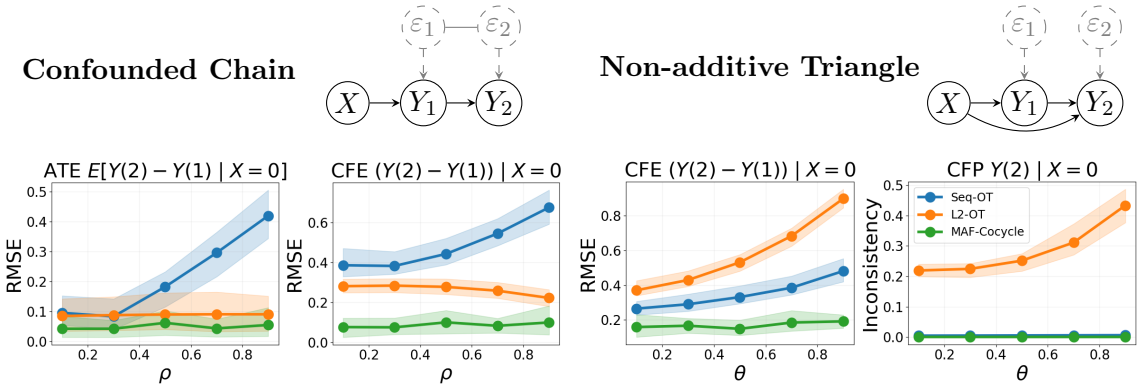


Figure 9: Left and middle left (Confounded Chain): Mean ± SD (20 trials) of error in estimated average incremental effect (left) and incremental counterfactual effect (middle left) of second treatment, for observed units $\{Y^{(i)}(0)\}_{i=1}^{n}$ in control group. blue = sequential optimal transport (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 $\rho := \mathrm{Corr}(\xi_1, \xi_2)$ increases, whereas the MAF-cocycle does not, due to the robustness of our estimation procedure. Right and middle right (Non-additive Triangle): Mean ± SD (20 trials) error in estimated incremental counterfactual effect (middle right) and inconsistency in counterfactual predictions $Y(2)$ when imputing (right) 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. θ controls degree of non-additivity in true transport maps. As θ increases, OT performance degrades and transport inconsistency increases, reflecting underlying model incoherence.

separate flow per treatment level $x \in \{0, 1, 2\}$, i.e., $f_x := \mathrm{MAF}[\theta_x]$ with $\theta_x = (\theta_0, \theta_1, \theta_2)[x]$. We optimize the cocycle to target (DA) w.r.t. $\mathbb{P}_{Y_1, Y_2 \mid X}$, assuming the causal ordering $X \prec Y_1 \prec Y_2$. Training details are as in 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), which factorizes the joint transport into a set of (1D) conditional transports 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 (ii)).

Results For design (i), Fig. 9 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 ρ 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., $(\hat{Y}(1),\hat{Y}(2)) = \big(T_{1,0}(Y(0)), T_{2,0}(Y(0))\big)$.

For design (ii), Fig. 9 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 θ 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 considered 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 BCMs.

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 $\xi_1 = \mathsf{N}(0,1)$ and $\xi_{d:2} = \{\mathsf{IG}(1,1), \mathsf{Rad}(1/2), \frac{1}{2}\mathsf{N}(-\sqrt{3}/2, 1/2) + \frac{1}{2}\mathsf{N}(\sqrt{3}/2, 1/2), \mathsf{Ga}(1,1)\}$.

We implement flow-based cocycles on $(X,Y) = (V_1, V_{>1})$ with the CMMD-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 $\hat{\mathbb{P}}_X$. 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 Sec-

tion 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 ± SD of $\mathrm{KS}_{\mathrm{int}}$ and $\mathrm{KS}_{\mathrm{CF}}$ on the *linear* SCMs.

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

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

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

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 $do(V_1 = 0)$, by computing the *average marginal Kolmogorov-Smirnov (KS) distance*:

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

Here $\hat{F}$ is the CDF, which for flow-based cocycles is estimated empirically on counterfactual samples, and for baselines is estimated via the abduct-act-predict procedure (see Section 2.1) Tables 4 and 5 report the mean ± SD results from 10 trials. Cocycles achieves the best $\mathrm{KS}_{\mathrm{int}}$ and $\mathrm{KS}_{\mathrm{CF}}$ in all SCMs except one (`Triangle-nonlin` for $\mathrm{KS}_{\mathrm{CF}}$ and `Fork-nonlin` for $\mathrm{KS}_{\mathrm{int}}$). The performance gap is generally greatest when the true DGP is linear, reflecting our method's ability to be well-specified with simpler selected architectures. Out of the baselines, CAREFL generally performed worst, which likely reflects its restriction to affine flows. BGM generally performed best out of the baselines. This is to be expected, as under the intervention $do(V_1 = 0)$, the only part of the flow used at test time is the conditional component $f : (x, y) \mapsto f_x(y)$ on $Y := V_{>1}$, which is directly optimized by BGM.

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

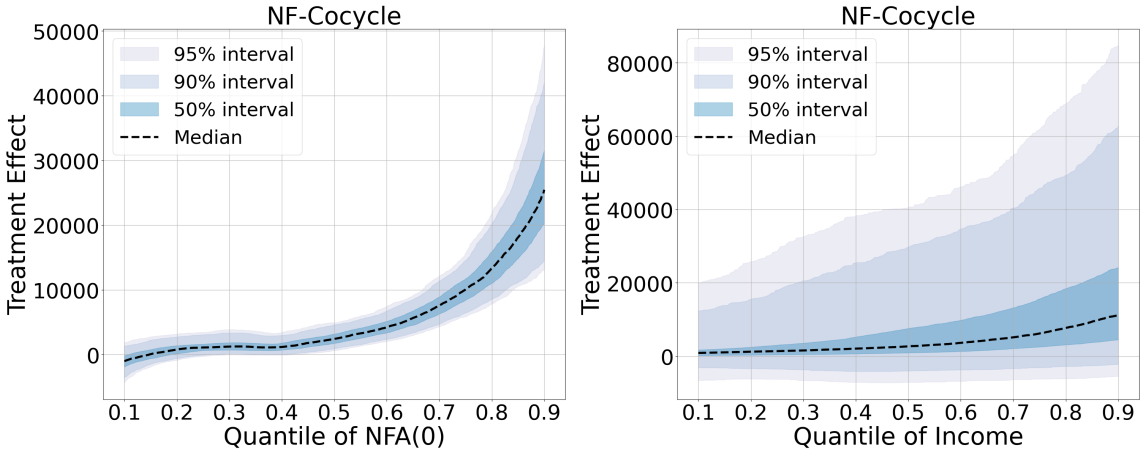


Figure 10: Estimated conditional distribution of the treatment effect $Y(1) - Y(0)$ (i.e., change in net financial assets) given quantiles of (i) $Y(0)$ (i.e., net financial assets under the no treatment scenario) and (ii) income, for different cocycle models. NF-Cocycle = normalizing flow based cocycle. Mean = conditional mean, and $\alpha\%$ interval = middle $\alpha\%$ of the conditional distribution.

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 the full distribution 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 $Y(0)$. The conditional distribution $\mathbb{P}(Y(1) - Y(0)|I)$ 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. The conditional distribution $\mathbb{P}(Y(1) - Y(0)|Y(0))$ will help us to determine the extent to which those who benefit most are those who were better off already without the scheme.

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, cross-validating over different architecture design choices and using the same training settings as in Section 8.1. We use the estimated cocycle to impute the counterfactual outcomes $\hat{Y}(1) := \hat{Y}(1, Z), \hat{Y}(0) := \hat{Y}(0, Z)$ and use them to estimate the required conditional distributions using nonparametric smoothing, as described in Section 4.3. We use the Nadaraya–Watson estimator with a Gaussian kernel for the smoothing weights. The kernel bandwidths are learned by performing gradient-descent on the K-fold cross-validation loss from regressing the dependent vector $\{\mathbf{1}(\hat{Y}(1) - \hat{Y}(0)) \geq t_i : i \in [m]\}$ on $\hat{Y}(0)$. Here t_i is chosen as the i/m empirical quantile of $\hat{Y}(1) - \hat{Y}(0)$. We use the ADAM optimizer with a learning rate of 0.01 and 1000 gradient steps.

The estimated conditional distributions of the treatment effects are displayed in Fig. 10. Both distributions are consistent with the "rich get richer" phenomenon. In particular, the effect is 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 much larger treatment effect variance at each income quantile.

The fact that the treatment effect $Y(1)-Y(0)$ is on average increasing over the quantile τ of $Y(0)$ implies that the effect of treatment on the quantile, $ETQ(\tau) = Q_{Y(1)}(\tau)-Q_{Y(0)}(\tau)$ is also increasing over τ. Increasing profiles of the latter have been reported in previous studies (Belloni et al., 2017; Chernozhukov and Hansen, 2004), supporting our findings.

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-based approaches, our key idea was to model a set of transports between counterfactuals that satisfy the necessary 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).

We showed that any counterfactual 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.

Appendix A. Supplementary 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 $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}), \mathbb{P}_{Y|X=x_0})$. We can therefore define the random variable $Y(x_0) : \mathbb{Y} \to \mathbb{Y}$ as the identity map, which is

Borel measurable. By construction, $Y(x_0) \sim \mathbb{P}_{Y|X=x_0}$. Now, define

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

By (DA) we have $\widetilde{Y}(x) \sim \mathbb{P}_{Y|X=x}$ 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 $T_{x_0,x} \circ T_{x,x_0} = \mathrm{id}$, $\mathbb{P}_{Y|X=x_0}$-a.s. This means that $T_{x_0,x}\big(\widetilde{Y}(x)\big) =_{\text{a.s.}} Y(x_0)$, $\forall x \in \mathbb{X}$. Applying T_{x',x_0} to both sides and noting that x and x' are arbitrary gives

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

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

Proof [Theorem 3] For each $x \in \mathbb{X}$ set $\tilde{f}_x := T_{x,x_0} : \mathbb{Y} \to \mathbb{Y}$ and $\tilde{f}_x^+ := T_{x_0,x} : \mathbb{Y} \to \mathbb{Y}$, where $x_0 \in \mathbb{X}$ is an arbitrary reference point. Choose $\mathbb{Y}_0 := \mathbb{Y}_{x_0}$ as the set of full $\mathbb{P}_{Y|X=x_0}$-measure used in the definition of (ID) and (PI) and note that $\mathbb{Y}_0 \subseteq \mathbb{Y}$, as required. Let $f_x := \tilde{f}_x\big|_{\mathbb{Y}_0} : \mathbb{Y}_0 \to \mathbb{Y}$ be the restriction of f_x to $\mathbb{Y}_0$ and $f_x^+ := \tilde{f}_x^+\big|_{\mathrm{Im}(f_x)}$ be the restriction of $\tilde{f}_x^+$ to the image of the restriction of $\tilde{f}_x$. Then, f_x^+ is a left inverse of $f_x : \mathbb{Y}_0 \to \mathbb{Y}$ because by (ID) we have $f_x^+ \circ f_x(y) = T_{x_0,x} \circ T_{x,x_0}(y) = y$ for every $y \in \mathbb{Y}_0$. To conclude the proof note that for any $x, x' \in \mathbb{X}$, we have by (PI):

$$f_x \circ f_{x'}^+(y) = T_{x,x_0} \circ T_{x_0,x'}(y) = T_{x,x'}(y) \quad \forall y \in \mathbb{Y}_{x'}$$

■

Proof [Theorem 5] By Theorem 3, any $\mathbb{P}_{Y|X}$-adapted, $\mathbb{G}$-valued cocycle can be written as $T_{x,x'} = f_x \circ f_{x'}^{-1}$, where $f_x := T_{x,x_0}$ for arbitrary $x_0 \in \mathbb{X}$ and the exact inverse reflects that $f_x \in \mathbb{G}$, a group of transformations $\mathbb{Y} \to \mathbb{Y}$. Now, by (DA), this immediately implies $(f_x^{-1})_{\#}\mathbb{P}_{Y|X=x} = \mathbb{P}_{Y|X=x_0}$. The choice of x_0 is arbitrary, with a different choice of x_0 leading to a different coboundary map but the same cocycle.

Now, assume that $T_{x,x'} = f_x \circ f_{x'}^{-1}$ and $\tilde{T}_{x,x'} = \tilde{f}_x \circ \tilde{f}_{x'}$ are two $\mathbb{P}_{Y|X}$-adapted and $\mathbb{G}$-valued cocycles. Since both cocycles are $\mathbb{P}_{Y|X}$-adapted,

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

Hence, $b_x := \tilde{f}_x^{-1} \circ f_x \in \mathrm{Aut}(\mathbb{P}_{Y|X=x_0})|_{\mathbb{G}}$ and likewise for $b_x^{-1} = f_x^{-1} \circ \tilde{f}_x$, which satisfy (9) for all $x, x' \in \mathbb{X}$. Conversely, if $T_{x,x'} = f_x \circ f_{x'}^{-1}$ is a $\mathbb{G}$-valued cocycle adapted to $\mathbb{P}_{Y|X}$, then it is easy to check that $\tilde{T}_{x,x'} = \tilde{f}_x \circ \tilde{f}_{x'}^{-1}$ as defined in (9) satisfies (ID), (PI), and (DA). ■

Proof [Theorem 6] By Theorem 5, it suffices to prove that $\mathrm{Aut}(\mathbb{P}_{Y|X=x_0})|_{\mathbb{G}_{\mathrm{TMI}}} \subseteq [\mathrm{id}]_{\mathbb{P}_{Y|X=x_0}}$ 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 ν be a Borel probability measure on $\mathbb{R}$. If $S : \mathbb{R} \to \mathbb{R}$ is Borel, non-decreasing and satisfies $S_{\#}\nu = \nu$, then $S(y) = y$, ν-almost surely.*

Proof [Lemma 14] Write F for the distribution function of ν. Since S is monotone increasing and preserves ν, we know by Portmanteau's Lemma (Van der Vaart, 2000) 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 $y \in \mathbb{R}$:

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

Where the second inequality follows from Portmanteau's 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 $Q \circ F = \mathrm{id}$ ν-a.s. Since S preserves the null-sets of ν by definition, this means that $Q \circ F \circ S = S$ ν-a.s. Since $Q \circ F \circ S = Q \circ F = \mathrm{id}$ ν-a.s., this immediately implies the result. ■

Now we use the Lemma to prove the main result by induction. Let $T = (T_1, \ldots, T_p) \in \mathbb{G}_{\mathrm{TMI}}$ satisfy $T_{\#}P = P$. We show $T_k(y) = y_k$ P-a.s. for $k = 1, \ldots, p$. 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., $\pi_{1:k}(y) = (y_1, \ldots, y_k)$). 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, $T = \mathrm{id}$ P-a.s. Therefore, any $T \in \mathbb{G}_{\mathrm{TMI}}$ with $T_{\#}P = P$ belongs to the P-a.s. equivalence class of the identity, so $\mathrm{Aut}_{\mathrm{TMI}}(P) \subseteq [\mathrm{id}]_P$. Because $x_0 \in \mathbb{X}$ was arbitrary, the statement holds for every conditional law $\mathbb{P}_{Y|X=x_0}$, completing the proof. ■

Proof [Theorem 7] Assume that the counterfactuals satisfy Assumption 1 and (CC) with cocycle T. By Theorem 3 there is a set of functions $(f_x)_{x \in \mathbb{X}}$ with $f_x : \mathbb{Y}_0 \to \mathbb{Y}_x$ and $\mathbb{Y}_0 \subseteq \mathbb{Y}$, and $T_{x,x'} = f_x \circ f_{x'}^{+}$ on $\mathbb{Y}_{x'}$. Therefore, fix any $x_0 \in \mathbb{X}$ and note $Y(x) = f_x \circ f_{x_0}^{+}\big(Y(x_0)\big)$ by (CC). Setting $\xi := f_{x_0}^{+}\big(Y(x_0)\big)$, we then have $Y = f(X, \xi)$ by the consistency property. Since independence is preserved under arbitrary transformations, the the exchangeability property implies that $X \perp\!\!\!\perp \xi$. Conversely, if Y is generated by an injective SCM, $Y(x') = f_{x'}(\xi)$ by definition, and so

$$Y(x) = f_x(f_{x'}^{+}(Y(x'))) =: T_{x,x'}(Y(x'))$$

This shows that (CC) is satisfied. Since $X \perp\!\!\!\perp \xi$, it holds that $Y(x) = f(x, \xi) \perp\!\!\!\perp X$. Moreover, $Y(X) = f(X, \xi)$ which verifies the consistency property. This completes the proof. ■

A.2 Proofs for Section 5

Proof [Proposition 8] First note that by the definition of T, we have that

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

for every probability measure $\mu \in \mathcal{P}(\mathbb{X})$, every measurable set $A \in \mathcal{B}(\mathbb{Y})$ (the Borel σ-algebra on $\mathbb{Y}$) and $x \in \mathbb{X}$. Therefore, since we can write

$$\ell(T) = \mathbb{E}_{x\sim\mathbb{P}_X} D\left(\mathbb{P}_{Y|X}(\bullet|x), \int_{\mathbb{X}} \mathbb{P}_{Y|X}(T_{x,x'}^{-1}\{\bullet\}|x')\mathbb{P}_X(dx')\right)^2 = 0$$

where $D = \mathrm{MMD}$, we have $T \in M := \operatorname{arginf}_{\theta\in\Theta}\ell(\theta)$ 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 $T^*_{x,x} = f^*_x \circ f^{*-1}_x$ (i.e., (ID)) and $T^*_{x,x'} = f^*_x \circ f^{*-1}_{x''} \circ f^*_{x''} \circ f^{*-1}_{x'} = f^*_x \circ f^{*-1}_{x'}$ (i.e., (PI)) for every $x, x', x'' \in \mathbb{X}$. 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 $x \in \mathbb{X}$:

$$\begin{aligned}
\mathbb{P}_{Y|X}(A|x) &= \int_{\mathbb{X}\times A} \mathbb{P}_{Y|X}(T^{*-1}_{x,x'}\{dy'\}|x')\mathbb{P}_X(dx') \\
&= \int_{\mathbb{X}\times\mathbb{Y}} \mathbf{1}\{y' \in A\}\mathbb{P}_{Y|X}(T^*_{x,x'}\{dy'\}|x')\mathbb{P}_X(dx') \\
&= \int_{\mathbb{X}\times\mathbb{Y}} \mathbf{1}\{T^*_{x,x'}(y') \in A\}\mathbb{P}_{Y|x'}(dy'|x')\mathbb{P}_X(dx') \\
&= \mathbb{E}\mathbf{1}\{T^*_{x,X}(Y) \in A\}
\end{aligned}$$

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

Proof [Proposition 9] Note that for any cocycle $T : \mathbb{X}^2 \times \mathbb{Y} \to \mathbb{Y}$, $\ell^V_n(T)$ and $\ell^U_n(T)$ 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

$$\begin{aligned}
\ell^V_n(T) &= \frac{1}{n^3}\sum_{i,j,k}^{n} h_T(Z^{(i)}, Z^{(j)}, Z^{(k)}) \\
\ell^U_n(T) &= \frac{1}{\binom{n}{3}}\sum_{i<j<k}^{n} h_T(Z^{(i)}, Z^{(j)}, Z^{(k)})
\end{aligned}$$

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 ($|k| \leq 1$), $\tilde{h}_T$ is uniformly bounded and therefore so is h_T. Therefore, under the assumption that $Z^{(1)}, \ldots, Z^{(n)} \overset{\text{iid}}{\sim} \mathbb{P}_Z$, 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 $Z, Z', Z'' \overset{\text{iid}}{\sim} \mathbb{P}_Z$ are independent copies.

All that remains is to show that $\mathbb{E}h_T(Z, Z', Z'') = \ell(T) + \beta$, where β 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 $Y_T := T_{X,X'}(Y')$.

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

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

Proof [Theorem 10] For notational convenience let $Z := (X, Y)$, $Z^{(1)}, \ldots, Z^{(n)} \overset{\text{iid}}{\sim} \mathbb{P}_Z$ 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 θ_0. Now, since $\ell_n^U(\theta)$ is an Order-3 U-statistic, it is a standard fact (e.g., 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 ℓ is continuous and M is compact. By Assumption 4.4 the kernel k is continuous and bounded, so the composite h_θ is uniformly bounded by a constant $C < \infty$ and, by Assumption 4.2, continuous in θ. Therefore, for any sequence $(\theta_n)_{n\geq 1} \in \Theta$ such that $\theta_n \to \theta$, we have point-wise convergence $h_{\theta_n}(z, z', z'') \to h_\theta(z, z', z'')$ and the uniform bound $|h_{\theta_n}(z, z', z'')| \leq C$. Dominated Convergence then gives

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

Hence ℓ is continuous on Θ and, since Θ is compact, by the Weierstrass extreme-value theorem ℓ 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 Θ and compact by the Heine–Borel Theorem. Now, for strong consistency of $\hat{\theta}_n$, we will show the following properties:

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

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

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 $d_M:\mathbb{R}^d\to[0,\infty)$ is 1-Lipschitz, hence continuous. Therefore $A_\varepsilon=d_M^{-1}([\varepsilon,\infty))$ is closed in $\mathbb{R}^d$ and is compact since Θ is compact. Now, define

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

Continuity of ℓ on compact A_ε guarantees ℓ^ε exists and, as A_ε 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

$$\inf_{\theta\in\Theta:d_M(\theta)\geq\varepsilon}\big[\ell(\theta)-\min_{\vartheta\in\Theta}\ell(\vartheta)\big]=\delta_\varepsilon>0. \tag{27}$$

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 θ, and Θ 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., see Example 19.8 in Van der Vaart (2000)). 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 holds for the U-statistic $\ell_n^U(\theta)$ (see Corollary 5.2.5 in De la Pena and Giné (2012)):

$$\sup_{\theta\in\Theta}|\ell_n(\theta)-\ell(\theta)|\to 0 \quad \text{a.s.}$$

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_ε such that, for all $n\geq N_\varepsilon$,

$$\sup_{\theta\in\Theta}\big|\ell_n(\theta)-\ell(\theta)\big|<\delta_\varepsilon/2. \tag{28}$$

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 (28) with the well-separatedness inequality (27) gives

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

whereas for any minimiser $\theta^\star\in M$ we have $\ell_n(\theta^\star)\leq\ell^\star+\delta_\varepsilon/2$ by (28), 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 $d_M(\hat{\theta}_n)\to 0$ 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 $\|\hat{\theta}_n-\eta_n\|\to 0$ almost surely. By continuity of T in θ, this implies

$$T_{\hat{\theta}_n,x,x'}(y)\to T_{\eta_n,x,x'}(y) \quad \text{a.s.} \tag{29}$$

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 (29) yields point-wise convergence $T_{\hat{\theta}_n,x,x'}(y) \to T_{\theta_0,x,x'}(y)$ on a full-measure set, which completes the proof. ■

Proof [Theorem 12] For notational convenience, let $Z := (X, Y)$, $Z^{(1)}, \ldots, Z^{(n)} \overset{\text{iid}}{\sim} \mathbb{P}_Z$ 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

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

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_θ:

$$\begin{aligned} \nabla_\theta \ell_n(\theta) - \nabla_\theta \ell(\theta) &= \frac{1}{n(n-1)(n-2)} \sum_{i \neq j \neq k} H_\theta(Z^{(i)}, Z^{(j)}, Z^{(k)}) \\ &= \tfrac{1}{\binom{n}{3}} \sum_{i<j<k} H_\theta(Z^{(i)}, Z^{(j)}, Z^{(k)}) \end{aligned} \tag{30}$$

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_θ (e.g., 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_θ 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

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

Similarly, by the Lipschitz cocycle Assumption 5.1, we have

$$\nabla_\theta T_{\theta,X^{(1)},X^{(3)}}(Y^{(3)}) \le L_T(X^{(1)},X^{(3)},X^{(3)}) \,.$$

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_θ in (30) with its representation in terms of the projections, we get the standard ANOVA decomposition

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

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

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

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

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

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 (e.g., Serfling (2009) Sec 5.1.5., pp. 178.). Additionally, since $\Theta \subset \mathbb{R}^d$ is compact, the packing number is $\mathcal{M}(\epsilon,\Theta,\|\bullet\|_2) \le D/\epsilon^d$, where $D = Diam(\Theta)$. This polynomial dependence on ϵ lets us apply Sherman's maximal inequality for degenerate U-statistics (Sherman, 1994, Corr 4):

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

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

$$\sup_{\theta\in U_\delta}\left\| \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(g_{1,\theta}(Z^{(i)}) - \mathbb{P}(g_{1,\theta})\right\| = O_p(1). \tag{D}$$

Combining (HD), (U) and (D) yields

$$\sup_{\theta\in U_\delta} \|\nabla \ell_n(\theta) - \nabla \ell(\theta)\| = O_p(n^{-1/2}). \tag{S}$$

Since $\hat\theta_n$ minimizes ℓ_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 $\|\nabla \ell(\hat\theta_n)\| \ge c\, d_{\hat\theta_n}(M)$ for all $n \ge N(\delta)$, where $N(\delta) \in \mathbb{N}$. Hence, for all such n, we have

$$d_{\hat\theta_n}(M) \le c^{-1}\|\nabla \ell(\hat\theta_n)\| = O_p(n^{-1/2}) \,.$$

■

A.3 Proofs for Section 6

Proof [Theorem 13] By definition of $\mathbb{G}_{f^{(g)}}$, one of its generators is $f_{x_0} \circ g = \mathrm{id}_{\mathbb{Y}} \circ g = g$. Hence $g, g^{-1} \in \mathbb{G}_{f^{(g)}}$. Since for each $x \in \mathbb{X}$, we have $f_x = \big(f_x \circ g\big) \circ g^{-1}$, 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, $\mathbb{G}_{f^{(g)}} \not\subset \mathbb{G}_f$. 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 $\mathcal{F}_{\mathbb{G}_{f^{(g)}}} \not\subset \mathcal{F}_{\mathbb{G}_f}$. Combining with (i) completes the proof. ■

Appendix B. Algorithms

Data: Samples $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$, cocycle T_θ, Positive-definite kernel $k(\cdot,\cdot)$, Batch size B, epochs E, learning rate η
Result: Optimized cocycle T_θ^*

for *epoch* $= 1, \ldots, E$ **do**
  $s \leftarrow 0$; `// samples counter`
  while $s < n$ **do**
    Sample without replacement a set $\mathcal{B} \subset \{1, \ldots, n\}$, $|\mathcal{B}| = B$;
    Define $X_{\mathcal{B}} \leftarrow \{x^{(i)}\}_{i \in \mathcal{B}}$, $Y_{\mathcal{B}} \leftarrow \{y^{(i)}\}_{i \in \mathcal{B}}$;
    forall $i, j \in \mathcal{B}$ **do**
      $Y_\theta^{(i,j)} \leftarrow T_{\theta, x^{(i)}, x^{(j)}}(y^{(j)})$; `// outer product of cocycle evaluations`
    end

$$\ell_b^V(\theta) \leftarrow -\frac{2}{B^2} \sum_{i,j \in \mathcal{B}} k\big(y^{(i)}, Y_\theta^{(i,j)}\big) + \frac{1}{B^3} \sum_{i,j,k \in \mathcal{B}} k\big(Y_\theta^{(i,j)}, Y_\theta^{(i,k)}\big).$$

    Compute $\nabla_\theta \ell_b^V(\theta)$ by backpropagation;
    $\theta \leftarrow \theta - \eta \nabla_\theta \ell_b^V(\theta)$;
    $s \leftarrow s + B$;
  end
end
return T_θ^*

Algorithm 1: Scalable Minibatch CMMD-V Optimization

Data: Cocycle Models $\{\mathcal{M}_m\}_{m=1}^{M}$, Dataset $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$, Folds K
Result: Best model $\mathcal{M}^*$, CV losses $\{\ell_{m,k}\}$

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 copy $\widehat{\mathcal{M}} \leftarrow \mathcal{M}_m$ on $\{i \in I_{\text{tr}}^{(k)}\}$ by minimizing CMMD;
      2.2. Evaluate $\ell_{m,k} \leftarrow \text{CMMD}\big(\widehat{\mathcal{M}}, \{i \in I_{\text{val}}^{(k)}\}\big)$;
   end
   2.3. Compute $\bar{\ell}_m \leftarrow \frac{1}{K}\sum_{k=1}^{K} \ell_{m,k}$;
end
3. $m^* \leftarrow \arg\min_m \bar{\ell}_m$;
4. Retrain $\mathcal{M}^* \leftarrow \mathcal{M}_{m^*}$ on all data by minimizing CMMD;
5. **return** $\mathcal{M}^*$, $\{\ell_{m,k}\}$

Algorithm 2: Cocycle Model Selection Procedure with 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. Gallic. 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. Kobyzev, 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, 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. Kobyzev, 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. Pawlowski, D. Coelho de Castro, and B. Glocker. 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.

A. Simon, Herbert. 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.

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.

T. Teshima, I. Ishikawa, K. Tojo, K. Oono, M. Ikeda, and M. Sugiyama. Coupling-based invertible neural networks are universal diffeomorphism approximators. *Advances in Neural Information Processing Systems*, 33:3362–3373, 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.