Flow IV: Counterfactual Inference In Nonseparable Outcome Models Using Instrumental Variables

Marc Braun MARC.BRAUN@LIU.SE
Department of Computer and Information Science, Linköping University
Jose M. Peña JOSE.M.PENA@LIU.SE
Department of Computer and Information Science, Linköping University
Adel Daoud ADEL.DAOUD@LIU.SE
Institute for Analytical Sociology, Linköping University
Chalmers University of Technology

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

To reach human level intelligence, learning algorithms need to incorporate causal reasoning. But identifying causality, and particularly counterfactual reasoning, remains elusive. In this paper, we make progress on counterfactual inference in nonseparable outcome models by utilizing instrumental variables (IVs). IVs are a classic tool for mitigating bias from unobserved confounders when estimating causal effects. While IV methods for effect estimation have been extended to nonseparable outcome models under different assumptions, existing IV approaches to counterfactual prediction typically assume one-dimensional outcomes and additive noise. In this paper, we show that under standard IV assumptions, along with the assumption that the outcome function is invertible and has a triangular structure, the treatment–outcome relationship becomes identifiable from observed data. We furthermore propose a method to learn the outcome function utilizing normalizing flows. This outcome function estimator can then be used to perform counterfactual inference. We refer to the method as *Flow IV*.

Keywords: Counterfactual Inference, Instrumental Variables

1. Introduction

Estimating causal effects is a central goal in many scientific fields, including the social and medical sciences. Unlike traditional machine learning methods, which rely on associational relationships, causal inference aims to answer questions about the effect of an intervention, such as how physical exercise affects cholesterol levels. Randomized controlled trials are the gold standard for such questions, but ethical, logistical, or financial constraints often necessitate reliance on observational data.

Causal inference from observational data requires assumptions about the data-generating process (DGP), most notably the absence of hidden confounders—unobserved variables that influence both treatment and outcome (Hernán and Robins, 2025; Lin et al., 2023). Denoting unobserved factors affecting the treatment and outcome by $\boldsymbol{U_a}$ and $\boldsymbol{U_y}$, respectively, this assumption implies independence between these variables (Pearl, 1995).

While average treatment effects are informative, many applications require more granular conclusions. Counterfactual inference goes beyond average or even conditional average causal effects and targets individual-level outcomes by reasoning about latent characteristics. For example, knowing the cholesterol level and amount of exercise, one may ask what an individual's cholesterol level

would have been had they exercised a different amount. Counterfactual reasoning allows personalized decision-making in domains such as medicine, education, and policy (Kino et al., 2021). A formal definition of counterfactual inference is given in Section 2.1.

When hidden confounders are present, instrumental variables (IVs) provide a principled approach to causal identification. An instrument $\boldsymbol{Z}$ affects the outcome only through its influence on the treatment. While classical IV methods address linear settings (Wald, 1940; Angrist et al., 1996), more recent work extends IV approaches to non-linear settings (Chesher, 2003; Imbens and Newey, 2009; Guo and Small, 2016; Puli and Ranganath, 2020). However, existing IV-based approaches to counterfactual inference typically assume separable outcome functions and one-dimensional outcomes (Hartford et al., 2017; Singh et al., 2019).

In practice, outcome functions are often nonseparable. For instance, the effect of age on cholesterol may differ depending on the level of exercise. In this work, we introduce a new identifiability result for counterfactuals under the assumptions that (i) a strong instrument is available and (ii) the outcome function follows a known triangular structure and is strictly monotonic in $\boldsymbol{U_y}$. Building on this result, we propose *Flow IV*, a method based on normalizing flows or flow matching that flexibly models complex, high-dimensional outcome functions beyond additive noise assumptions.

The remainder of the paper is organized as follows. Section 2 introduces the problem setup, IV assumptions, and related work. Section 3 presents our identifiability result, and Section 4 describes the proposed Flow IV method. Experimental results, including a real-world application to aid allocation in Africa (Dreher et al., 2021), are presented in Section 5. Section 6 concludes with a discussion and future directions.

2. Background and Related Work

This section defines the general structure of the type of DGPs this paper aims to perform counterfactual inference for and reviews related work.

2.1. Problem Setup

Let $\boldsymbol{Z}$, $\boldsymbol{A}$, and $\boldsymbol{Y}$ be continuous random variables of dimension k, m, and n respectively. Let $\boldsymbol{U_y}$ be an unobserved continuous random variable of dimension n and let $\boldsymbol{U_a}$ be an unobserved random variable of dimension w. With p_X we denote the probability density function (pdf) of random variable X. We call $\boldsymbol{A}$ the treatment and $\boldsymbol{Y}$ the outcome. $\boldsymbol{Z}$ is called instrument or IV if it satisfies the following three conditions where $\boldsymbol{Y}(\boldsymbol{a}, \boldsymbol{z})$ is the potential outcome of $\boldsymbol{Y}$ under intervention $\boldsymbol{A} = \boldsymbol{a},\ \boldsymbol{Z} = \boldsymbol{z}$.

1. *Relevance condition* $\boldsymbol{Z} \not\perp\!\!\!\perp \boldsymbol{A}$ states that the instrument is associated with the treatment,
2. *exclusion restriction* $\boldsymbol{Y}(\boldsymbol{z}, \boldsymbol{a}) = \boldsymbol{Y}(\boldsymbol{z}', \boldsymbol{a})$ for all $\boldsymbol{z},\ \boldsymbol{z}'$ requires that the instrument has no direct effect on the outcome, and
3. *marginal exchangeability* $\boldsymbol{Y}(\boldsymbol{a}, \boldsymbol{z}) \perp\!\!\!\perp \boldsymbol{Z}$ for all $\boldsymbol{z},\ \boldsymbol{a}$, requires that the instrument and outcome do not share common causes.

Let the DGP for $(\boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{Y})$ be defined by the following structural causal model (SCM) (Pearl, 2009).

SCM 1

$$\begin{aligned}\boldsymbol{Z} &= f_{\boldsymbol{z}}(\boldsymbol{U_z})\\ \boldsymbol{A} &= f_{\boldsymbol{a}}(\boldsymbol{Z}, \boldsymbol{U_a})\\ \boldsymbol{Y} &= f_{\boldsymbol{y}}(\boldsymbol{A}, \boldsymbol{U_y})\end{aligned}$$

where $\boldsymbol{U_z}$ is independent of $\boldsymbol{U_a}$ and $\boldsymbol{U_y}$. The joint distribution of $\boldsymbol{U_a}$ and $\boldsymbol{U_y}$ represents possible latent confounding between $\boldsymbol{A}$ and $\boldsymbol{Y}$. The corresponding causal graph of this SCM can be seen in Figure 1.

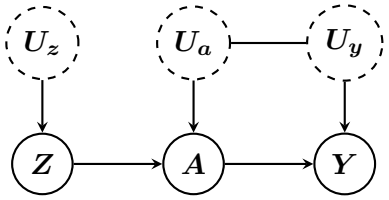


Figure 1: Causal graph illustrating the structure of the type of DGPs considered in this paper.

Note that, under this DGP, $\boldsymbol{Z}$ is an IV and the outcome function $f_{\boldsymbol{y}}$ can be any separable or nonseparable function. Formally, $f_{\boldsymbol{y}}$ is called separable if it can be written as $f_{\boldsymbol{y}}(\boldsymbol{A}, \boldsymbol{U_y}) = \tilde{f}_{\boldsymbol{y}}(\boldsymbol{A}) + \boldsymbol{U_y}$ where $\tilde{f}_{\boldsymbol{y}}$ is a possibly nonlinear function.

The goal of this paper is to perform counterfactual inference of the form "What would the value of $\boldsymbol{Y}_i$ for individual i have been, had the value of $\boldsymbol{A}_i$ been at the unobserved level $\boldsymbol{a}'$ when we actually observe $\boldsymbol{Y}_i = \boldsymbol{y},\ \boldsymbol{A}_i = \boldsymbol{a}$?" We express this counterfactual outcome formally as $\boldsymbol{Y}_i(\boldsymbol{a}') \mid \boldsymbol{Y}_i = \boldsymbol{y},\ \boldsymbol{A}_i = \boldsymbol{a}$. Answering these types of counterfactual queries requires three steps (Pearl, 2009).

1. *Abduction*, where we infer the latent value $\boldsymbol{U}_{\boldsymbol{y}_i}$ from the observed $(\boldsymbol{A}_i, \boldsymbol{Y}_i)$,
2. *action*, where we replace the value of $\boldsymbol{A}_i = \boldsymbol{a}$ with $\boldsymbol{a}'$, and
3. *prediction*, where we estimate the counterfactual outcome as $\boldsymbol{Y}_i(\boldsymbol{a}') = f_{\boldsymbol{y}}(\boldsymbol{a}', \boldsymbol{U}_{\boldsymbol{y}_i})$.

Note that while *probabilistic* counterfactual inference is possible under some assumptions, the counterfactual outcome is *uniquely* identified if and only if the following condition for the outcome function $f_{\boldsymbol{y}}$ holds for all $\boldsymbol{a}, \boldsymbol{a}', \boldsymbol{u_y}$, and $\boldsymbol{u}'_{\boldsymbol{y}}$

$$f_{\boldsymbol{y}}(\boldsymbol{a}, \boldsymbol{u_y}) = f_{\boldsymbol{y}}(\boldsymbol{a}, \boldsymbol{u_y}') \iff f_{\boldsymbol{y}}(\boldsymbol{a}', \boldsymbol{u_y}) = f_{\boldsymbol{y}}(\boldsymbol{a}', \boldsymbol{u_y}') \tag{1}$$

The condition says that if the abduction step results in a set $\{\boldsymbol{u_y}, \boldsymbol{u_y}'\}$ of possible values for $\boldsymbol{U_y}$ that produce the factual $\boldsymbol{y}$ given $\boldsymbol{a}$, then the counterfactual value of $\boldsymbol{y}$ under $\boldsymbol{u_y}$ has to be the same as the counterfactual value of $\boldsymbol{y}$ under $\boldsymbol{u_y}'$ for all alternative treatments $\boldsymbol{a}'$. Otherwise, the counterfactual is not uniquely identified. Note that when the condition is true, then we can rewrite SCM 1 such that $f_{\boldsymbol{y}}$ is invertible while the model stays counterfactually equivalent (proof in Appendix A).

2.2. Related Work

Utilizing IVs to estimate causal effects has a long tradition (Wald, 1940; Angrist et al., 1996). The classical IV estimator assumes a linear relationship between $\boldsymbol{Z}$, $\boldsymbol{A}$, and $\boldsymbol{Y}$ such that $\boldsymbol{Y} = \boldsymbol{\beta}\boldsymbol{A} + \boldsymbol{U_y}$. In the simplest setup with a single instrument, β can then be estimated using the so called Wald estimator $\hat{\beta} = \mathrm{Cov}(Y, Z)/\mathrm{Cov}(A, Z)$ (Wald, 1940). Following Hartford et al. (2017), we refer to it as the Usual IV approach. When the linearity assumption is violated, the Usual IV estimator can be used to estimate the complier average treatment effect (Liu et al., 2025). There is a vast body of literature that utilizes IVs to identify such causal effects for specific subpopulations (local average treatment effects) (Imbens and Angrist, 1994; Robins, 1994; Imbens, 2010).

Other literature aims at finding a *control function* $C(\boldsymbol{Z}, \boldsymbol{A})$ such that conditioning on the control variable $\boldsymbol{V} \coloneqq C(\boldsymbol{Z}, \boldsymbol{A})$ renders the treatment $\boldsymbol{A}$ independent of $\boldsymbol{U_y}$ (Chesher, 2003; Imbens and Newey, 2009; Guo and Small, 2016; Hahn and Ridder, 2017; Puli and Ranganath, 2020). The control variable then blocks all backdoor paths relative to $(\boldsymbol{A}, \boldsymbol{Y})$ which allows identification of causal effects. While Figure 1 shows that in the setup of this paper, a control variable exists in the form of $\boldsymbol{U_a}$, the method we propose does not directly aim at finding a control function, but yields a control function as a by-product. Puli and Ranganath (2020) propose what they call the *general control function* method (GCFN). They learn a control function using an autoencoder, where the encoder learns a distribution $F_{\boldsymbol{V}|\boldsymbol{Z},\boldsymbol{A}}$ and the decoder learns to reconstruct $\boldsymbol{A}$ from $\boldsymbol{Z}$ and $\boldsymbol{V}$. However, guaranteeing that $\boldsymbol{V}$ is a valid control function with the GCFN method requires knowledge about the structural form of the treatment function like additivity or multiplicity and that this structural form is reflected by the decoder. After training the control function, a neural network can be used to regress the outcome on the treatment and control function. This estimator of $\mathbb{E}\left[\boldsymbol{Y} \mid \boldsymbol{A} = \boldsymbol{a}, \boldsymbol{V} = v\right]$ can then be used to estimate the average effect as $\mathbb{E}\left[\boldsymbol{Y} \mid do(\boldsymbol{A} = \boldsymbol{a})\right] = \mathbb{E}_{v \sim F_{\boldsymbol{V}}}\left[\mathbb{E}\left[\boldsymbol{Y} \mid do(\boldsymbol{A} = \boldsymbol{a}), \boldsymbol{V} = v\right]\right] = \mathbb{E}_{v \sim F_{\boldsymbol{V}}}\left[\mathbb{E}\left[\boldsymbol{Y} \mid \boldsymbol{A} = \boldsymbol{a}, \boldsymbol{V} = v\right]\right]$ as $\boldsymbol{V}$ blocks all backdoor paths. The distribution of $\boldsymbol{V}$ is obtained from the decoder and the observed data distribution. How control functions can be used for counterfactual inference with nonseparable outcome functions is not the scope of above mentioned papers.

There exists literature that explicitly targets counterfactual inference with IVs that has relaxed the linearity assumption. However, Hartford et al. (2017) and Singh et al. (2019) assume separable outcome functions and Nasr-Esfahany et al. (2023) assume that instrument and treatment are discrete. Note that separability of the outcome function is a sufficient condition for the (necessary and sufficient) condition for counterfactual inference in Equation 1. Hartford et al. (2017) propose the *Deep IV* framework, where the problem is divided into two prediction tasks. In the first stage, the conditional distribution of the treatment under the instrument is learned whereas the second stage network models $g_{\boldsymbol{y}}(\boldsymbol{A}) + \mathbb{E}[\boldsymbol{U_y}]$. To counteract the spurious correlations between $\boldsymbol{A}$ and $\boldsymbol{Y}$, their loss function when training the second network involves integrating over the treatment distribution conditioned on the instrument. Note that under the separability assumption, we can also use other methods such as control function approaches to perform counterfactual inference as the abduction and prediction becomes trivial.

Table 1 summarizes the additional assumptions (beyond the existence of an IV) made by existing IV approaches, and includes our Flow IV method, whose assumptions are formally introduced in Section 3. The table also indicates whether an approach supports interventional inference, i.e., predicting outcome distributions under interventions for a (sub-) population, and counterfactual inference, i.e., predicting individual-level potential outcomes.

Approach	Assumptions	Interv.	Counterf.
Usual IV	Linear outcome function	✓	✓
Deep IV	Separable outcome function	✓	✓
GCFN	Structural treatment process assumptions	✓	✗
Flow IV (ours)	Triangular and strictly monotonic outcome function	✓	✓

Table 1: Comparison of assumptions and capabilities across IV approaches. In accordance with Pearl's causal hierarchy, the last two columns refer to the capability of performing interventional (Interv.) and counterfactual (Counterf.) inference. All approaches additionally require the existence of an IV.

3. Identifiability of the Outcome Function

In this section, we introduce a new identifiability result for counterfactual outcomes that relaxes the assumptions made in existing literature. First, we formally state the two assumptions required for the identifiability result to hold.

Assumption 1 (Triangular Monotonicity) *For $f_{\boldsymbol{y}}$ in SCM 1 we assume that $\boldsymbol{u_y} \mapsto f_{\boldsymbol{y}}(\boldsymbol{a}, \boldsymbol{u_y})$ is triangular, i.e.*

$$f_{\boldsymbol{y}}(\boldsymbol{a}, \boldsymbol{u_y}) = \begin{pmatrix} f_{\boldsymbol{y}}^{(1)}(\boldsymbol{a}, u_{\boldsymbol{y}}^{(1)}) \\ f_{\boldsymbol{y}}^{(2)}(\boldsymbol{a}, u_{\boldsymbol{y}}^{(2)}, u_{\boldsymbol{y}}^{(1)}) \\ \vdots \\ f_{\boldsymbol{y}}^{(n)}(\boldsymbol{a}, u_{\boldsymbol{y}}^{(n)}, u_{\boldsymbol{y}}^{(n-1)}, \dots, u_{\boldsymbol{y}}^{(1)}) \end{pmatrix}$$

and that each $f_{\boldsymbol{y}}^{(i)}$ is strictly monotonic with respect to $u_{\boldsymbol{y}}^{(i)}$.

Note that the separability assumption is a special case of the triangular monotonicity assumption and we therefore strictly relax the assumptions of existing approaches. In the setting where $\boldsymbol{Y}$ is one-dimensional, triangular monotonicity is equivalent to invertibility. Then, Assumption 1 is a necessary and sufficient condition for unique counterfactual inference and is therefore not specific to our approach (recall Section 2.1). For one-dimensional outcomes, previous research has argued why monotonicity and thus invertibility in the confounder is not unreasonable to hold in reality (Ogburn and VanderWeele, 2012). For higher-dimensional outputs, Assumption 1 is sufficient for unique counterfactual inference.

The triangularity assumption is inspired by triangular systems, which are common in econometrics, and is assumed to hold in real-world applications where the effect unfolds over time. In such cases, the outcome $y^{(t)}$ at time t is a function of the treatment and the hidden causes $\boldsymbol{u_y}$ up to time t, i.e., $y^{(t)} = f_{\boldsymbol{y}}^{(t)}(\boldsymbol{a}, u_{\boldsymbol{y}}^{(t)}, u_{\boldsymbol{y}}^{(t-1)}, \dots, u_{\boldsymbol{y}}^{(1)})$ (see Figure 7 in Appendix C for causal graph).

Assumption 2 (Strong IV) *We assume that $p_{\boldsymbol{A}|\boldsymbol{Z}}(\boldsymbol{A} = \boldsymbol{a} \mid \boldsymbol{Z} = \boldsymbol{z}) > 0$ for every $\boldsymbol{a}$ and $\boldsymbol{z}$ such that $p_{\boldsymbol{A}}(\boldsymbol{A} = \boldsymbol{a}) > 0$ and $p_{\boldsymbol{Z}}(\boldsymbol{Z} = \boldsymbol{z}) > 0$.*

Assumption 2 is similarly used in other IV literature and typically referred to as *strong IV* assumption. It states that any value of the treatment can be observed given any value of the instrument

with non-zero probability, which can be assessed by domain experts. This leads to the novel identifiability result in the following theorem.

Theorem 1 (Identifiability) *Let Assumption 1 and 2 be true. Let $g_{\boldsymbol{y}} : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n$ be a function where $\tilde{\boldsymbol{u}}_{\boldsymbol{y}} \mapsto g_{\boldsymbol{y}}(\boldsymbol{a}, \tilde{\boldsymbol{u}}_{\boldsymbol{y}})$ is triangular monotonic following the same triangular structure as $f_{\boldsymbol{y}}$ and let $\tilde{\boldsymbol{Y}} \coloneqq g_{\boldsymbol{y}}(\boldsymbol{A}, \tilde{\boldsymbol{U}}_{\boldsymbol{y}})$ be a random variable. Let $\tilde{\boldsymbol{U}}_{\boldsymbol{y}}$ be a random variable that has the same independencies as $\boldsymbol{U}_{\boldsymbol{y}}$ in the causal graph in Figure 1.*

Then for every $g_{\boldsymbol{y}}$ such that $(\tilde{\boldsymbol{Y}}, \boldsymbol{A}, \boldsymbol{Z}) \overset{d}{=} (\boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{Z})$, we have that

$$g_{\boldsymbol{y}}(\boldsymbol{a}, \tilde{\boldsymbol{u}}_{\boldsymbol{y}}) = f_{\boldsymbol{y}}(\boldsymbol{a}, \psi(\tilde{\boldsymbol{u}}_{\boldsymbol{y}}))$$

where ψ is an invertible function.

The proof can be found in Appendix A. Theorem 1 states that any such function $g_{\boldsymbol{y}}$ that produces the correct observed distribution differs from the true outcome function $f_{\boldsymbol{y}}$ only by an invertible transformation of the latent term $\tilde{\boldsymbol{u}}_{\boldsymbol{y}}$. This immediately implies that using the function $g_{\boldsymbol{y}}$ in the abduction and prediction steps results in the same counterfactual value as using the function $f_{\boldsymbol{y}}$ (proof in Appendix A). We can therefore use $g_{\boldsymbol{y}}$ to predict counterfactuals.

4. Implementation of Flow IV

In this section we introduce *Flow IV*, an approach that makes use of Theorem 1 to identify counterfactuals under relaxed assumptions compared to existing literature. The method uses normalizing flows or flow matching to model an invertible mapping between a latent space $(\tilde{\boldsymbol{U}}_{\boldsymbol{z}}, \tilde{\boldsymbol{U}}_{\boldsymbol{a}}, \tilde{\boldsymbol{U}}_{\boldsymbol{y}})$ and the observed space $(\boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{Y})$. The training procedure as pseudocode can be found in Appendix D.

According to Theorem 1, the latent variable $\tilde{\boldsymbol{U}}_{\boldsymbol{y}}$ has to respect the same independencies as $\boldsymbol{U}_{\boldsymbol{y}}$ and by definition of $g_{\boldsymbol{y}}$, we know that $\tilde{\boldsymbol{Y}}$ must follow the same independencies as $\boldsymbol{Y}$ in Figure 1. We therefore specify the flow model such that it respect the same independencies specified in the causal graph in Figure 1. Normalizing flows that respect the independencies specified in a causal graph have been used before and are referred to as causal graphical normalizing flows (cGNFs) (Balgi et al., 2022; Javaloy et al., 2023; Balgi et al., 2024). They learn a conditional normalizing flow g_i from standard normal noise $\mathbf{U}_i$ to each observed variable V_i conditioned on the parents of V_i in the corresponding causal graph and thereby resemble autoregressive-flows (Kingma et al., 2016). The conditional flow transformations in our setup are the following.

$$\begin{aligned}
\tilde{\boldsymbol{Z}} &= g_{\boldsymbol{z}}(\tilde{\boldsymbol{U}}_{\boldsymbol{z}}; \theta_{\boldsymbol{z}}) \\
\tilde{\boldsymbol{A}} &= g_{\boldsymbol{a}}(\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{U}}_{\boldsymbol{a}}; \theta_{\boldsymbol{a}}) \\
\tilde{\boldsymbol{Y}} &= g_{\boldsymbol{y}}(\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{U}}_{\boldsymbol{y}}; \theta_{\boldsymbol{y}})
\end{aligned}$$

Note how closely the conditional flow transformations resemble SCM 1. In contrast to existing cGNF approaches that assume no hidden confounding, in our setup $\tilde{\boldsymbol{U}}_{\boldsymbol{y}}$ and $\tilde{\boldsymbol{U}}_{\boldsymbol{a}}$ need not be independent. We therefore introduce another normalizing flow $h : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m \times \mathbb{R}^n$ from independent standard normal variables $\varepsilon_{\boldsymbol{a}}$ and $\varepsilon_{\boldsymbol{y}}$ to model the joint distribution of $(\tilde{\boldsymbol{U}}_{\boldsymbol{a}}, \tilde{\boldsymbol{U}}_{\boldsymbol{y}})$. We call the parameters of this transformation $\theta_{\tilde{\boldsymbol{u}}}$. $\tilde{\boldsymbol{U}}_{\boldsymbol{z}}$ is independent of the other noise terms. Let the distribution of $\tilde{\boldsymbol{U}}_{\boldsymbol{z}}$ be the standard normal distribution.

4.1. Flow IV with Normalizing Flows

By using autoregressive normalizing flows, the structure of $g_{\boldsymbol{y}}$ can be chosen such that it has the same monotonic triangular structure as $f_{\boldsymbol{y}}$ as it is required by Theorem 1. The parameters $\theta = (\theta_{\boldsymbol{z}}, \theta_{\boldsymbol{a}}, \theta_{\boldsymbol{y}}, \theta_{\tilde{\boldsymbol{u}}})$ of the transformations can be optimized via maximum likelihood estimation given observations $\mathcal{D} = \{(\boldsymbol{z}_i, \boldsymbol{a}_i, \boldsymbol{y}_i)\}_{i=1}^{\ell}$.

$$\begin{aligned}
\hat{\theta} &= \arg\max_{\theta} \log \mathcal{L}(\theta \mid \mathcal{D}) \\
\log \mathcal{L}(\theta \mid \mathcal{D}) &= \sum_{i=1}^{\ell} \log p_{\text{obs}}(\boldsymbol{z}_i, \boldsymbol{a}_i, \boldsymbol{y}_i \mid \theta) \\
&= \sum_{i=1}^{\ell} \log p_{\tilde{\boldsymbol{U}}} \left(g^{-1}(\boldsymbol{z}_i, \boldsymbol{a}_i, \boldsymbol{y}_i; \theta) \mid \theta_{\tilde{\boldsymbol{u}}} \right) + \log \left| \det \mathbb{J}_{g^{-1}(\boldsymbol{z}_i, \boldsymbol{a}_i, \boldsymbol{y}_i; \theta)} \right|
\end{aligned}$$

Here, p_{obs} and $p_{\tilde{\boldsymbol{U}}}$ are the pdfs of the target variables $(\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{A}}, \tilde{\boldsymbol{Y}})$ and the source variable $\tilde{\boldsymbol{U}} = (\tilde{\boldsymbol{U}}_{\boldsymbol{z}}, \tilde{\boldsymbol{U}}_{\boldsymbol{a}}, \tilde{\boldsymbol{U}}_{\boldsymbol{y}})$ respectively, $g^{-1}(\boldsymbol{z}_i, \boldsymbol{a}_i, \boldsymbol{y}_i) = (g_{\boldsymbol{z}}^{-1}(\boldsymbol{z}_i), g_{\boldsymbol{a}}^{-1}(\boldsymbol{z}_i, \boldsymbol{a}_i), g_{\boldsymbol{y}}^{-1}(\boldsymbol{a}_i, \boldsymbol{y}_i))$ is the inverse transformation, and $\mathbb{J}_{g^{-1}(\boldsymbol{z}_i, \boldsymbol{a}_i, \boldsymbol{y}_i)}$ denotes the Jacobian matrix of g^{-1} evaluated at $\boldsymbol{z}_i, \boldsymbol{a}_i, \boldsymbol{y}_i$.

As $(\tilde{\boldsymbol{U}}_{\boldsymbol{a}}, \tilde{\boldsymbol{U}}_{\boldsymbol{y}})$ is modelled with another normalizing flow, the density $p_{\tilde{\boldsymbol{U}}}$ in the likelihood equation can be derived similarly to the density p_{obs} above. With φ being the pdf of the standard normal distribution we have

$$\begin{aligned}
\log p_{\tilde{\boldsymbol{U}}}(\tilde{\boldsymbol{u}}_{\boldsymbol{z}}, \tilde{\boldsymbol{u}}_{\boldsymbol{a}}, \tilde{\boldsymbol{u}}_{\boldsymbol{y}} \mid \theta_{\tilde{\boldsymbol{u}}}) &= \log \varphi(\tilde{\boldsymbol{u}}_{\boldsymbol{z}}) + \log p_{\tilde{\boldsymbol{U}}_{\boldsymbol{a}}, \tilde{\boldsymbol{U}}_{\boldsymbol{y}}}(\tilde{\boldsymbol{u}}_{\boldsymbol{a}}, \tilde{\boldsymbol{u}}_{\boldsymbol{y}} \mid \theta_{\tilde{\boldsymbol{u}}}) \\
&= \log \varphi(\tilde{\boldsymbol{u}}_{\boldsymbol{z}}) + \log \varphi(\varepsilon_{\boldsymbol{a}}) + \log \varphi(\varepsilon_{\boldsymbol{z}}) + \log |\mathbb{J}_{h^{-1}(\tilde{\boldsymbol{u}}_{\boldsymbol{a}}, \tilde{\boldsymbol{u}}_{\boldsymbol{y}}; \theta_{\tilde{\boldsymbol{u}}})}| \\
\text{with } (\varepsilon_{\boldsymbol{a}}, \varepsilon_y) &= h^{-1}(\tilde{\boldsymbol{u}}_{\boldsymbol{a}}, \tilde{\boldsymbol{u}}_{\boldsymbol{y}}; \theta_{\tilde{\boldsymbol{u}}})
\end{aligned}$$

Normalizing flows converge to the true data distribution for $\ell \to \infty$ if the transformations are sufficiently flexible (Papamakarios et al., 2021). According to our identifiability result in Theorem 1 we can then use $g_{\boldsymbol{y}}$ to obtain consistent estimates of counterfactuals.

4.2. Flow IV with Flow Matching

While normalizing flows can in theory be used to model distributions of any dimensionality, flow matching has been shown to be very effective for modelling high-dimensional distributions. The method builds on continuous normalizing flows, where the time dependent flow transformation $g_{\boldsymbol{y}}$ is expressed indirectly through a vector field and an ordinary differential equation (ODE) (Lipman et al., 2023). The exact ODE is defined later in this section.

It is easy to show that choosing a neural network architecture for the vector field that has the same triangular structure as the outcome function implies that the solution of the ODE (i.e. the function $g_{\boldsymbol{y}}$) will have the same triangular structure. Encoding such triangular structure in architectures like multiplayer perceptrons (MLPs) is straightforward, but how to do this for more complex architecture remains to be investigated by future research. We show empirically in Section 5 that Flow IV outperforms associational models even when the triangular structure is not enforced.

We propose a two step process for training a flow matching model for Flow IV. First, optimize $\theta_{\boldsymbol{z}}, \theta_{\boldsymbol{a}}$ of the normalizing flows $g_{\boldsymbol{Z}}$ and $g_{\boldsymbol{A}}$ defined in the previous subsection via the maximum-likelihood objective and the observations $\{(\boldsymbol{z}_i, \boldsymbol{a}_i)\}_{i=1}^{\ell}$. For sufficiently flexible normalizing flow

transformations, the generated variables $(\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{A}})$ then converge to the true distribution of $(\boldsymbol{Z}, \boldsymbol{A})$ for $\ell \to \infty$.

In a second step, define the distribution of $(\tilde{\boldsymbol{U}}_{\boldsymbol{a}}, \tilde{\boldsymbol{U}}_{\boldsymbol{y}})$ via a normalizing flow transformation as in the previous subsection. However, now the marginal distribution of $\tilde{\boldsymbol{U}}_{\boldsymbol{a}}$ is already fixed in the first step to train the normalizing flow $g_{\boldsymbol{A}}$. We therefore use a conditional flow h from $\varepsilon_{\boldsymbol{Y}}$ to $\tilde{\boldsymbol{U}}_{\boldsymbol{y}}$ conditioned on $\tilde{\boldsymbol{U}}_{\boldsymbol{a}}$. We can do this without loss of generality since the marginal distribution of $\tilde{\boldsymbol{U}}_{\boldsymbol{a}}$ does not restrict the expressivity of our model.

Then, optimize the parameters $\theta_{\tilde{\boldsymbol{u}}}$ of h jointly with the parameters $\theta_{\boldsymbol{y}}$ of the vector field v_t via the conditional flow matching objective (Lipman et al., 2023) by sampling a source sample $\tilde{\boldsymbol{U}}_{\boldsymbol{y}}$ from the conditional normalizing flow h conditioned on $\tilde{\boldsymbol{U}}_{\boldsymbol{a}} = g_{\boldsymbol{A}}^{-1}(\boldsymbol{z}_i, \boldsymbol{a}_i)$ and matching it with $\boldsymbol{y}_i$ from the training set $\{(\boldsymbol{z}_i, \boldsymbol{a}_i, \boldsymbol{y}_i)\}_{i=1}^{\ell}$.

We thereby perform flow matching between the latent space $\tilde{\boldsymbol{U}}_{\boldsymbol{y}} \mid \tilde{\boldsymbol{U}}_{\boldsymbol{a}}$ at time $t = 0$ and an observed space $\tilde{\boldsymbol{Y}} \mid \tilde{\boldsymbol{A}}, \tilde{\boldsymbol{Z}}$ at time $t = 1$. Note that in the setup of this paper, the flow generating the outcome variable is conditioned on the treatment variable $\tilde{\boldsymbol{A}}$ to respect the given independencies. The vector field $v : [0, 1] \times \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n$ then constructs $g_{\boldsymbol{y}}$ via the following ODE.

$$\begin{aligned} \frac{\partial}{\partial t} g_{\boldsymbol{y}}(\boldsymbol{a}, \boldsymbol{u}_{\boldsymbol{y}}, t) &= v(t, \boldsymbol{a}, g_{\boldsymbol{y}}(\boldsymbol{a}, \boldsymbol{u}_{\boldsymbol{y}}, t)) \\ g_{\boldsymbol{y}}(\boldsymbol{a}, \boldsymbol{u}_{\boldsymbol{y}}, 0) &= \boldsymbol{u}_{\boldsymbol{y}} \end{aligned}$$

To perform counterfactual inference, the backward time ODE from $t = 1$ to $t = 0$ can be numerically solved in the abduction step and the forward time ODE from $t = 0$ to $t = 1$ can be solved numerically in the prediction step. The flow matching model that generates $\tilde{\boldsymbol{Y}} \mid \tilde{\boldsymbol{A}}, \tilde{\boldsymbol{Z}}$ converges to the true distribution of $\boldsymbol{Y} \mid \boldsymbol{A}, \boldsymbol{Z}$ for $\ell \to \infty$. Consequently, the transformation $g_{\boldsymbol{y}}$ implied by the vector field v can be used to predict counterfactuals according to Theorem 1.

5. Experiments

In this section, we evaluate our Flow IV method in different experimental settings with synthetic data by comparing it to the Deep IV and GCFN[1] approaches. Furthermore, we illustrate how Flow IV can be used to produce counterfactual predictions for high-dimensional outputs like images. Lastly, we reanalyse a study performed by Dreher et al. (2021) that investigates the impact of Chinese foreign aid on a region's economic development. Some additional experiments and details about the experimental setup can be found in the Appendix B.

5.1. Synthetic Data Experiments

We consider three different synthetic DGPs where all variables are 1D. In the first DGP, the outcome function is separable and therefore triangular monotonic. We call it DGP 1 and it satisfies the assumptions of Flow IV, Deep IV and GCFN. In the second DGP, Y is generated from a nonseparable but invertible (in 1D equivalent to triangular monotonic) outcome function. Therefore, only the assumptions of Flow IV are satisfied. In a third setup, we consider a noninvertible (and therefore

1. The GCFN method does not aim at performing counterfactual inference. The simplest approach to enable it to perform counterfactual inference is to assume separable outcome functions which is what we do in our experiments in order to compare it to Flow IV which does not require separability.

also nonseparable) outcome function. Therefore, the assumptions of none of the approaches are fulfilled.

$$
\begin{aligned}
&\text{DGP 1:} && Y = 0.6 \cdot A + U_y && U_y \propto \alpha \cdot U_a^2 + \frac{1}{8}\eta \quad \text{with } U_a, \eta \overset{iid}{\sim} \mathcal{N}(0,1) \\
&\text{DGP 2:} && Y = (\sin(A + 1.5) + 1) \cdot U_y && U_y \propto \alpha \cdot U_a^2 + \frac{1}{8}\eta \quad \text{with } U_a, \eta \overset{iid}{\sim} \mathcal{N}(0,1) \\
&\text{DGP 3:} && Y = 0.6 \cdot (A + U_y)^2 && (U_y, U_a) \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \\
& && && \text{with } \rho = -\exp(-\alpha) + 1
\end{aligned}
$$

All terms U_y are scaled to have mean zero and standard deviation one. The variable $\alpha \in \mathbb{R}^+$ controls the strength of confounding between A and Y. The full DGPs can be found in Appendix B.

As an evaluation metric, we use the expected squared prediction error for counterfactual outcomes or cf-MSE for short. We define it as follows.

$$
\text{cf-MSE} := \mathbb{E}_{(A,Y)\sim p_{A,Y}}\left[\mathbb{E}_{A'\sim p_A}\left[(Y(A') - \hat{Y}(A'))^2 \mid A, Y\right]\right]
$$

We draw 20,000 samples from these distributions and use Monte Carlo estimation to estimate the expected value. Note that this is only possible for synthetic data, because only then can we calculate the true counterfactual outcome $Y(a') \mid A = a, Y = y$. A low counterfactual MSE indicates good predictive quality for counterfactuals.

Figure 2 shows the estimated cf-RMSEs (defined as the square root of the cf-MSE) for the three different DGPs.

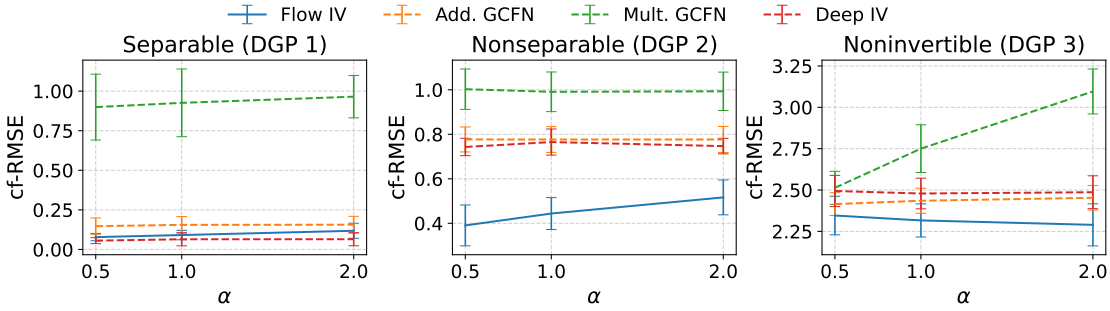


Figure 2: Comparison between Flow IV, Deep IV and GCFN in three different setups. Add. GCFN denotes the GCFN method with an additive decoder and Mult. GCFN with a multiplicative decoder. α controls the strength of confounding.

We can see that Flow IV is on par or outperforms Deep IV and GCFN in all setups and for almost all levels of confounding that were tested. The difference is the largest for DGP 2, because there, the triangular monotonicity assumption of Flow IV is satisfied whereas the separability assumption of Deep IV and GCFN is violated. According to the condition in Equation 1 in Section 2.1, unique

identification of counterfactuals is impossible for DGP 3, implying that there exists no model that could achieve a cf-MSE of zero in this setup. Nonetheless, Flow IV outperforms existing methods in DGP 3 which could be due to the higher flexibility by having relaxed the separability assumption. Figure 2 furthermore shows that the performance of GCFN depends on the choice of the decoder structure.

5.2. High-dimensional Outcome

In this section, we illustrate how Flow IV can be used to perform counterfactual inference for high-dimensional outcome variable $\boldsymbol{Y}$ as described in Section 4.2. This can be useful for tasks like counterfactual image editing (Pan and Bareinboim, 2024) under hidden confounding. Specifically, we perform counterfactual image editing on the MNIST dataset (LeCun et al., 1998).

We consider the following DGP.

$$\text{DGP:} \quad Z \sim \text{Unif}(0,1), \qquad A = (Z + 0.2 \cdot U_a + \frac{(D+1)}{10})\frac{1}{2.2}, \qquad \boldsymbol{Y} = A \cdot \boldsymbol{U_y}$$
$$\text{with } D \sim \text{Unif}\{0, 1 \ldots, 9\},\ U_a \sim \text{Unif}(0,1),\ \boldsymbol{U_y} \sim \text{Unif}\{\text{MNIST dataset} \mid \text{Digit} = D\}$$

The outcome variable is an MNIST image scaled by the treatment variable A which takes values on $(0, 1)$. This implies that the triangular monotonicity assumption 1 is satisfied as each output component $Y^{(i)}$ is only a function of A and $U_{\boldsymbol{y}}^{(i)}$. The confounder is the digit D shown on the outcome sample, where A tends to be larger for larger values of D. Therefore, higher digits tend to be multiplied by values of A close to one and are therefore similar to the original MNIST images, whereas lower digits tend to be multiplied by values of A close to zero and are therefore much dimmer than the original MNIST images. We compare the Flow IV method using flow matching to a purely associational model. We do not enforce the triangular structure of the true outcome function in our flow matching model, thereby violating the assumption made in Theorem 1 (more detail in Appendix B.2.3). The experiment serves as an ablation study to illustrate the effect of not enforcing the triangular structure. We compare the Flow IV model to an associational model, which also uses flow matching with the same architecture as in the Flow IV model but is only optimized to generate $\boldsymbol{Y} \mid A = a$ without utilizing the instrument Z.

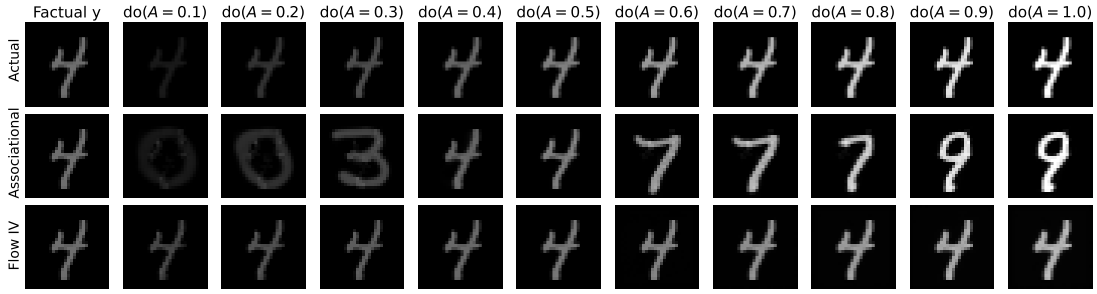


Figure 3: Comparison of counterfactual predictions for image data generated with Flow IV and a flow matching based associational model.

Figure 3 shows that the associational model seems to replicate the spurious association through the confounder (i.e. the digit shown in the image). When increasing the value of A, the associational model produces images of increasing digits, starting from digit zero at counterfactual level $A = 0.1$ up to digit nine at counterfactual level $A = 1.0$. Despite the fact that the triangular structure is not enforced, the Flow IV model correctly produces the digit four that can be seen on the factual image at all tested counterfactual levels of A. It seems like it slightly underestimates the effect of A as the dimming effect is less pronounced compared to the ground truth counterfactual images. The better performance is quantitively confirmed by a cf-MSE of the Flow IV model of $2.025 \cdot 10^{-3}$ compared to $14.919 \cdot 10^{-3}$ of the associational flow matching model.

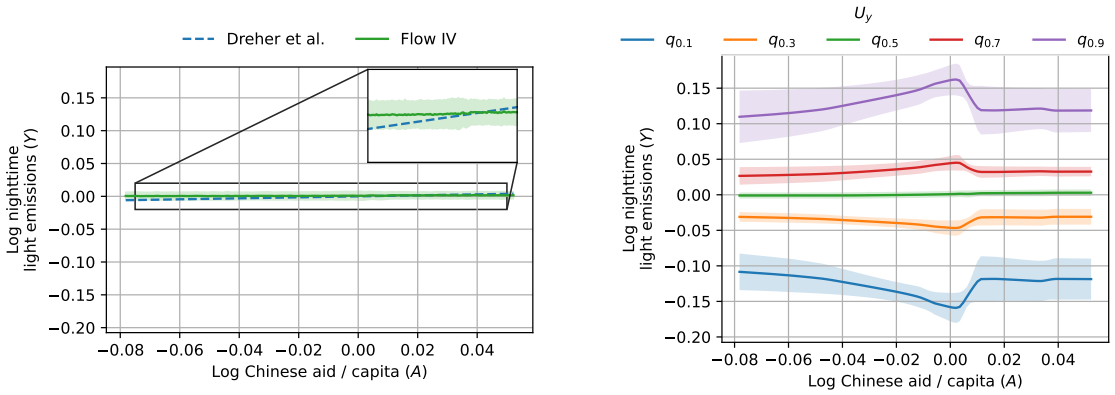
5.3. Real-world data: The Impact of Chinese Foreign Aid on Economic Development

IVs are common in social sciences and we replicate a foundational study in development economics, which influenced many subsequent studies. Specifically, we reanalyse an article investigating the impact of Chinese foreign aid on the economic development of different regions in Africa (Dreher et al., 2021). The article focuses on the role of political influence on the effectiveness of the supplied aid. In a first step, the authors estimate the causal effect of Chinese foreign aid in USD per capita provided in the previous year (treatment A) on the nighttime light emission of that region which is used as a proxy for economic development (outcome Y). Both A and Y are in log-scale. The nighttime light data used in the study comes from a dataset that was collected using weather satellites from the US Air Force (NOAA, 2014) and the data on Chinese foreign aid was introduced by Dreher et al.. The study uses the Chinese steel production as an instrument Z (data from WSA 2010; 2014).

The authors of the original article perform their analysis with the Usual IV approach and control for fixed effects capturing the yearly production volumes of steel and the recipient region's probability of receiving aid. We replicate this approach but use the Flow IV with normalizing flows to model the structural equations.

Dreher et al. (2021) found that the regression coefficient of the treatment variable was 0.0753 (Table 1 column 1 in Dreher et al. (2021)) which corresponds to the constant treatment effect according to their model. Figure 4(*a*) shows $\mathbb{E}[Y \mid do(A)]$ obtained from the Flow IV and the Usual IV estimate by Dreher et al. (2021). According to the Flow IV model, we cannot conclude that the effect of A on Y is non-zero, whereas Dreher et al. (2021) find a small but significant positive effect. However, the estimates of Dreher et al. (2021) lie within the bootstrap confidence interval of the Flow IV method for most values of A, suggesting that the Flow IV model does not differ significantly from the linear model.

Since Flow IV aims at performing counterfactual inference, the method allows us to go beyond an analysis of average effects among all analysed African regions and to investigate the effect of Chinese foreign aid provided to specific individual regions. Specifically, the learned outcome functions of Flow IV illustrated in Figure 4(*b*) can be used to assess counterfactuals according to the model. The abduction step corresponds to finding the (constant U_y) line which an observed pair (y, a) lies on. For example, if we want to perform counterfactual inference on a region with $A = -0.06$ and $Y = 0.12$, we can see in Figure 4(*b*) that the only value of U_y that can produce this observation corresponds to the purple curve. The action step corresponds to moving on the purple line (i.e. the value of U_y is unchanged) from the observed value of $A = a$ to the counterfactual value a'. For example, if we want to know what the Y value of the aforementioned region would have been had A been equal to zero, then moving on the purple line to $A = 0$ yields $Y'(a = 0 \mid A = -0.06, Y = 0.12) = 0.16$.



(*a*) $\mathbb{E}[Y \mid do(A = a)]$ under different values of a.

(*b*) Outcome function learned by the Flow IV model. q_i represents the i-quantile of U_y.

Figure 4: The outcome function learned with the Flow IV model in (b) and the expectation over U_y in (a).

Counterfactual inference allows us to make a personalized decision for this specific region: It would have seen the largest value in night-time light emissions (maximum of purple curve) for $A = 0$.

In general, Figure 4(*b*) suggests that regions with high levels of nighttime light emissions (e.g. red and purple curves) might benefit from Chinese foreign aid up to a certain level (up to $A = 0$, nighttime light emissions increase as the foreign aid increases) whereas countries with lower levels of nighttime light emissions might be harmed (orange and blue curve). This sort of counterfactual analysis can help decision-makers make more personalized interventions that account for regional heterogeneity in aid effectiveness.

6. Discussion

We proposed Flow IV, a method that allows to perform counterfactual inference in nonseparable and high-dimensional outcome settings. In addition to common IV assumptions, Flow IV relies on the assumption that the outcome function is triangular monotonic with respect to the latent factors, which for 1D outcomes is equivalent to invertibility. Invertibility is a necessary and sufficient condition for unique counterfactual inference. We showed that any correctly specified triangular monotonic outcome function that produces the observed distributions and that respects the independencies required by the IV setup is consistent with the true counterfactual. Therefore, any triangular monotonic generative model like autoregressive normalizing flows can make use of this result with Flow IV if the triangular ordering is known. We illustrated empirically how Flow IV with flow matching allows counterfactual image editing in the presence of hidden confounders even without encoding the triangular structure. How triangular monotonicity can be encoded in the structure of different network architectures in the flow matching approach and how $(\tilde{\boldsymbol{U}}_{\boldsymbol{a}}, \tilde{\boldsymbol{U}}_{\boldsymbol{y}})$ can be modelled with flow matching instead of normalizing flows was not addressed in this paper and is thus an open question for future research.

References

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996. ISSN 01621459. URL http://www.jstor.org/stable/2291629.

Sourabh Balgi, Adel Daoud, and José M. Peña. Personalized public policy analysis in social sciences using causal-graphical normalizing flows. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11810–11818, 2022. doi: 10.1609/aaai.v36i11.21437. URL https://cdn.aaai.org/ojs/21437/21437-13-25450-1-2-20220628.pdf.

Sourabh Balgi, Adel Daoud, José M. Peña, Geoffrey Wodtke, and Jesse Zhou. Deep learning with dags. *Sociological Methods & Research*, 2024. doi: 10.1177/00491241251319291. URL https://doi.org/10.1177/00491241251319291.

Andrew Chesher. Identification in nonseparable models. *Econometrica*, 71:1405–1441, 02 2003. doi: 10.1111/1468-0262.00454.

Axel Dreher, Andreas Fuchs, Roland Hodler, Bradley C. Parks, Paul A. Raschky, and Michael J. Tierney. African leaders and the geography of china's foreign assistance. *Journal of Development Economics*, 140:44–71, 2019. ISSN 0304-3878. doi: https://doi.org/10.1016/j.jdeveco.2019.04.003. URL https://www.sciencedirect.com/science/article/pii/S030438781831099X.

Axel Dreher, Andreas Fuchs, Roland Hodler, Bradley C. Parks, Paul A. Raschky, and Michael J. Tierney. Is favoritism a threat to chinese aid effectiveness? a subnational analysis of chinese development projects. *World Development*, 139:105291, 2021. ISSN 0305-750X. doi: https://doi.org/10.1016/j.worlddev.2020.105291. URL https://www.sciencedirect.com/science/article/pii/S0305750X20304186.

Zijian Guo and Dylan S. Small. Control function instrumental variable estimation of nonlinear causal effect models. *Journal of Machine Learning Research*, 17(100):1–35, 2016. URL http://jmlr.org/papers/v17/14-379.html.

Jinyong Hahn and Geert Ridder. Instrumental variable estimation of nonlinear models with nonclassical measurement error using control variables. *Journal of Econometrics*, 200(2):238–250, 2017. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2017.06.008. URL https://www.sciencedirect.com/science/article/pii/S0304407617300945. Measurement Error Models.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: a flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1414–1423, Sydney, NSW, Australia, 2017. JMLR.org.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. CRC Press, 2025. URL https://miguelhernan.org/whatifbook.

Guido W. Imbens. Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009). *Journal of Economic Literature*, 48(2):399–423, 2010. ISSN 00220515. URL http://www.jstor.org/stable/20778730.

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2951620.

Guido W. Imbens and Whitney K. Newey. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512, 2009. ISSN 00129682, 14680262. URL e.

Adrián Javaloy, Pablo Sanchez-Martin, and Isabel Valera. Causal normalizing flows: from theory to practice. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 58833–58864. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/b8402301e7f06bdc97a31bfaa653dc32-Paper-Conference.pdf.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/ddeebdeefdb7e7e7a697e1c3e3d8ef54-Paper.pdf.

Shiho Kino, Yu-Tien Hsu, Koichiro Shiba, Yung-Shin Chien, Carol Mita, Ichiro Kawachi, and Adel Daoud. A scoping review on the use of machine learning in research on social determinants of health: Trends and research prospects. *SSM - Population Health*, 15:100836, 2021. ISSN 2352-8273. doi: https://doi.org/10.1016/j.ssmph.2021.100836. URL https://www.sciencedirect.com/science/article/pii/S2352827321001117.

Yann LeCun, Corinna Cortes, and Christopher J.C. Burges. The mnist database of handwritten digits. http://yann.lecun.com/exdb/mnist/, 1998.

Cheng Lin, José M. Peña, and Adel Daoud. Assessing the unobserved: Enhancing causal inference in sociology with sensitivity analysis. *arXiv preprint arXiv:2311.13410*, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Jiewen Liu, Chan Park, Yonghoon Lee, Yunshu Zhang, Mengxin Yu, James M. Robins, and Eric J. Tchetgen Tchetgen. The multiplicative instrumental variable model, 2025. URL https://arxiv.org/abs/2507.09302.

Arash Nasr-Esfahany, MohammadIman Alizadeh, and Devavrat Shah. Counterfactual identifiability of bijective causal models. In *International Conference on Machine Learning*, 2023.

NOAA. National oceanic and atomspheric administration , version 4 dmsp-ols nighttime lights time series. boulder, co: National geophysical data center. https://eogdata.mines.edu/products/dmsp/, 2014. Accessed: 2014-04-05.

Elizabeth L. Ogburn and Tyler J. VanderWeele. On the nondifferential misclassification of a binary confounder. *Epidemiology*, 23(3):433–439, May 2012. ISSN 1044-3983. doi: 10.1097/ede.0b013e31824d1f63.

Yushu Pan and Elias Bareinboim. Counterfactual image editing. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39087–39101. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/pan24a.html.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.*, 22(1), January 2021. ISSN 1532-4435.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2337329.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2009. URL https://dl.acm.org/doi/book/10.5555/1642718.

Aahlad Puli and Rajesh Ranganath. General control functions for causal effect estimation from ivs. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8440–8451. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/604f2c31e67034642b288d76a8df11d5-Paper.pdf.

James M. Robins. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics - Theory and Methods*, 23(8):2379–2412, 1994. doi: 10.1080/03610929408831393. URL https://doi.org/10.1080/03610929408831393.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/17b3c7061788dbe82de5abe9f6fe22b3-Paper.pdf.

Abraham Wald. The Fitting of Straight Lines if Both Variables are Subject to Error. *The Annals of Mathematical Statistics*, 11(3):284 – 300, 1940. doi: 10.1214/aoms/1177731868. URL https://doi.org/10.1214/aoms/1177731868.

WSA. World steel association, statistical yearbook 2010, brussels, belgium: Worldsteel committee on economic studies. https://worldsteel.org/wp-content/uploads/Steel-Statistical-Yearbook-2010.pdf, 2010.

WSA. World steel association, statistical yearbook 2014, brussels, belgium: Worldsteel committee on economic studies. https://worldsteel.org/wp-content/uploads/Steel-Statistical-Yearbook-2014.pdf, 2014.

Appendix A. Technical Proofs

We start by proving the following claim made in Section 2.1.

Claim 1 (Justification of Assumption 1) *When the necessary and sufficient condition for unique counterfactual inference stated in Equation 1 is true, then we can rewrite SCM 1 such that the outcome function is invertible while the model stays counterfactually equivalent.*

Example 1 *Assume the outcome function is $f(A, U_y) = A + U_y^2$. Then we can only determine u_y in the abduction step up to its sign. However, note that all $\pm u_y$ result in the same counterfactual outcome under all alternative treatments (i.e. Condition 1 is true). We can therefore just replace $f(A, U_y)$ with $f^\star(A, U_y) = A + U_y^\star$ for $U_y^\star = U_y^2$ where then f and $f^\star$ are counterfactually equivalent.*

Proof The equivalency in Equation 1 states the following necessary condition for unique counterfactual inference.

$$f_{\boldsymbol{y}}(\boldsymbol{a}, \boldsymbol{u_y}) = f_{\boldsymbol{y}}(\boldsymbol{a}, \boldsymbol{u_y}') \iff f_{\boldsymbol{y}}(\boldsymbol{a}', \boldsymbol{u_y}) = f_{\boldsymbol{y}}(\boldsymbol{a}', \boldsymbol{u_y}')$$

Assume this condition is true. Now let $\tilde{f}_{\boldsymbol{y}} : \mathcal{X} \to \mathcal{Y}$ be a function with domain $\mathcal{X} \subseteq \mathcal{A} \times \mathcal{U}_y$ such that $\tilde{f}_{\boldsymbol{y}}(\boldsymbol{a}, \boldsymbol{u_y}) = f_{\boldsymbol{y}}(\boldsymbol{a}, \boldsymbol{u_y})$ for all $(\boldsymbol{a}, \boldsymbol{u_y}) \in \mathcal{X}$. Let $\mathcal{X}$ contain all elements of $\mathcal{A} \times \mathcal{U}_y$ except for pairs $(\boldsymbol{a}, \boldsymbol{u_y})$ and $(\boldsymbol{a}, \boldsymbol{u}_{\boldsymbol{y}}')$ for which $f_{\boldsymbol{y}}(\boldsymbol{a}, \boldsymbol{u_y}) = f_{\boldsymbol{y}}(\boldsymbol{a}, \boldsymbol{u_y}')$. For those pairs, only keep $(\boldsymbol{a}, \boldsymbol{u_y})$. This ensures that $\tilde{f}_{\boldsymbol{y}}$ is invertible.

The counterfactual outcome under alternative treatment $\boldsymbol{a}'$ given the observations $\boldsymbol{y}$ and $\boldsymbol{a}$ with the model $f_{\boldsymbol{y}}$ is

$$\begin{aligned}
\{\boldsymbol{u_y}, \boldsymbol{u}_{\boldsymbol{y}}'\} &= f_{\boldsymbol{y}}^{-1}(\boldsymbol{a}, \boldsymbol{y}) && \text{(Abduction)} \\
\boldsymbol{Y}(\boldsymbol{a}') &= f_{\boldsymbol{y}}(\boldsymbol{a}', \boldsymbol{u_y}) = f_{\boldsymbol{y}}(\boldsymbol{a}', \boldsymbol{u}_{\boldsymbol{y}}') && \text{(Action), (Cond. in Eq. 1)}
\end{aligned}$$

where $f_{\boldsymbol{y}}^{-1}$ is the multivalued inverse of $f_{\boldsymbol{y}}$.

The same counterfactual with the model $\tilde{f}_{\boldsymbol{y}}$ yields

$$\begin{aligned}
\boldsymbol{u_y} &= \tilde{f}_{\boldsymbol{y}}^{-1}(\boldsymbol{a}, \boldsymbol{y}) && \text{(Abduction), (Def. of } \tilde{f}_{\boldsymbol{y}}\text{)} \\
\tilde{\boldsymbol{Y}}(\boldsymbol{a}') &= \tilde{f}_{\boldsymbol{y}}(\boldsymbol{a}', \boldsymbol{u_y}) && \text{(Action)} \\
&= f_{\boldsymbol{y}}(\boldsymbol{a}', \boldsymbol{u_y}) && \text{(Def. of } \tilde{f}_{\boldsymbol{y}}\text{)} \\
&= \boldsymbol{Y}(\boldsymbol{a}')
\end{aligned}$$

which concludes the proof. ■

Before we prove Theorem 1, we will prove the following lemma.

Lemma 2 *Let $f_\alpha : \mathbb{R}^n \to \mathbb{R}^n$ be a function with parameter α and $g_\beta : \mathbb{R}^n \to \mathbb{R}^n$ be a function with parameter β. Let f_α and g_β be triangular monotonic as defined in Assumption 1. Let X be a continuous random variable and define $\tau \coloneqq f_\alpha(X)$ and $\eta \coloneqq g_\beta(X)$. For any choice of α and β let*

$$f_\alpha(X) \overset{d}{=} g_\beta(X)$$

It then follows that

$$f_\alpha(x) = g_\beta(x) \qquad a.s.$$

Proof By assumption we have

$$\tau \overset{d}{=} \eta$$

which implies that

$$\begin{gathered}\tau^{(1)} \overset{d}{=} \eta^{(1)} \\ \tau^{(2)} \mid \tau^{(1)} \overset{d}{=} \eta^{(2)} \mid \eta^{(1)} \\ \dots \\ \tau^{(n)} \mid \tau^{(n-1)} \dots \tau^{(1)} \overset{d}{=} \eta^{(n)} \mid \eta^{(n-1)} \dots \eta^{(1)}\end{gathered}$$

However, because each $(f_\alpha^{(i)}, f_\alpha^{(i-1)}, \dots f_\alpha^{(1)})$ is a bijective mapping between $(\tau^{(i)}, \tau^{(i-1)}, \dots, \tau^{(1)})$ and $(X^{(i)}, X^{(i-1)}, \dots, X^{(1)})$ due to the strict monotonicity assumption, we can write each $\tau^{(i)} \mid \tau^{(i-1)}, \dots, \tau^1$ as $\tau^{(i)} \mid x^{(i-1)}, \dots, x^1$. The same is true for η and g_β. We then have

$$\begin{aligned}&\tau^{(i)} \mid x^{(i-1)}, \dots, x^1 \overset{d}{=} \eta^{(i)} \mid x^{(i-1)}, \dots, x^1 \\ \iff\ & f_\alpha^{(i)}(X^{(i)}, x^{(i-1)}, \dots, x^{(1)}) \overset{d}{=} g_\beta^{(i)}(X^{(i)}, x^{(i-1)}, \dots, x^{(1)}) \\ \iff\ & f_\alpha^{(i)}(x^{(i)}, x^{(i-1)}, \dots, x^{(1)}) = g_\beta^{(i)}(x^{(i)}, x^{(i-1)}, \dots, x^{(1)}) \qquad \text{a.s.}\end{aligned}$$

where the last equality follows from the strict monotonicity of $f_\alpha^{(i)}$ and $g_\beta^{(i)}$ w.r.t. $X^{(i)}$. ■

Next, we show the proof of Theorem 1.

Proof (**Theorem** 1)

Let $\boldsymbol{Z}$, $\boldsymbol{A}$, and $\boldsymbol{Y}$ be random variables with nonzero continuous densities in $\mathcal{Z} \subseteq \mathbb{R}^k$, $\mathcal{A} \subseteq \mathbb{R}^m$, and $\mathcal{Y} \subseteq \mathbb{R}^n$ respectively. Let $\boldsymbol{U_y}$, and $\tilde{\boldsymbol{U}}_{\boldsymbol{y}}$ be unobserved random variables with nonzero continuous densities in $\mathcal{U}_y \subseteq \mathbb{R}^n$ and let $\boldsymbol{U_a}$ be an unobserved random variable with nonzero continuous density in $\mathcal{U}_a \subseteq \mathbb{R}^w$. Given the independencies specified by the graph in Figure 1, assume that the output $\boldsymbol{Y}$ is generated by an unknown function $f_{\boldsymbol{y}} : \mathcal{A} \times \mathcal{U}_y \to \mathcal{Y}$ such that $\boldsymbol{Y} \coloneqq f_{\boldsymbol{y}}(\boldsymbol{A}, \boldsymbol{U_y})$. We observe the conditional distribution of $\boldsymbol{Y} \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{A} = \boldsymbol{a}$. We want to find a function $g_{\boldsymbol{y}} : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n$ such that $(\boldsymbol{Y} \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{A} = \boldsymbol{a}) \overset{d}{=} g(\boldsymbol{a}, (\tilde{\boldsymbol{U}}_{\boldsymbol{y}} \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{A} = \boldsymbol{a}))$ where $\tilde{\boldsymbol{U}}_{\boldsymbol{y}}$ has the same independencies as $\boldsymbol{U_y}$ and $\overset{d}{=}$ denotes equality in distribution.

By Assumption 1, $f_{\boldsymbol{y}}$ is invertible and we have

$$(\boldsymbol{U_y} \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{A} = \boldsymbol{a}) \overset{d}{=} f_{\boldsymbol{y}}^{-1}(\boldsymbol{a}, g_{\boldsymbol{y}}(\boldsymbol{a}, (\tilde{\boldsymbol{U}}_{\boldsymbol{y}} \mid \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{A} = \boldsymbol{a}))) \tag{2}$$

where $f_{\boldsymbol{y}}^{-1}(\cdot, \boldsymbol{u_y})$ denotes the inverse of $f_{\boldsymbol{y}}(\cdot, \boldsymbol{u_y})$.

Let $f_{\boldsymbol{a}}^{-1} : \mathcal{Z} \times \mathcal{A} \rightrightarrows \mathcal{U}_a$ be the possibly multivalued inverse of the treatment function such that $f_{\boldsymbol{a}}^{-1}(\boldsymbol{z}, \boldsymbol{a}) = \{\boldsymbol{u_a} \in \mathcal{A} \mid \boldsymbol{a} = f_{\boldsymbol{a}}(\boldsymbol{z}, \boldsymbol{u_a})\}$ where $f_{\boldsymbol{a}}$ is the treatment function.

Let $\boldsymbol{U}_{\boldsymbol{y}}^c$ be a random variable with the same distribution as $\boldsymbol{U_y} \mid \boldsymbol{U_a} \in f_{\boldsymbol{a}}^{-1}(\boldsymbol{z}, \boldsymbol{a})$ and let $\tilde{\boldsymbol{U}}_{\boldsymbol{y}}^c$ be similarly defined for $\tilde{\boldsymbol{U}}_{\boldsymbol{y}} \mid \boldsymbol{U_a} \in f_{\boldsymbol{a}}^{-1}(\boldsymbol{z}, \boldsymbol{a})$. We then have

$$\boldsymbol{U}_{\boldsymbol{y}}^c \overset{d}{=} f_{\boldsymbol{y}}^{-1}\left(\ \boldsymbol{a},\ g\left(\boldsymbol{a},\ (\tilde{\boldsymbol{U}}_{\boldsymbol{y}}^c)\right)\ \right) \tag{3}$$

Let $\boldsymbol{x}^{(1:i)} \coloneqq (x^{(1)}, \dots, x^{(i)})^T$ and $\boldsymbol{x}^{(1:0)} = \emptyset$ and define

$$q(\boldsymbol{u_y}; \boldsymbol{u_a}) \coloneqq (q^{(1)}(u_y^{(1)}; \boldsymbol{u_a}, \boldsymbol{u_y}^{(1:0)}), q^{(2)}(u_y^{(2)}; \boldsymbol{u_a}, \boldsymbol{u_y}^{(1:1)}), \dots, q^{(n)}(u_y^{(n)}; \boldsymbol{u_a}, \boldsymbol{u_y}^{(1:n-1)}))^T \tag{4}$$

$$r(\tilde{\boldsymbol{u}}_{\boldsymbol{y}}; \boldsymbol{u_a}) \coloneqq (r^{(1)}(\tilde{u}_y^{(1)}; \boldsymbol{u_a}, \tilde{\boldsymbol{u}}_{\boldsymbol{y}}^{(1:0)}), r^{(2)}(\tilde{u}_y^{(2)}; \boldsymbol{u_a}, \tilde{\boldsymbol{u}}_{\boldsymbol{y}}^{(1:1)}), \dots, r^{(n)}(\tilde{u}_y^{(n)}; \boldsymbol{u_a}, \tilde{\boldsymbol{u}}_{\boldsymbol{y}}^{(1:n-1)}))^T \tag{5}$$

as the vectors of unobserved (univariate) conditional cumulative distribution functions (CDFs). This means each component is defined as

$$q^{(i)}(u_{\boldsymbol{y}}^{(i)}; \boldsymbol{u_a}, \boldsymbol{u_y}^{(1:i-1)}) \coloneqq \Pr(U_{\boldsymbol{y}}^{(i)} \leq u_{\boldsymbol{y}}(i) \mid \boldsymbol{U_a} \in \boldsymbol{u_a}, \boldsymbol{U_y}^{(1:i-1)} = \boldsymbol{u_y}^{(1:i-1)}) \tag{6}$$

$$r(\tilde{\boldsymbol{u}}_{\boldsymbol{y}}^{(i)}; \boldsymbol{u_a}, \tilde{\boldsymbol{u}}_{\boldsymbol{y}}^{(1:i-1)}) \coloneqq \Pr(\tilde{U}_{\boldsymbol{y}}^{(i)} \leq \tilde{\boldsymbol{u}}_{\boldsymbol{y}}(i) \mid \boldsymbol{U_a} \in \boldsymbol{u_a}, \tilde{\boldsymbol{U}}_{\boldsymbol{y}}^{(1:i-1)} = \tilde{\boldsymbol{u}}_{\boldsymbol{y}}^{(1:i-1)}) \tag{7}$$

Let the inverse of $q(\boldsymbol{u_y}; \boldsymbol{u_a})$ be called $q^{-1}(u; \boldsymbol{u_a})$.

We can then use the Rosenblatt transform with q and r to transform $(\boldsymbol{U_y} \mid \boldsymbol{U_a} = \boldsymbol{u_a})$ and $(\tilde{\boldsymbol{U}}_{\boldsymbol{y}} \mid \boldsymbol{U_a} = \boldsymbol{u_a})$ respectively to a multivariate uniform random variable

$$q((\boldsymbol{U_y} \mid \boldsymbol{U_a} \in \boldsymbol{u_a}); \boldsymbol{u_a}) \overset{d}{=} r((\tilde{\boldsymbol{U}}_{\boldsymbol{y}} \mid \boldsymbol{U_a} \in \boldsymbol{u_a}); \boldsymbol{u_a}) \sim \mathrm{Unif}([0,1]^n) \tag{8}$$

$$\iff (\boldsymbol{U_y} \mid \boldsymbol{U_a} \in \boldsymbol{u_a}) \overset{d}{=} q^{-1}(r((\tilde{\boldsymbol{U}}_{\boldsymbol{y}} \mid \boldsymbol{U_a} \in \boldsymbol{u_a}); \boldsymbol{u_a}); \boldsymbol{u_a}) \tag{9}$$

Together with Equation 3 this results in the following equality.

$$q^{-1}(r(\tilde{\boldsymbol{U}}_{\boldsymbol{y}}^c; f_{\boldsymbol{a}}^{-1}(\boldsymbol{z}, \boldsymbol{a})); f_{\boldsymbol{a}}^{-1}(\boldsymbol{z}, \boldsymbol{a})) \overset{d}{=} f_{\boldsymbol{y}}^{-1}\left(\boldsymbol{a}, g_{\boldsymbol{y}}(\boldsymbol{a}, (\tilde{\boldsymbol{U}}_{\boldsymbol{y}}^c))\right) \tag{10}$$

Because the left-hand side (LHS) is triangular monotonic as a consequence of the Rosenblatt transform and the right-hand side (RHS) expression is triangular monotonic according to Assumption 1, from Lemma 2 it follows that that for all $\tilde{\boldsymbol{u}}_{\boldsymbol{y}} \in \{\tilde{\boldsymbol{u}}_{\boldsymbol{y}} \in \mathcal{U}_y \mid \Pr(\tilde{\boldsymbol{U}}_{\boldsymbol{y}}^c) > 0\}$

$$q^{-1}(r(\tilde{\boldsymbol{u}}_{\boldsymbol{y}}; f_{\boldsymbol{a}}^{-1}(\boldsymbol{z}, \boldsymbol{a})); f_{\boldsymbol{a}}^{-1}(\boldsymbol{z}, \boldsymbol{a})) = f_{\boldsymbol{y}}^{-1}\left(\boldsymbol{a}, g_{\boldsymbol{y}}(\boldsymbol{a}, \tilde{\boldsymbol{u}}_{\boldsymbol{y}})\right) \tag{11}$$

Because $f_{\boldsymbol{a}}$ is a deterministic function and because $\boldsymbol{A}$ and $\boldsymbol{Z}$ are not independent according to the causal graph in Figure 1, Assumption 2 implies that for every tuple $(\boldsymbol{a}, \boldsymbol{u_a})$ with $\boldsymbol{a} \in \mathcal{A},\ \boldsymbol{u_a} \in \mathcal{U}_a$ there exists $\boldsymbol{z} \in \mathcal{Z}$ such that $\boldsymbol{u_a} \in f_{\boldsymbol{a}}^{-1}(\boldsymbol{z}, \boldsymbol{a})$.

Then, we know that for all $\boldsymbol{u_a} \in \mathcal{U}_a$ we have that

$$q^{-1}(r(\tilde{\boldsymbol{u}}_{\boldsymbol{y}}; \boldsymbol{u_a}); \boldsymbol{u_a}) = f_{\boldsymbol{y}}^{-1}\left(\boldsymbol{a}, g_{\boldsymbol{y}}(\boldsymbol{a}, \tilde{\boldsymbol{u}}_{\boldsymbol{y}})\right) \tag{12}$$

Because the LHS does not depend on $\boldsymbol{a}$ and the RHS does not depend on $\boldsymbol{u_a}$, and because both sides are invertible w.r.t. $\tilde{\boldsymbol{u}}_{\boldsymbol{y}}$, we can conclude that there has to exist an invertible function $\psi : \mathcal{Y} \to \mathcal{Y}$ that only depends on $\tilde{\boldsymbol{u}}_{\boldsymbol{y}}$ such that

$$f_{\boldsymbol{y}}^{-1}\left(\boldsymbol{a}, g_{\boldsymbol{y}}(\boldsymbol{a}, \tilde{\boldsymbol{u}}_{\boldsymbol{y}})\right) = \psi(\tilde{\boldsymbol{u}}_{\boldsymbol{y}}) \tag{13}$$

$$\iff g_{\boldsymbol{y}}(\boldsymbol{a}, \tilde{\boldsymbol{u}}_{\boldsymbol{y}}) = f_{\boldsymbol{y}}\left(\boldsymbol{a}, \psi(\tilde{\boldsymbol{u}}_{\boldsymbol{y}})\right) \tag{14}$$

We have shown that if Assumptions 1 and 2 are true, then all triangular monotonic $g_{\boldsymbol{y}}$ (following the same triangular ordering as $f_{\boldsymbol{y}}$) that produce the correct observed distribution differs from $f_{\boldsymbol{y}}$ only by some invertible transformation of the parameter $\tilde{\boldsymbol{u}}_{\boldsymbol{y}}$ which concludes the proof of Theorem 1.

■

Claim 2 (Justification of the implication after Theorem 1) *Under the conditions of Theorem 1, using $g_{\boldsymbol{y}}$ in the abduction and prediction steps yields the same counterfactual value as using $f_{\boldsymbol{y}}$.*

Proof Let $g_{\boldsymbol{y}}(\boldsymbol{a}, \tilde{\boldsymbol{u}}_{\boldsymbol{y}}) = f_{\boldsymbol{y}}\left(\boldsymbol{a}, \psi(\tilde{\boldsymbol{u}}_{\boldsymbol{y}})\right)$. The counterfactual outcome under the model $\boldsymbol{y} = f_{\boldsymbol{y}}(\boldsymbol{a}, \boldsymbol{u}_{\boldsymbol{y}})$ is

$$\begin{aligned}
\boldsymbol{u}_{\boldsymbol{y}} &= f_{\boldsymbol{y}}^{-1}(\boldsymbol{a}, \boldsymbol{y}) && \text{(Abduction)} \\
\boldsymbol{Y}(\boldsymbol{a}') &= f_{\boldsymbol{y}}(\boldsymbol{a}', \boldsymbol{u}_{\boldsymbol{y}}) && \text{(Action)} \\
&= f_{\boldsymbol{y}}(\boldsymbol{a}', f_{\boldsymbol{y}}^{-1}(\boldsymbol{a}, \boldsymbol{y}))
\end{aligned}$$

The counterfactual outcome under the model $\hat{\boldsymbol{y}} = g_{\boldsymbol{y}}(\boldsymbol{a}, \tilde{\boldsymbol{u}}_{\boldsymbol{y}}) = f_{\boldsymbol{y}}(\boldsymbol{a}, \psi(\tilde{\boldsymbol{u}}_{\boldsymbol{y}}))$ is

$$\begin{aligned}
\tilde{\boldsymbol{u}}_{\boldsymbol{y}} &= g^{-1}(\boldsymbol{a}, \boldsymbol{y}) && \text{(Abduction)} \\
&= \psi^{-1}(f_{\boldsymbol{y}}^{-1}(\boldsymbol{a}, \boldsymbol{y})) \\
\hat{\boldsymbol{Y}}(\boldsymbol{a}') &= g_{\boldsymbol{y}}(\boldsymbol{a}', \tilde{\boldsymbol{u}}_{\boldsymbol{y}}) && \text{(Action)} \\
&= f_{\boldsymbol{y}}(\boldsymbol{a}', \psi(\tilde{\boldsymbol{u}}_{\boldsymbol{y}})) \\
&= f_{\boldsymbol{y}}(\boldsymbol{a}', \psi(\psi^{-1}(f_{\boldsymbol{y}}^{-1}(\boldsymbol{a}, \boldsymbol{y})))) \\
&= f_{\boldsymbol{y}}(\boldsymbol{a}', f_{\boldsymbol{y}}^{-1}(\boldsymbol{a}, \boldsymbol{y})) \\
&= \boldsymbol{Y}(\boldsymbol{a}')
\end{aligned}$$

which concludes the proof. ■

Appendix B. Experiments

B.1. Synthetic Data Generation

In Section 5 we introduce three different DGPs that represent the assumptions that the different IV approaches make. We specify the full DGPs here.

DGP 1:
$$
\begin{aligned}
Z &= U_Z \\
A &= Z + U_A \\
Y &= 0.6 \cdot A + U_Y \\
U_Y &= \left(\alpha \cdot U_A^2 + \frac{1}{8}\eta - \alpha\right)\left(2\alpha^2 + \frac{1}{64}\right)^{-\frac{1}{2}} \\
&\text{with } U_A, \eta \overset{iid}{\sim} \mathcal{N}(0,1)
\end{aligned}
$$

DGP 2:
$$
\begin{aligned}
Z &= U_Z \\
A &= Z + \sin(U_A) \\
Y &= (\sin(A + 1.5) + 1) \cdot U_Y \\
U_Y &= \left(\alpha \cdot U_A^2 + \frac{1}{8}\eta - \alpha\right)\left(2\alpha^2 + \frac{1}{64}\right)^{-\frac{1}{2}} \\
&\text{with } U_A, \eta \overset{iid}{\sim} \mathcal{N}(0,1)
\end{aligned}
$$

DGP 3:
$$
\begin{aligned}
Z &= U_Z \\
A &= Z + U_A \\
Y &= 0.6 \cdot (A + U_Y)^2 \\
(U_A, U_Y)^T &\sim \mathcal{N}(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}) \\
&\text{with } \rho = -\exp(-\alpha) + 1
\end{aligned}
$$

B.2. Additional Experiments

In addition to the experiments in Section 5, we performed ablation studies to investigate the predictive quality of Flow IV for counterfactuals when different assumptions are violated. We furthermore investigated the convergence of Flow IV under finite samples and the performance of Flow IV with flow matching when we enforce triangularity of the vector field network.

B.2.1. VIOLATION OF FLOW IV'S ASSUMPTIONS

We present two experiments that investigate the performance of Flow IV when the two of its core assumptions are violated.

Wrong Triangular Structure One of the assumptions of Flow IV is, that the outcome function of the true underlying DGP has a triangular structure and that the learned outcome function has the same triangular structure. We use the following DGP to train two Flow IV models, one with the correct and one with the reversed triangular structure, and a purely associational model in the sense that it assumes unconfoundedness.

$$
\begin{aligned}
Z &= U_Z \\
A &= 0.5(Z + U_A) \\
Y_1 &= A + U_{Y_1} \\
Y_2 &= (|A| + 0.1)U_{Y_1} + U_{Y_2} \\
\boldsymbol{Y} &= \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \\
\text{with} \qquad U_Z &\sim \mathcal{N}(0, 1) \\
\begin{pmatrix} U_A \\ U_{Y_1} \\ U_{Y_2} \end{pmatrix} &\sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho^2 \\ \rho & \rho^2 & 1 \end{pmatrix}\right), \quad \rho = 0.5
\end{aligned}
$$

For the model Flow IV (correct triangularity) we assume that Y_1 is a function only of U_{Y_1} and Y_2 is a function of U_{Y_1} and U_{Y_2} just like it is the case in the DGP. For the model Flow IV (wrong triangularity) we assume the opposite, i.e. Y_2 is a function only of U_{Y_1} and Y_1 is a function of U_{Y_1} and U_{Y_2}. The results can be seen in Table 2.

Model	Flow IV (correct triangularity)	Flow IV (wrong triangularity)	Associational Model
cf-RMSE	0.180 ± 0.047	0.388 ± 0.014	0.860 ± 0.023

Table 2: Comparison of counterfactual RMSE across models assuming wrong triangular structure.

We can see that in this example Flow IV outperforms the associational model that assumes unconfoundedness even if the wrong triangular structure is used. However, the results also suggest that choosing the correct triangular structure is (in general) a necessary assumption as the model with the correct triangular structure outperformed the model that assumed the wrong triangular structure.

Non-Triangular Structure We furthermore ran an experiment with the following DGP where the triangularity assumption is violated. Note that the outcome function is invertible, so counterfactual inference is in principle possible.

$$
\begin{aligned}
Z &= U_Z \\
A &= 0.5(Z + U_A) \\
Y_1 &= (|A| + 0.1)U_{Y_2} + U_{Y_1} \\
Y_2 &= (|A| + 0.1)U_{Y_1} + U_{Y_2} \\
\boldsymbol{Y} &= \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \\
\text{with} \qquad U_Z &\sim \mathcal{N}(0, 1) \\
\begin{pmatrix} U_A \\ U_{Y_1} \\ U_{Y_2} \end{pmatrix} &\sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho^2 \\ \rho & \rho^2 & 1 \end{pmatrix}\right), \quad \rho = 0.5
\end{aligned}
$$

From the results in Table 3 we can see that Flow IV does not outperform a model that assumes unconfoundedness in this experiment where the true DGP is not triangular, underlying the importance of the triangularity assumption.

From the experiments in this section it can be concluded that the triangularity assumption and the correct triangular specification of the Flow IV model are important to ensure that we can predict counterfactuals consistently. Otherwise, Flow IV may be outperformed even by purely associational models.

Model	Flow IV	Associational Model
cf-RMSE	0.594 ± 0.099	0.454 ± 0.057

Table 3: Comparison of counterfactual RMSE when the DGP is non-triangular but invertible.

B.2.2. FINITE-SAMPLE BEHAVIOUR

Using the following DGP, we investigated the convergence of Flow IV as the amount of training data increases. The results can be seen in Figure 5.

$$
\begin{aligned}
Z &= U_Z \\
A &= 0.5(Z + U_A) \\
\boldsymbol{Y} &= \begin{pmatrix} A \\ A \end{pmatrix} + \boldsymbol{U_y} \\
\text{with} \qquad U_Z &\sim \mathcal{N}(0, 1) \\
\begin{pmatrix} U_A \\ U_{Y_1} \\ U_{Y_2} \end{pmatrix} &\sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho^2 \\ \rho & \rho^2 & 1 \end{pmatrix}\right), \quad \rho = 0.5
\end{aligned}
$$

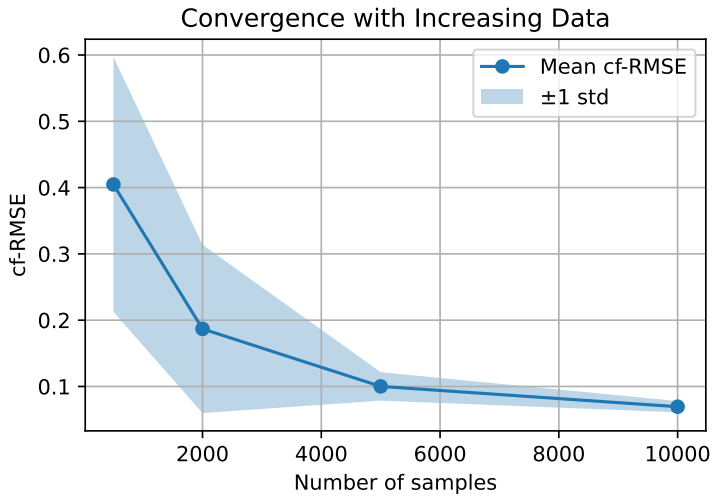


Figure 5: Finite sample convergence of Flow IV.

B.2.3. TRIANGULAR FLOW MATCHING

In Section 5.2 we present an experiment where we use flow matching with Flow IV but do not enforce that the outcome function has a triangular structure. We rerun the same experiment where we use a triangular MLP network for the vector field. However, as the examples in Figure 6 illustrates, the model failed to produce realistic looking images. Our assumption is that the MLP is not flexible enough for the task. How the triangular structure can be enforced with more flexible network architectures remains an open question for future research.

B.3. Training Details and Model Architecture

Training the models in this study was performed on a RTX 3060 GPU with 12 GB of VRAM running on Linux. For the normalizing flows we used conditional quadratic rational spline transformations with one flow step and 32 spline bins for all experiments. The conditioning network had three hidden layers with 24 hidden units and we used a learning rate of 0.00544 and a batch size of 256 for the Dreher example. For the synthetic data examples, we used a network with two hidden layers with ten hidden units and a learning rate of 0.001. The flow matching model was a UNet with seven convolutional layers (batch size 128). The Deep IV model is a MLP with two layers with ten hidden units and a learning rate of 0.001. For the GCFN we use a categorical control variable with 50 categories using two MLPs with two hidden layers with 100 hidden units for the encoder and decoder, followed by an outcome model using another 2-layer, 50-unit MLP. The learning rates are all set to 0.01 for encoder, decoder, marginal, and outcome model. The training set for the synthetic data was of size 15,000 and the batch size used for all models was 500. We used early stopping on a validation set of size 9,000 for all models. All details and the full replication code and models can be found in the repository https://github.com/Marbr987/flow-iv.

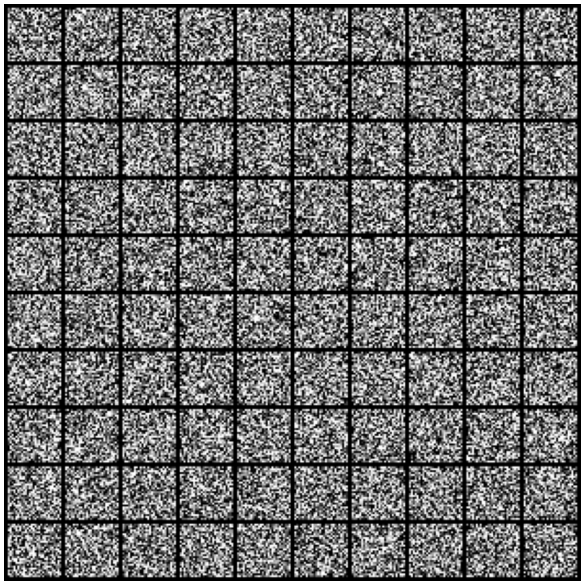


Figure 6: 10×10 grid of samples generated from a flow matching model with triangular MLP as vector field network.

Appendix C. Justification of Assumption 1

As stated in Section 3, the triangularity assumption of Assumption 1 is justified for example when a proces unfolds over time. Such process is illustrated in the following figure.

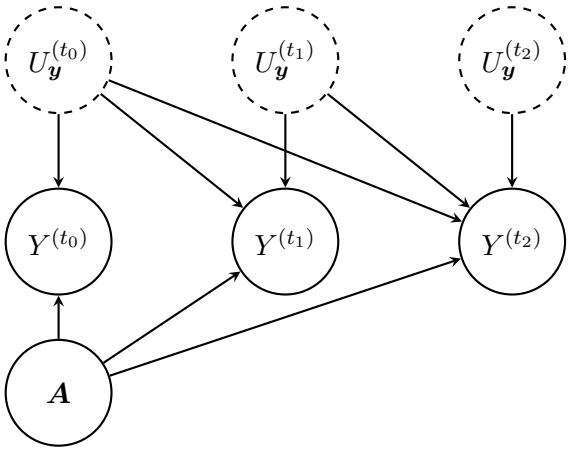


Figure 7: Causal graph illustrating how a triangular structure can arise from temporal processes.

Appendix D. Algorithms

Next we present the training procedures for Flow IV with normalizing flows and for Flow IV with flow matching as pseudocode.

Algorithm 1: Flow IV Training with Normalizing Flows

Input: Training dataset $\mathcal{D} = \{(\boldsymbol{z}_i, \boldsymbol{a}_i, \boldsymbol{y}_i)\}_{i=1}^{\ell}$
Hyperparameters: Learning rate η, number of iterations T
Output: Trained parameters $\theta = (\theta_{\boldsymbol{z}}, \theta_{\boldsymbol{a}}, \theta_{\boldsymbol{y}}, \theta_{\tilde{\boldsymbol{u}}})$
Randomly initialize parameters θ
for $t \leftarrow 1$ **to** T **do**
  Sample a minibatch $\mathcal{B} \subset \mathcal{D}$
  $\log \mathcal{L} \leftarrow 0$
  foreach $(\boldsymbol{z}_i, \boldsymbol{a}_i, \boldsymbol{y}_i) \in \mathcal{B}$ **do**
    $(\tilde{\boldsymbol{u}}_{\boldsymbol{z}}, \tilde{\boldsymbol{u}}_{\boldsymbol{a}}, \tilde{\boldsymbol{u}}_{\boldsymbol{y}}) \leftarrow g^{-1}(\boldsymbol{z}_i, \boldsymbol{a}_i, \boldsymbol{y}_i; \theta)$
    $(\varepsilon_{\boldsymbol{a}}, \varepsilon_{\boldsymbol{y}}) \leftarrow h^{-1}(\tilde{\boldsymbol{u}}_{\boldsymbol{a}}, \tilde{\boldsymbol{u}}_{\boldsymbol{y}}; \theta_{\tilde{u}})$
    $\log \mathcal{L} \leftarrow \log \mathcal{L} + \log \varphi(\tilde{\boldsymbol{u}}_{\boldsymbol{z}}) + \log \varphi(\varepsilon_{\boldsymbol{a}}) + \log \varphi(\varepsilon_{\boldsymbol{y}}) + \log |\det J_{h^{-1}}(\tilde{\boldsymbol{u}}_{\boldsymbol{a}}, \tilde{\boldsymbol{u}}_{\boldsymbol{y}}; \theta_{\tilde{u}})| + \log \left|\det J_{g^{-1}}(z_i, a_i, y_i; \theta)\right|$
  end
  $\theta \leftarrow \theta + \eta \nabla_\theta \log \mathcal{L}$ `// gradient ascent (MLE)`
end

Algorithm 2: Flow IV Training with Flow Matching

Input: Training dataset $\mathcal{D} = \{(\boldsymbol{z}_i, \boldsymbol{a}_i, \boldsymbol{y}_i)\}_{i=1}^{\ell}$,
Hyperparameters: Learning rates $\eta_1, \eta_2, \eta_3, \eta_4$, number of iterations T_1, T_2, noise schedule $\sigma(t, \boldsymbol{y})$, conditional mean function $\mu(t, \boldsymbol{y})$
Output: Trained parameters $\theta = (\theta_{\boldsymbol{z}}, \theta_{\boldsymbol{a}}, \theta_{\boldsymbol{y}}, \theta_{\tilde{\boldsymbol{u}}})$
Randomly initialize parameters θ
`// Step 1: Optimize normalizing flows` $g_{\boldsymbol{z}}$ `and` $g_{\boldsymbol{a}}$
for $t \leftarrow 1$ **to** T_1 **do**
  Sample a minibatch $\mathcal{B} \subset \mathcal{D}$
  $\log \mathcal{L} \leftarrow 0$
  foreach $(\boldsymbol{z}_i, \boldsymbol{a}_i, \boldsymbol{y}_i) \in \mathcal{B}$ **do**
    $\tilde{\boldsymbol{u}}_{\boldsymbol{z}} \leftarrow g_{\boldsymbol{z}}^{-1}(\boldsymbol{z}_i; \theta_{\boldsymbol{z}})$
    $\tilde{\boldsymbol{u}}_{\boldsymbol{a}} \leftarrow g_{\boldsymbol{a}}^{-1}(\boldsymbol{z}_i, \boldsymbol{a}_i; \theta_{\boldsymbol{a}})$
    $\log \mathcal{L} \leftarrow \log \mathcal{L} + \log \varphi(\tilde{\boldsymbol{u}}_{\boldsymbol{z}}) + \log \varphi(\tilde{\boldsymbol{u}}_{\boldsymbol{a}}) + \log \left| \det J_{(g_{\boldsymbol{z}}^{-1}, g_{\boldsymbol{a}}^{-1})}(z_i, a_i; \theta) \right|$
  end
  $\theta_{\boldsymbol{z}} \leftarrow \theta_{\boldsymbol{z}} + \eta_1 \nabla_{\theta_{\boldsymbol{z}}} \log \mathcal{L}$ `// gradient ascent (MLE)`
  $\theta_{\boldsymbol{a}} \leftarrow \theta_{\boldsymbol{a}} + \eta_2 \nabla_{\theta_{\boldsymbol{a}}} \log \mathcal{L}$
end
`// Step 2: Optimize normalizing flow` h `and vector field` v
for $t \leftarrow 1$ **to** T_2 **do**
  Sample a minibatch $\mathcal{B} \subset \mathcal{D}$
  $\mathcal{L} \leftarrow 0$
  foreach $(\boldsymbol{z}_i, \boldsymbol{a}_i, \boldsymbol{y}_i) \in \mathcal{B}$ **do**
    $\tilde{\boldsymbol{u}}_{\boldsymbol{a}} \leftarrow g_{\boldsymbol{a}}^{-1}(\boldsymbol{z}_i, \boldsymbol{a}_i; \theta_{\boldsymbol{a}})$
    Sample $\varepsilon_{\boldsymbol{y}} \sim \mathcal{N}(0, \mathbf{I})$
    $\tilde{\boldsymbol{u}}_{\boldsymbol{y}} \leftarrow h(\tilde{\boldsymbol{u}}_{\boldsymbol{a}}, \varepsilon_{\boldsymbol{y}}; \theta_{\tilde{\boldsymbol{u}}})$
    Sample $t \sim \mathrm{Unif}[0, 1]$
    $\boldsymbol{y}_t \leftarrow \mu(t, \boldsymbol{y}_i) + \sigma(t, \boldsymbol{y}_i) \tilde{\boldsymbol{u}}_{\boldsymbol{y}}$
    $v_{\text{cond}} \leftarrow \frac{\sigma'(t, \boldsymbol{y}_i)}{\sigma(t, \boldsymbol{y}_i)} (\boldsymbol{y}_t - \mu(t, \boldsymbol{y}_i)) + \mu'(t, \boldsymbol{y}_i)$
    $\mathcal{L} \leftarrow \mathcal{L} + (v_{\text{cond}} - v(t, a, y_t; \theta_{\boldsymbol{y}}))^2$
  end
  $\theta_{\tilde{\boldsymbol{u}}} \leftarrow \theta_{\tilde{\boldsymbol{u}}} - \eta_3 \nabla_{\theta_{\tilde{\boldsymbol{u}}}} \mathcal{L}$ `// gradient descent`
  $\theta_{\boldsymbol{y}} \leftarrow \theta_{\boldsymbol{y}} - \eta_4 \nabla_{\theta_{\boldsymbol{y}}} \mathcal{L}$
end