

# Diffusion Model in Causal Inference with Unmeasured Confounders

Tatsuhiro Shimizu

Department of Political Science and Economics

Waseda University

Tokyo, Japan

t.shimizu432@akane.waseda.jp

**Abstract**—We study how to extend the use of the diffusion model to answer the causal question from the observational data under the existence of unmeasured confounders. In Pearl’s framework of using a Directed Acyclic Graph (DAG) to capture the causal intervention, a Diffusion-based Causal Model (DCM) was proposed incorporating the diffusion model to answer the causal questions more accurately, assuming that all of the confounders are observed. However, unmeasured confounders in practice exist, which hinders DCM from being applicable. To alleviate this limitation of DCM, we propose an extended model called Backdoor Criterion based DCM (BDCM), whose idea is rooted in the Backdoor criterion to find the variables in DAG to be included in the decoding process of the diffusion model so that we can extend DCM to the case with unmeasured confounders. Synthetic data experiment demonstrates that our proposed model captures the counterfactual distribution more precisely than DCM under the unmeasured confounders.

**Index Terms**—Diffusion Model, Causal Inference, Unmeasured Confounders

## I. INTRODUCTION

Causal inference is the study of identifying the causal relationships between variables of one’s interest and developing the estimator for the estimands, such as the Average Treatment Effect (ATE), from the observational data. With ATE, for instance, we can use observational data to determine the personalized medicine [1] that maximizes the outcome, such as recovery from a disease. There are two mainstreams in causal inference: the Potential Outcome (PO) framework [2] and the Directed Acyclic Graph (DAG) framework [3]. In the DAG framework, Chao et al. (2023) [4] proposed the algorithm called the *Diffusion-based Causal Model (DCM)* that allows us to sample from the target distribution of our interest, by which we can calculate the approximation of ATE, outperforming the state-of-the-art algorithms [5] and [6]. However, only under causal sufficiency can the DCM sample from the target distribution, which requires the complete observation of all the confounders, which often does not hold in practice where confounders are the variables that affect both the cause and outcome variables of our interest. For instance, we often cannot observe stress levels, physical activities, mental health, sleep patterns, and genetic factors. To overcome the limitation of DCM, we extend it and propose a new algorithm to be able to estimate the ATE even under the existence of the unmeasured confounders by including the nodes that satisfy the backdoor criterion [3] in both training and sampling phases of the algorithm, which tells us which

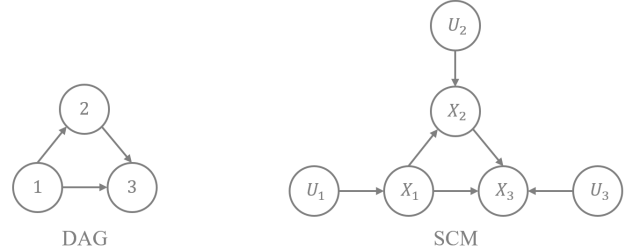


Fig. 1. DAG with three nodes and edges and the corresponding SCM with three exogenous and endogenous nodes

variables we should adjust. To illustrate the applicability of a new algorithm where unmeasured confounders exist, we conduct the synthetic data experiment for both simple and complex underlying data-generating processes. The experiment shows that our new algorithm samples precisely from the ground truth target distribution where DCM fails for both cases.

## II. BACKGROUND

We formulate the data-generating process, intervention, and causal effect in Pearl’s [3] framework. Firstly, DAG is the main element of Pearl’s framework and is defined as follows.

**Definition 1** (Directed Acyclic Graph). DAG  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  is a pair of the set of nodes  $\mathcal{V}$  and the set of edges  $\mathcal{E}$  where  $\mathcal{V} = \{1, \dots, d\}$  and  $\mathcal{E} = \{(i, j) : \exists \text{ edge from node } i \text{ to } j\}$ . DAG expresses variables by nodes and causal relationships by edges.

DAG only represents the relationship between nodes on which nodes affect which nodes. To quantify how the model generates the variables in terms of distributions and functions, we introduce a *structural causal model (SCM)*. We assume that we sample the observational data from the underlying SCM.

**Definition 2** (Structural Causal Model). Structural Causal Model (SCM)  $\mathcal{M} = (\mathcal{U}, \mathcal{V}, f)$  is the tuple of the set of exogenous variables  $\mathcal{U} = \{U_1, \dots, U_d\}$ , the set of endogenous variables  $\mathcal{V} = \{X_1, \dots, X_d\}$ , and the set of structural equations  $f = \{f_1, \dots, f_d\}$  such that for each  $i \in [d]$ , the endogenous variable satisfies  $X_i = f_i(\text{Pa}(X_i), U_i)$  where  $\text{Pa}(X_i)$  is the set of the parent nodes of  $X_i$  and  $d$  is the number of endogenous or exogenous nodes.

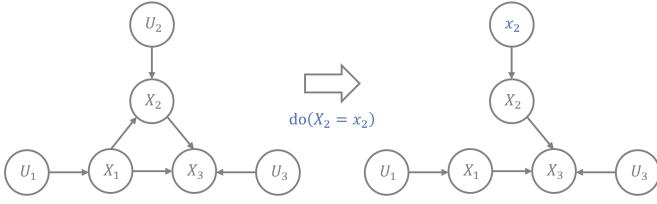


Fig. 2. do operator where we intervened in the node  $X_2 = x_2$  in SCM with three exogenous and endogenous nodes

Fig. 1 illustrates the examples of DAG  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  where  $\mathcal{V} = \{1, 2, 3\}$  and  $\mathcal{E} = \{(1, 2), (1, 3), (2, 3)\}$  and SCM  $\mathcal{M} = (\mathcal{U}, \mathcal{V}, f)$  with three endogenous and exogenous nodes where  $\mathcal{U} = \{U_1, U_2, U_3\}$ ,  $\mathcal{V} = \{X_1, X_2, X_3\}$ , and  $f = \{f_1, f_2, f_3\}$ .

In SCM, we assume that there exist unknown distributions of exogenous variables. We generate  $n$  independent and identically distributed samples from the distribution for each exogenous variable, but we do not observe them. Then, we observe endogenous variables according to the underlying structural equation for each. Thus, the observational data is  $X \in \mathbb{R}^{n \times d}$  where  $n$  is the number of samples and  $d$  is the number of nodes

Only with the observational data  $X \in \mathbb{R}^{n \times d}$  are we not sure about the parent nodes of each endogenous variable. Such information can be defined by the *topological order* as follows.

**Definition 3** (Topological Order). Topological order  $\pi = (\pi_1, \dots, \pi_d)$  is a permutation of  $d$  nodes in SCM such that  $\pi_i < \pi_j \iff X_j \in \text{De}(X_i)$  for all  $X_i, X_j \in \mathcal{V}$  such that  $i \neq j$  where  $\text{De}(X_i)$  is the set of the descendant nodes of  $i$ .

The problem of finding the topological order given the observational data is called *causal discovery* [7], [8]. As this problem is computationally intensive and NP-hard [9], most of the methods focus on the approximation of it. We assume we know the topological order of endogenous variables in SCM from which we get the observational data as we can get the estimated topological order by using the algorithm such as SCORE [10] or DiffAN [11], which use the properties of the leaf nodes and iteratively extract the leaf nodes to construct the topological order from the observational data.

Furthermore, we introduce *do-operator* that represents the intervention on SCM  $\mathcal{M}$  as follows.

**Definition 4** (do-operator). For any  $i \in [d]$ , We define  $\text{do}(X_i = x_i)$  by setting the corresponding exogenous variable to the intervened value  $U_i = x_i$  and deleting all the edges coming into  $X_i$  from the endogenous variables on SCM.

Fig. 2 shows the example of the do-operator where we intervene in the endogenous variable  $X_2$  to  $x_2$  on the SCM in Fig. 1.

The following defines the *average treatment effect (ATE)*, one of the causal effects we are interested in, using the do-operator.

**Definition 5** (Average Treatment Effect). For all  $X_i, X_j \in \mathcal{V}$  in SCM such that  $i \neq j$ , we define the ATE of the variable

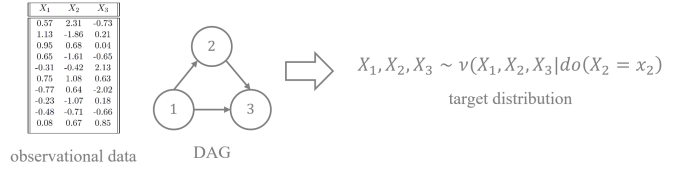


Fig. 3. Illustration of our problem: sampling from the target distribution after the intervention on a node in the SCM given the observational data and the underlying DAG

(cause)  $X_i$  on the variable (outcome)  $X_j$  when we compare two counterfactual situations  $X_i = x_i$  and  $X_i = 0$  by

$$\begin{aligned} ATE(x_i, 0) &:= \mathbb{E}[x_j | \text{do}(X_i = x_i)] - \mathbb{E}[x_j | \text{do}(X_i = 0)] \\ &= \int_{x_j} x_j \nu(X_j = x_j | \text{do}(X_i = x_i)) dx_j \\ &\quad - \int_{x_j} x_j \nu(X_j = x_j | \text{do}(X_i = 0)) dx_j \end{aligned}$$

where  $\nu(X_j | \text{do}(X_i = x_i))$  is the probability density function of  $X_j$  after the surgery on the SCM by do operator  $\text{do}(X_i = x_i)$ .

As we aim to figure out the causal effect of an arbitrary node on an arbitrary node, our problem boils down to how to approximately sample from the target distribution  $\nu(X_j | \text{do}(X_i = x_i))$  given observational data  $X \in \mathbb{R}^{n \times d}$  and underlying DAG for all  $i, j \in [d]$  such that  $i \neq j$  shown in Fig. 3. Note that we can estimate the underlying DAG by the topological order  $\pi$  and edge pruning algorithm [12] that uses the feature selection.

### III. EXISTING ALGORITHM

We introduce a diffusion-based algorithm called Diffusion-based Causal Model (DCM) proposed by Chao et al. (2023) [4], that can sample from the target distribution  $\nu(X_j | \text{do}(X_i = x_i))$  more accurately than existing state-of-the-art algorithms [5] and [6] under the following causal sufficiency.

**Assumption 6** (Causal Sufficiency). We say that the data-generating process satisfies causal sufficiency if no unmeasured confounders exist.

DCM uses Denoising Diffusion Implicit Model (DDIM) [13], a more efficient sampling algorithm than Denoising Diffusion Probabilistic Model (DDPM) [14] [15], which attained the groundbreaking performance in generating image and audio data [16], [17], [18]. DCM trains the diffusion model at each node to capture the characteristics of the exogenous nodes in SCM. In the forward diffusion process for each endogenous node, where we gradually add the isotropic Gaussian noise, we obtain the standard Gaussian distribution. Then, in the reverse diffusion process, we decode it by adding the Gaussian distribution with a learned parameter  $\theta$  to sample from the target distribution. As [19] shows that learning the parameter in the reverse diffusion process is equivalent to learning how much noise we add at each step, we also construct the neural network that captures how much noise  $\epsilon$  we should add according

---

**Algorithm 1**  $\text{Dec}_i(Z_i, X_{\text{Pa}_i})$  [4]

---

Input:  $Z_i, X_{\text{Pa}_i}$   
 Sample  $\hat{X}^T \sim Z_i$   
**for**  $t = T, \dots, 1$  **do**

$$\hat{X}_i^{t-1} \leftarrow \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} \hat{X}_i^t - \epsilon_\theta^i(\hat{X}_i^t, X_{\text{Pa}_i}, t) \\ \times \left( \sqrt{\frac{\alpha_{t-1}(1-\alpha_t)}{\alpha_t}} - \sqrt{1-\alpha_{t-1}} \right)$$

**end for**  
 Output:  $\hat{X}_i^0$

---



---

**Algorithm 2** DCM Training [4]

---

Input: target distribution  $\nu$ , scale factors  $\{\alpha_t\}_{t=1}^T$ , DAG  $\mathcal{G}$  whose node  $i$  is represented by  $X_i$

**while** not converge **do**

Sample  $X^0 \sim \nu$

**for**  $i = 1, \dots, d$  **do**

$t \sim \text{Unif}[\{1, \dots, T\}]$

$\epsilon \sim \mathcal{N}(0, 1)$

Update the parameter of the node  $i$ 's diffusion model  $\epsilon_\theta^i$  by minimization of the following loss function by Adam optimizer

$$\|\epsilon - \epsilon_\theta^i(\sqrt{\alpha_t}X_i^0 + \sqrt{1-\alpha_t}\epsilon, X_{\text{Pa}_i}^0, t)\|_2^2$$

**end for**  
**end while**

---

to the time  $t$  and the already sampled values of the parent nodes  $\hat{X}_{\text{Pa}_i}$  where  $X_{\text{Pa}_i}$  is the set of the parent nodes of  $X_i$  in SCM  $\mathcal{M}$ . After the training, we can sample from the target distribution. We sample the root node  $X_i$  in SCM from the empirical distribution  $E_i$ . For the intervened node  $X_i$ , we set it to the intervened value  $\gamma_i$ . For other nodes  $X_i$ , we sample by the reverse diffusion process  $\text{Dec}_i(Z_i, \text{Pa}(X_i))$  using the trained neural network  $\epsilon_\theta$  with parent nodes  $\text{Pa}(X_i)$  and the corresponding proxy exogenous nodes  $Z_i \sim \mathcal{N}(0, 1)$ . Algorithms 1, 2, and 3 show the comprehensive procedure of decoding, training, and sampling processes, respectively.

One of the crucial limitations of DCM [4] is that we cannot cope with the situation where there exist unmeasured confounders, which is often the case with the data collection for business, public health and social science where causal inference makes a significant contribution.

#### IV. PROPOSED ALGORITHM

##### A. Backdoor Diffusion-based Causal Model

To overcome the problem of DCM and use the observational data as much as possible, we introduce the novel **Backdoor Criterion-based DCM (BDCM)** algorithm inspired by the backdoor criterion proposed by Pearl [3]. To define the backdoor criterion, we introduce the notion of blocking a path in DAG.

---

**Algorithm 3** DCM Sampling [4]

---

$t$

Input: Intervened node  $j$  with value  $\gamma_j$ , noise  $Z_i \sim \mathcal{N}(0, 1)$   
**for all**  $i \in [d]$

**for**  $i = 1, \dots, d$  **do**

**if**  $i$  is a root node **then**

$\hat{X}_i \sim E_i$

**else if**  $i = j$  **then**

$\hat{X}_i \leftarrow \gamma_i$

**else**

$\hat{X}_i \leftarrow \text{Dec}_i(Z_i, \hat{X}_{\text{Pa}_i})$

**end if**

**end for**

**return**  $\hat{X} = (\hat{X}_1, \dots, \hat{X}_d)$

---



---

**Algorithm 4** BDCM Training

---

Input: target distribution  $\nu$ , scale factors  $\{\alpha_t\}_{t=1}^T$ , DAG  $\mathcal{G}$  whose node  $i$  is represented by  $X_i$  and intervened node  $j$  with intervened value  $\gamma_j$

**while** not converge **do**

Sample  $X^0 \sim \nu$

**for**  $i = 1, \dots, d$  **do**

$t \sim \text{Unif}[\{1, \dots, T\}]$

$\epsilon \sim \mathcal{N}(0, 1)$

Update the parameter of the node  $i$ 's diffusion model  $\epsilon_\theta^i$  by minimization of the following loss function depending on the nodes.

**if**  $X_j \in X_{\text{Pa}_i}$  **then**

$$\|\epsilon - \epsilon_\theta^i(\sqrt{\alpha_t}X_i^0 + \sqrt{1-\alpha_t}\epsilon, X_{\text{Pa}_i}^0, X_j, t)\|_2^2$$

**else**

$$\|\epsilon - \epsilon_\theta^i(\sqrt{\alpha_t}X_i^0 + \sqrt{1-\alpha_t}\epsilon, X_{\text{Pa}_i}^0, t)\|_2^2$$

**end if**

**end for**

**end while**

---

**Definition 7** (Block a Path). We say that the node  $Z$  blocks a path  $P$  if the path  $P$  includes a *chain*  $L \rightarrow Z \rightarrow R$ , or a *fork*  $L \leftarrow Z \rightarrow R$  where  $L$  and  $R$  are the nodes in the path  $P$ .

Then, using Definition 7, we define *backdoor criterion* as follows.

**Definition 8** (Backdoor Criterion). A set of variables  $\mathcal{B}$  satisfies backdoor criterion [3] for tuple  $(X, Y)$  in DAG  $\mathcal{G}$  if no node in  $\mathcal{B}$  is a descendant of  $X$  and  $\mathcal{B}$  blocks all paths between  $X$  (cause) and  $Y$  (outcome) which contains an arrow into  $X$ .

If unmeasured confounders exist, then the Backdoor criterion tells us which variable to adjust concerning tuple  $(X, Y)$ . Then, the idea of Backdoor DCM is that for each node  $X_i$  in SCM, instead of having the parents  $X_{\text{Pa}_i}$  and corresponding

**Algorithm 5** BDCM Sampling

---

Input: Intervened node  $j$  with value  $\gamma_j$ , noise  $Z_i \sim \mathcal{N}(0, 1)$   
for all  $i \in [d]$   
**for**  $i = 1, \dots, d$  **do**  
  **if**  $i = j$  **then**  
     $\hat{X}_i \leftarrow \gamma_i$   
  **else if**  $i$  is a root node **then**  
     $\hat{X}_i \sim E_i$   
  **else if**  $X_j \in X_{Pa_i}$  **then**  
     $\hat{X}_i \leftarrow \text{Dec}_i(Z_i, \hat{X}_{\mathcal{B}_i}, X_j)$   
  **else**  
     $\hat{X}_i \leftarrow \text{Dec}_i(Z_i, \hat{X}_{\mathcal{B}_i})$   
  **end if**  
**end for**  
**return**  $\hat{X} = (\hat{X}_1, \dots, \hat{X}_d)$

---

exogenous nodes  $Z_i$  as the input of the decoder of the diffusion model, we include the nodes which meet the backdoor criterion  $X_{\mathcal{B}_i}$  and the corresponding exogenous nodes  $Z_i$  and also include the intervened node  $X_j$  if it is the child of the intervened node ( $X_j \in X_{Pa_i}$ ). Furthermore, we change the training process accordingly. As the parent nodes of the outcome node always satisfy the backdoor criterion under Assumption 6 [3], including the nodes that meet the backdoor criterion instead of the parent nodes in the decoder of BDCM is the generalized algorithm of DCM. Algorithms 4 and 5 show the training and sampling process of BDCM. Then, we have the following conjecture.

**Conjecture 9** (Applicability of BDCM). Suppose sets of nodes satisfy the backdoor criterion for the intervened node and other nodes. In that case, we can generalize DCM to BDCM to sample from the target distribution even if Assumption 6 is violated.

### B. Experiment

To show that BDCM precisely samples from the target distribution where we cannot use DCM, we conduct an empirical analysis with the following settings where causal sufficiency does not hold. Python code for the experiment is available in <https://github.com/tatsu432/BDCM>.

Fig. 4 and Fig. 5 show the SCMs  $\mathcal{M}_1$  and  $\mathcal{M}_2$  respectively that do not satisfy Assumption 6 where  $X_1$  and  $X_4$  in Fig. 4 and  $X_2$  in Fig. 5 are the unobserved nodes. Note that we did not show the exogenous nodes in the figures for clarity. Examples 10 and 11 show the concrete structural equations for  $\mathcal{M}_1$  and Examples 12 and 13 for  $\mathcal{M}_2$ . We create simple and complex structural equations for both cases. The simple cases are the *additive noise models* (ANM) [20], [21], [22], [12] whereas the complex ones are not ANM.

**Example 10.** We define the set of simple structural equations

$f = \{f_i\}_{i \in [5]}$  for SCM  $\mathcal{M}_1$  in Fig. 4 as follows.

$$\begin{aligned} X_1 &= f_1(U_1) = U_1 \\ X_2 &= f_2(X_1, U_2) = X_1^2 + U_2 \\ X_3 &= f_3(X_1, U_3) = 2X_1 + U_3 \\ X_4 &= f_4(X_3, U_4) = X_3 + U_4 \\ X_5 &= f_5(X_2, X_4, U_5) = X_2 + 2X_4 + U_5 \end{aligned}$$

**Example 11.** We define the set of complex structural equations  $f = \{f_i\}_{i \in [5]}$  for SCM  $\mathcal{M}_1$  in Fig. 4 as follows.

$$\begin{aligned} X_1 &= f_1(U_1) = U_1 \\ X_2 &= f_2(X_1, U_2) = \frac{\sqrt{|X_1|}(|U_2| + 0.1)}{2} + |X_1| + \frac{U_2}{5} \\ X_3 &= f_3(X_1, U_3) = \frac{1}{1 + (|U_3| + 0.1) \exp(-X_2)} \\ X_4 &= f_4(X_3, U_4) = X_3 + X_3 U_4 + U_4 \\ X_5 &= f_5(X_2, X_4, U_5) = X_2 + X_4 + X_2 X_4 U_5 + U_5 \end{aligned}$$

**Example 12.** We define the set of simple structural equations  $f = \{f_i\}_{i \in [6]}$  for SCM  $\mathcal{M}_2$  in Fig. 5 as follows.

$$\begin{aligned} X_1 &= f_1(U_1) = U_1 \\ X_2 &= f_2(X_1, U_2) = X_1^2 + U_2 \\ X_3 &= f_3(X_2, U_3) = X_2 + U_3 \\ X_4 &= f_4(X_3, U_4) = X_3^3 + X_3 + U_4 \\ X_5 &= f_5(X_3, U_5) = X_3^2 + 0.1 + U_5 \\ X_6 &= f_6(X_2, X_4, X_5, U_6) = X_2 X_4 + X_2 X_5 + X_4 X_5 + U_6 \end{aligned}$$

**Example 13.** We define the set of complex structural equations  $f = \{f_i\}_{i \in [6]}$  for SCM  $\mathcal{M}_2$  in Fig. 5 as follows.

$$\begin{aligned} X_1 &= f_1(U_1) = U_1 \\ X_2 &= f_2(X_1, U_2) = \frac{\sqrt{|X_1|}(|U_2| + 0.1)}{2} + |X_1| + \frac{U_2}{5} \\ X_3 &= f_3(X_2, U_3) = \frac{1}{1 + (|U_3| + 0.1) \exp(-X_2)} \\ X_4 &= f_4(X_3, U_4) = \frac{U_4(|X_3| + 0.3)}{5} + U_4 \\ X_5 &= f_5(X_3, U_5) = \frac{1}{\sqrt{|U_5 X_3|} + 0.1} + U_5 \\ X_6 &= f_6(X_2, X_4, X_5, U_6) \\ &= X_2^2 X_4 + X_2 X_5 + X_5 X_6 + X_2 U_6 \end{aligned}$$

For Examples 10, 11, 12, and 13, we sample exogenous nodes  $U_i$  from standard normal distribution  $\mathcal{N}(0, 1)$  for all  $i \in [5]$  in  $\mathcal{M}_1$  and all  $i \in [6]$  in  $\mathcal{M}_2$ . We normalized each endogenous variable as [4] did.

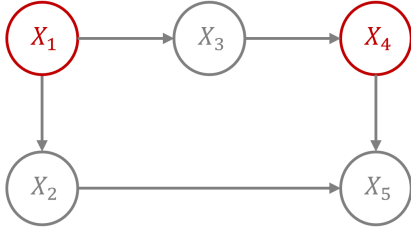


Fig. 4. SCM  $\mathcal{M}_1$  where the unobserved confounders  $X_1$  and  $X_4$  exist with five exogenous and endogenous nodes where we intervene in the node  $X_2 = x_2$

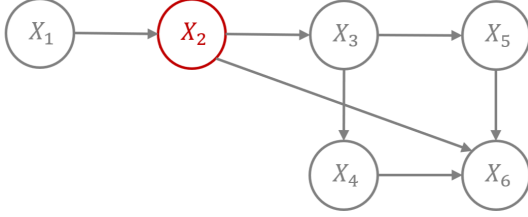


Fig. 5. SCM  $\mathcal{M}_2$  where the unobserved confounder  $X_2$  exists with six exogenous and endogenous nodes where we intervene in the node  $X_4 = x_4$

For both Examples 10 and 11 for Fig. 4, we aim to sample correctly from the target distribution  $\nu(X_5|\text{do}(X_2 = x_2))$  where  $X_2$  is the cause, and  $X_5$  is the outcome. For both DCM and BDCM, we set the intervened node  $X_2$  to intervened value  $x_2$  and sample  $X_3$  from the empirical distribution  $E_3$ . For the node of our interest  $X_5$ , DCM takes  $\hat{X}_2$  as the input for the decoder  $\text{Dec}_5(Z_5, \hat{X}_2)$  whereas BDCM takes  $\hat{X}_2$  and  $\hat{X}_3$  as the input for the decoder  $\text{Dec}_5(Z_5, \hat{X}_2, \hat{X}_3)$ .

For both Examples 12 and 13 for Fig. 5, we aim to sample correctly from the target distribution  $\nu(X_6|\text{do}(X_4 = x_4))$  where  $X_4$  is the cause, and  $X_6$  is the outcome. For both DCM and BDCM, we set the intervened node  $X_4$  to intervened value  $x_4$ , sample  $X_1$  and  $X_3$  from the empirical distribution  $E_1$  and  $E_3$  respectively, and sample  $X_5$  by the decoder  $\text{Dec}_5(Z_5, \hat{X}_3)$ . For the node of our interest  $X_6$ , DCM takes  $\hat{X}_4$  and  $\hat{X}_5$  as the inputs for the decoder  $\text{Dec}_6(Z_6, \hat{X}_4, \hat{X}_5)$  whereas BDCM takes  $\hat{X}_3$  and  $\hat{X}_4$  as the inputs for the decoder  $\text{Dec}_6(Z_6, \hat{X}_3, \hat{X}_4)$ .

For parameters in the algorithm, we set them to the following values, mostly the same as [4]. For the noise schedule  $\beta_t$  and  $\alpha_t$ , we set them to  $\beta_t = (0.1 - 10^{-4}) \frac{t-1}{T-1} + 10^{-4}$  and  $\alpha_t = \prod_{i=1}^t (1 - \beta_i)$  where we set  $T = 100$ . For the neural networks, we set the epochs to 500, batch size to 64, and learning rate to  $10^{-4}$  where each neural network consists of three hidden layers whose numbers of nodes are 128, 256, and 256 for the first, second and third layers, respectively. We extract 500 samples via DCM and BDCM, where we train them with 1000 samples. We calculate the Maximum Mean Discrepancy (MMD) [23] between the empirical distributions obtained from the algorithms and the ground truth target for both DCM and BDCM. Note that the lower MMD is, the closer the empirical distributions are, so the algorithm is more precise. We set the intervened values to ten different values sampled randomly from  $\text{Unif}(-3, 3)$ . We also conduct the simulation for

TABLE I  
AVERAGE  $\pm$  STANDARD DEVIATION OF MMD ( $\times 10^{-3}$ ) OF DCM AND BDCM COMPARED TO THE TRUE TARGET DISTRIBUTION IN EXAMPLES 10, 11, 12, 13, 14, 15, 16, 17, 18, 19

		BDCM (ours)	DCM
SCM $\mathcal{M}_1$	Example 10	<b><math>1.24 \pm 0.744</math></b>	$1.79 \pm 1.54$
	Example 11	<b><math>1.04 \pm 0.835</math></b>	$2.34 \pm 2.17$
SCM $\mathcal{M}_2$	Example 12	$5.08 \pm 2.51$	$5.07 \pm 2.17$
	Example 13	<b><math>1.55 \pm 1.91</math></b>	$2.89 \pm 2.08$
SCM $\mathcal{M}_3$	Example 14	<b><math>0.741 \pm 0.68</math></b>	$1.14 \pm 1.26$
	Example 15	<b><math>1.51 \pm 1.43</math></b>	$1.8 \pm 1.55$
SCM $\mathcal{M}_4$	Example 16	<b><math>1.69 \pm 1.49</math></b>	$2.12 \pm 1.34$
	Example 17	<b><math>1.46 \pm 1.11</math></b>	$2.38 \pm 1.81$
SCM $\mathcal{M}_5$	Example 18	<b><math>0.638 \pm 0.586</math></b>	$0.747 \pm 0.575$
	Example 19	<b><math>1.29 \pm 0.938</math></b>	$1.41 \pm 0.591$

five different seeds. Then, We output the average and standard deviation of MMDs.

Table I shows the results of the experiments. Table I demonstrates that BDCM output a more precise distribution than DCM, where unmeasured confounders exist for Examples 10, 11, 13. For Example 12, BDCM is almost as accurate as DCM. For both SCMs  $\mathcal{M}_1$  and  $\mathcal{M}_2$ , the more complex the structural equations become in SCM, the clear the difference in the performance between DCM and BDCM is. For SCM  $\mathcal{M}_1$  in Fig. 4, BDCM successfully considers the backdoor path  $X_2 \leftarrow X_1 \rightarrow X_3 \rightarrow X_4 \rightarrow X_5$  by including the node  $X_3$  that blocks the backdoor path in the decoder of the outcome meanwhile DCM does not consider this path when we sample the outcome  $X_5$  where we intervene in the node  $X_2$ , which creates the bias. Furthermore, for SCM  $\mathcal{M}_2$  in Fig. 5, BDCM carefully chooses the nodes  $X_3$  and  $X_4$  that block all the backdoor paths concerning the pair of the cause and outcome nodes as the input for the decoder of the outcome  $X_6$  of our interest. In contrast, DCM takes the parent nodes of the outcome we observe  $X_4$  and  $X_5$  without considering one of the backdoor paths:  $X_4 \leftarrow X_3 \leftarrow X_2 \rightarrow X_6$ , which incurs the bias in the sample by DCM.

Furthermore, Fig. 6 and Fig. 11 show ones of the empirical distributions sampled by DCM, BDCM, and target distribution for SCMs  $\mathcal{M}_1$  and  $\mathcal{M}_2$  where the structural equations are complex. The blue histograms are the ground truth distribution we want to sample from, whereas the red histograms are the outputs of the DCM (left) and BDCM (right). From Fig. 6 and Fig. 7, we can see that BDCM can sample from the target distribution  $\nu(X_5|\text{do}(X_2 = x_2))$  in  $\mathcal{M}_1$  and  $\nu(X_6|\text{do}(X_4 = x_4))$  in  $\mathcal{M}_2$  precisely where unmeasured confounders exist whereas DCM fails to do so.

## V. CONCLUSION AND FUTURE WORK

We extended the Diffusion-based causal Model (DCM) proposed by [4] to the case where unmeasured confounders exist. We proposed Backdoor Criterion-based DCM (BDCM) that can consider the unobserved confounders by including

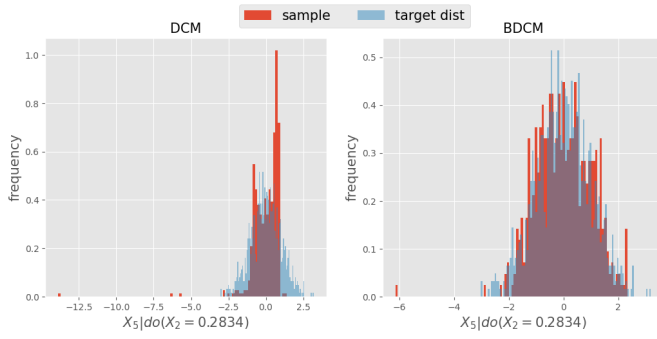


Fig. 6. Empirical distributions of the  $X_5$  sampled from DCM (left) and BDCM (right) compared to the ground-truth target distribution where we intervened in the node  $X_2 = 0.2834$  in Example 10

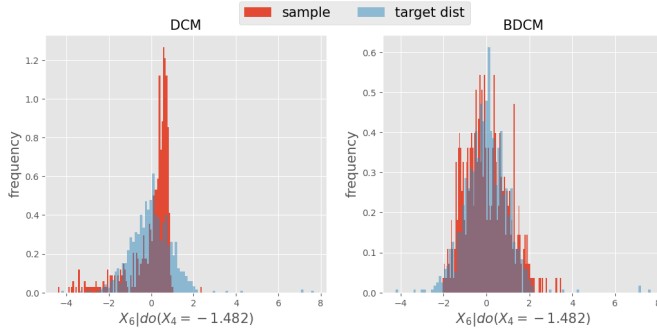


Fig. 7. Empirical distributions of the  $X_6$  sampled from DCM (left) and BDCM (right) compared to the ground-truth target distribution where we intervened in the node  $X_4 = -1.482$  in Example 12

the nodes that meet the backdoor criterion [3]. Synthetic data experiment demonstrates that BDCM can precisely sample from the target distribution of our interest where DCM fails to do so.

For future work, one of the intriguing topics would be to derive the convergence guarantee of BDCM. Implementing the comprehensive algorithm of BDCM in Python would also be interesting. Moreover, it would be intriguing to generalize BDCM using the Front-door criterion [3], another criterion to adjust the nodes where unobserved confounders exist.

#### ACKNOWLEDGEMENTS

We thank Dr. Andre Wibisono at Yale University Department of Computer Science for his help and guidance.

#### REFERENCES

- [1] P. Sanchez, J. P. Voisey, T. Xia, H. I. Watson, A. Q. O’Neil, and S. A. Tsafaris, “Causal machine learning for healthcare and precision medicine,” *Royal Society Open Science*, vol. 9, no. 8, p. 220638, 2022.
- [2] G. W. Imbens and D. B. Rubin, *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.
- [3] J. Pearl, M. Glymour, and N. P. Jewell, *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.
- [4] P. Chao, P. Blöbaum, and S. P. Kasiviswanathan, “Interventional and counterfactual inference with diffusion models,” *arXiv preprint arXiv:2302.00860*, 2023.
- [5] 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.
- [6] I. Khemakhem, R. Monti, R. Leech, and A. Hyvarinen, “Causal autoregressive flows,” in *International conference on artificial intelligence and statistics*. PMLR, 2021, pp. 3520–3528.
- [7] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*. MIT press, 2000.
- [8] C. Glymour, K. Zhang, and P. Spirtes, “Review of causal discovery methods based on graphical models,” *Frontiers in Genetics*, vol. 10, 2019.
- [9] D. M. Chickering, “Learning bayesian networks is np-complete,” *Learning from data: Artificial intelligence and statistics V*, pp. 121–130, 1996.
- [10] P. Rolland, V. Cevher, M. Kleindessner, C. Russell, D. Janzing, B. Schölkopf, and F. Locatello, “Score matching enables causal discovery of nonlinear additive noise models,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 18 741–18 753.
- [11] P. Sanchez, X. Liu, A. Q. O’Neil, and S. A. Tsafaris, “Diffusion models for causal discovery via topological ordering,” *arXiv preprint arXiv:2210.06201*, 2022.
- [12] P. Bühlmann, J. Peters, and J. Ernest, “Cam: Causal additive models, high-dimensional order search and penalized regression,” *Ann. Statist.*, 2014.
- [13] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” *arXiv preprint arXiv:2010.02502*, 2020.
- [14] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *International conference on machine learning*. PMLR, 2015, pp. 2256–2265.
- [15] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [16] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, “Diffwave: A versatile diffusion model for audio synthesis,” *arXiv preprint arXiv:2009.09761*, 2020.
- [17] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with clip latents,” *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.
- [18] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, “Photorealistic text-to-image diffusion models with deep language understanding,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 479–36 494, 2022.
- [19] C. Luo, “Understanding diffusion models: A unified perspective,” *arXiv preprint arXiv:2208.11970*, 2022.
- [20] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan, “A linear non-gaussian acyclic model for causal discovery,” *Journal of Machine Learning Research*, vol. 7, no. 10, 2006.
- [21] P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, “Nonlinear causal discovery with additive noise models,” *Advances in neural information processing systems*, vol. 21, 2008.
- [22] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf, “Causal discovery with continuous additive noise models,” *Journal of Machine Learning Research*, 2014.
- [23] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, “A kernel two-sample test,” *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.



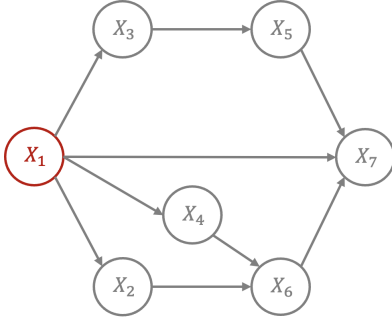


Fig. 8. SCM  $\mathcal{M}_3$  where the unobserved confounder  $X_1$  exists with seven exogenous and endogenous nodes where we intervene in the node  $X_6 = x_6$

## VI. DETAILS OF SYNTHETIC DATA EXPERIMENT

**Example 14.** We define the set of simple structural equations  $f = \{f_i\}_{i \in [7]}$  for SCM  $\mathcal{M}_3$  in Fig. 8 as follows.

$$\begin{aligned} X_1 &= f_1(U_1) = U_1 \\ X_2 &= f_2(X_1, U_2) = X_1^2 + U_2 \\ X_3 &= f_3(X_1, U_3) = X_1 + U_3 \\ X_4 &= f_4(X_1, U_4) = X_1^3 + X_1 + U_4 \\ X_5 &= f_5(X_3, U_5) = X_3^2 + 0.1 + U_5 \\ X_6 &= f_6(X_2, X_4, U_6) = X_2 X_4 + U_6 \\ X_7 &= f_7(X_1, X_5, X_6, U_7) = X_1 X_5 + X_6^2 + X_1 X_6 + U_7 \end{aligned}$$

**Example 15.** We define the set of complex structural equations  $f = \{f_i\}_{i \in [7]}$  for SCM  $\mathcal{M}_3$  in Fig. 8 as follows.

$$\begin{aligned} X_1 &= f_1(U_1) = U_1 \\ X_2 &= f_2(X_1, U_2) = \frac{\sqrt{|X_1|}(|U_2| + 0.1)}{2} + |X_1| + \frac{U_2}{5} \\ X_3 &= f_3(X_1, U_3) = \frac{1}{1 + (|U_3| + 0.1) \exp(-X_1)} \\ X_4 &= f_4(X_1, U_4) = \frac{U_4(|X_3| + 0.3)}{5} + U_4 \\ X_5 &= f_5(X_3, U_5) = \frac{1}{\sqrt{|U_5 X_3|} + 0.1} + U_5 \\ X_6 &= f_6(X_2, X_4, U_6) \\ &= X_2^2 X_4 + X_2 X_4 + X_2 U_6 \\ X_7 &= f_7(X_1, X_5, X_6, U_7) = \\ &= X_1^2 X_5 + X_1 X_6 + X_1 X_5 U_7 \end{aligned}$$

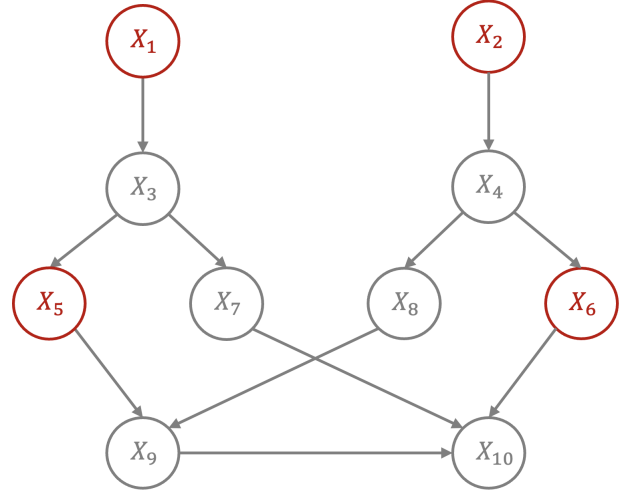


Fig. 9. SCM  $\mathcal{M}_4$  where the unobserved confounders  $X_1, X_2, X_5$ , and  $X_6$  exist with five exogenous and endogenous nodes where we intervene in the node  $X_9 = x_9$

**Example 16.** We define the set of simple structural equations  $f = \{f_i\}_{i \in [10]}$  for SCM  $\mathcal{M}_4$  in Fig. 9 as follows.

$$\begin{aligned} X_1 &= f_1(U_1) = U_1 \\ X_2 &= f_2(U_2) = U_2 \\ X_3 &= f_3(X_1, U_3) = X_1 + U_3 \\ X_4 &= f_4(X_2, U_4) = -X_2^3 + X_2 + U_4 \\ X_5 &= f_5(X_3, U_5) = X_3^2 + 0.1 + U_5 \\ X_6 &= f_6(X_4, U_6) = X_4^2 + X_4 + U_6 \\ X_7 &= f_7(X_3, U_7) = -X_3^2 + X_3 + U_7 \\ X_8 &= f_8(X_4, U_8) = 3X_4 + 0.1 + U_8 \\ X_9 &= f_9(X_5, X_8, U_9) = X_5 X_8 + X_5 + X_8 + U_9 \\ X_{10} &= f_{10}(X_6, X_7, X_9, U_{10}) = X_6 X_7 X_9 + X_6 X_7 + U_{10} \end{aligned}$$

**Example 17.** We define the set of complex structural equations

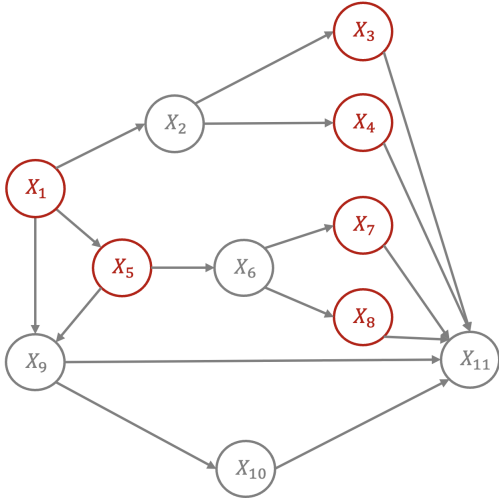


Fig. 10. SCM  $\mathcal{M}_5$  where the unobserved confounders  $X_1, X_3, X_4, X_5, X_7$ , and  $X_8$  exist with five exogenous and endogenous nodes where we intervene in the node  $X_9 = x_9$

$f = \{f_i\}_{i \in [10]}$  for SCM  $\mathcal{M}_4$  in Fig. 9 as follows.

$$\begin{aligned}
X_1 &= f_1(U_1) = U_1 \\
X_2 &= f_2(U_2) = U_2 \\
X_3 &= f_3(X_1, U_3) = \frac{\sqrt{|X_1|}(|U_3| + 0.1)}{2} + |X_1| + \frac{U_3}{5} \\
X_4 &= f_4(X_2, U_4) = \frac{U_4(|X_2| + 0.3)}{5} + U_4 \\
X_5 &= f_5(X_3, U_5) = -\frac{1}{1 + (|U_5| + 0.1) \exp(-X_3)} \\
X_6 &= f_6(X_4, U_6) = \frac{U_6(|X_4| + 0.3)}{5} + U_6 \\
X_7 &= f_7(X_3, U_7) = \frac{\sqrt{|X_3|}(|U_7| + 0.1)}{2} + |X_3| + \frac{U_7}{5} \\
X_8 &= f_8(X_4, U_8) = 3X_4 + 0.1 + U_8 \\
X_9 &= f_9(X_5, X_8, U_9) = X_5^2 X_8 + X_5 + X_8 + U_9 \\
X_{10} &= f_{10}(X_6, X_7, X_9, U_{10}) = X_6^2 X_7 X_9 + X_6 X_7 + U_{10}
\end{aligned}$$

**Example 18.** We define the set of simple structural equations

$f = \{f_i\}_{i \in [11]}$  for SCM  $\mathcal{M}_5$  in Fig. 10 as follows.

$$\begin{aligned}
X_1 &= f_1(U_1) = U_1 \\
X_2 &= f_2(X_1, U_2) = -X_1 + U_2 \\
X_3 &= f_3(X_2, U_3) = X_2 + 0.1 + U_3 \\
X_4 &= f_4(X_2, U_4) = -X_2 + 0.1 + U_4 \\
X_5 &= f_5(X_1, U_5) = 1.3X_1 + X_1 U_5 + U_5 \\
X_6 &= f_6(X_5, U_6) = -1.2(X_5 + 0.1) + X_5 + U_6 \\
X_7 &= f_7(X_6, U_7) = -X_6^2 + X_6 + U_7 \\
X_8 &= f_8(X_6, U_8) = 3X_6 + 0.1 + U_8 \\
X_9 &= f_9(X_1, X_6, U_9) = X_1 X_5 + X_1 - X_5^2 + 0.1 + U_9 \\
X_{10} &= f_{10}(X_9, U_{10}) = X_9^2 + U_{10} \\
X_{11} &= f_{11}(X_3, X_4, X_7, X_8, X_9, X_{10}, U_{11}) \\
&= X_3 X_4 + X_7 X_8 + X_9 X_{10} + X_3 X_9 - X_7 X_{10} - 0.1
\end{aligned}$$

**Example 19.** We define the set of complex structural equations  $f = \{f_i\}_{i \in [11]}$  for SCM  $\mathcal{M}_5$  in Fig. 10 as follows.

$$\begin{aligned}
X_1 &= f_1(U_1) = U_1 \\
X_2 &= f_2(X_1, U_2) = X_1(U_2 + 0.1) \\
X_3 &= f_3(X_2, U_3) = \frac{\sqrt{|X_2|}(|U_3| + 0.1)}{2} + |X_2| + \frac{U_3}{5} \\
X_4 &= f_4(X_2, U_4) = X_2 + \frac{U_4 + 0.1}{2} X_2 \\
X_5 &= f_5(X_1, U_5) = -\frac{1}{1 + (|U_5| + 0.1) \exp(-X_1)} \\
X_6 &= f_6(X_5, U_6) = \frac{U_6(|X_5| + 0.3)}{5} + U_6 \\
X_7 &= f_7(X_6, U_7) = X_6 U_7 + |X_6 + 0.01| |U_7| \\
X_8 &= f_8(X_6, U_8) = 3X_6 + 0.1 + U_8 \\
X_9 &= f_9(X_1, X_6, U_9) = X_5^3 X_8 + X_5 + X_8 + U_9 \\
X_{10} &= f_{10}(X_9, U_{10}) = X_9 U_{10} + (U_{10} + 0.1)^2 \\
X_{11} &= f_{11}(X_3, X_4, X_7, X_8, X_9, X_{10}, U_{11}) \\
&= X_3(X_8 - 0.1) + X_9 X_{10} + X_3 X_9 - X_7 X_{10} + X_3 X_8 \\
&\quad - X_4 X_9 + X_9 X_{10}
\end{aligned}$$

## VII. ADDITIONAL RESULT OF SYNTHETIC DATA EXPERIMENT



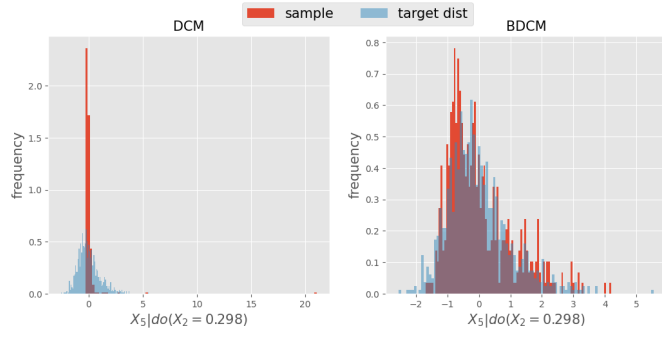


Fig. 11. Empirical distributions of the  $X_5$  sampled from DCM (left) and BDCM (right) compared to the ground-truth target distribution where we intervened in the node  $X_2 = 0.298$  in Example 11

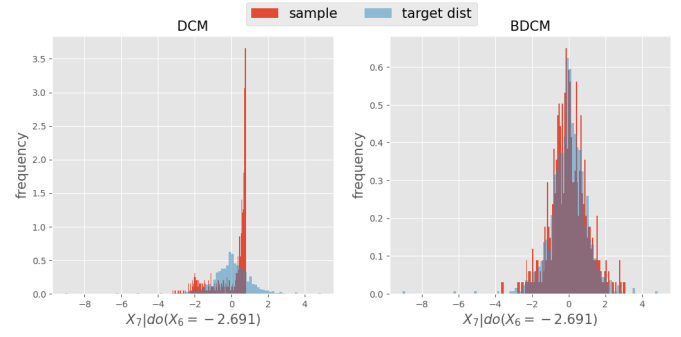


Fig. 14. Empirical distributions of the  $X_7$  sampled from DCM (left) and BDCM (right) compared to the ground-truth target distribution where we intervened in the node  $X_6 = -2.691$  in Example 15

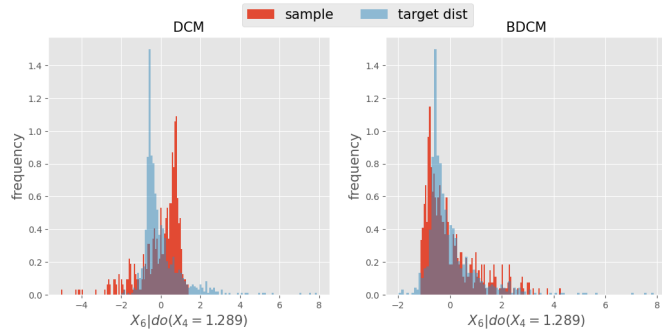


Fig. 12. Empirical distributions of the  $X_6$  sampled from DCM (left) and BDCM (right) compared to the ground-truth target distribution where we intervened in the node  $X_4 = 1.289$  in Example 13

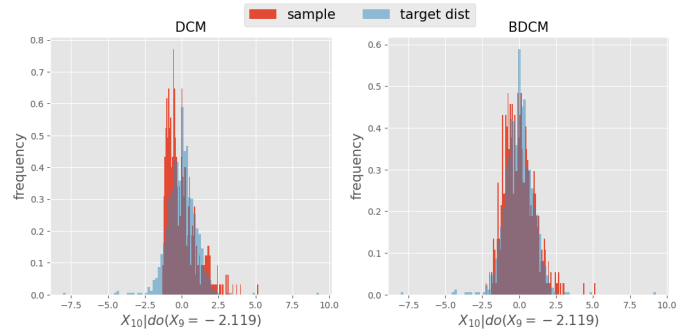


Fig. 15. Empirical distributions of the  $X_{10}$  sampled from DCM (left) and BDCM (right) compared to the ground-truth target distribution where we intervened in the node  $X_9 = -2.119$  in Example 16

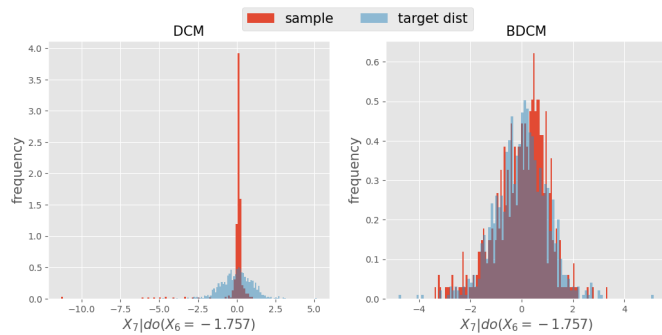


Fig. 13. Empirical distributions of the  $X_7$  sampled from DCM (left) and BDCM (right) compared to the ground-truth target distribution where we intervened in the node  $X_6 = -1.757$  in Example 14

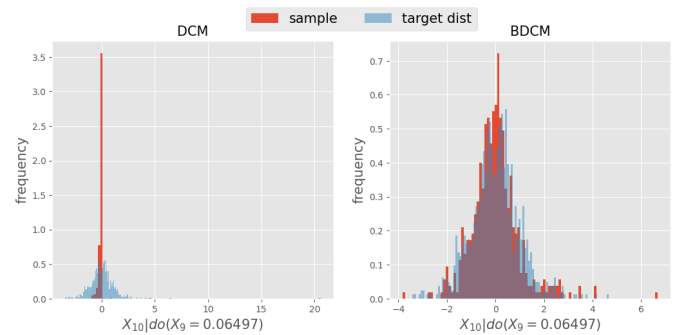


Fig. 16. Empirical distributions of the  $X_{10}$  sampled from DCM (left) and BDCM (right) compared to the ground-truth target distribution where we intervened in the node  $X_9 = 0.06497$  in Example 17

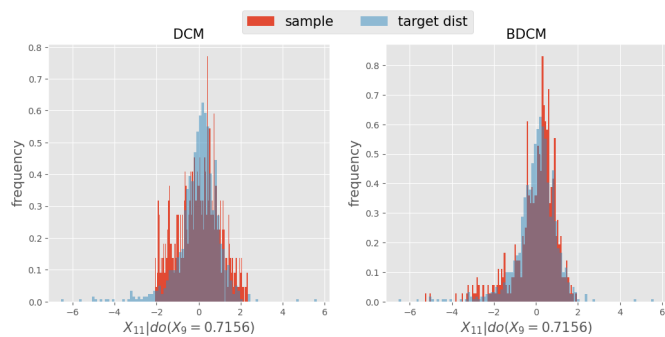


Fig. 17. Empirical distributions of the  $X_{11}$  sampled from DCM (left) and BDCM (right) compared to the ground-truth target distribution where we intervened in the node  $X_9 = 0.7156$  in Example 18

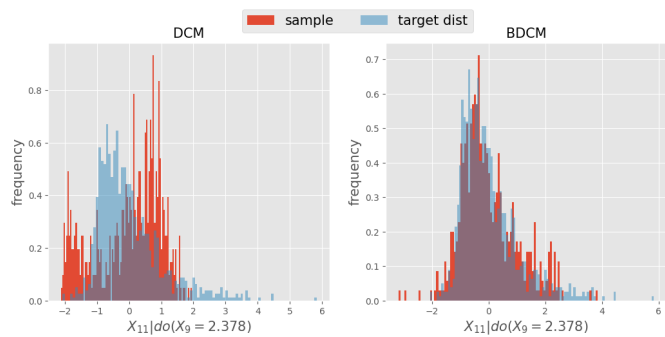


Fig. 18. Empirical distributions of the  $X_{11}$  sampled from DCM (left) and BDCM (right) compared to the ground-truth target distribution where we intervened in the node  $X_9 = 2.378$  in Example 19