Diffusion Model in Causal Inference

CPSC 486 Probabilistic Machine Learning
Final Project
Tatsuhiro Shimizu

Agenda

1. Setting of the problem
    1. Two mainstream of Causal Inference (Pearl and Rubin)
    2. Target distribution we want to sample from
2. Literature Review
    1. SCORE (Rolland et al 2022) and DiffAN (Sanchez et al 2022)
    2. DCM (Chao et al 2023)
3. Result of New Algorithm
    1. Backdoor DCM
    2. Experiment
4. Discussion and Future Work
    1. Contribution
    2. limitations
    3. Future work

Setting of the Problem

Two mainstreams of Causal Inference

- Pearl (2016)
  - Directed Acyclic Graph (DAG) framework
  - Create a model by graph
  - Analyze the causal effect on the graph
- Imbens and Rubin (2015)
  - Potential Outcome framework

We follow the Pearl's framework

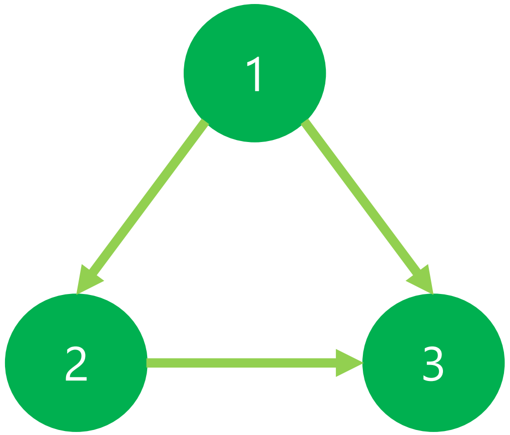
Definition of DAG and SCM

- DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
  - ➢Express variables by nodes and causal relationships by edges
  - ➢Set of nodes: $\mathcal{V} = \{1, 2, \cdots, d\}$
  - ➢Set of edges: $\mathcal{E} = \{(i, j) : \exists \text{ edge from node } i\ to\ j\}$

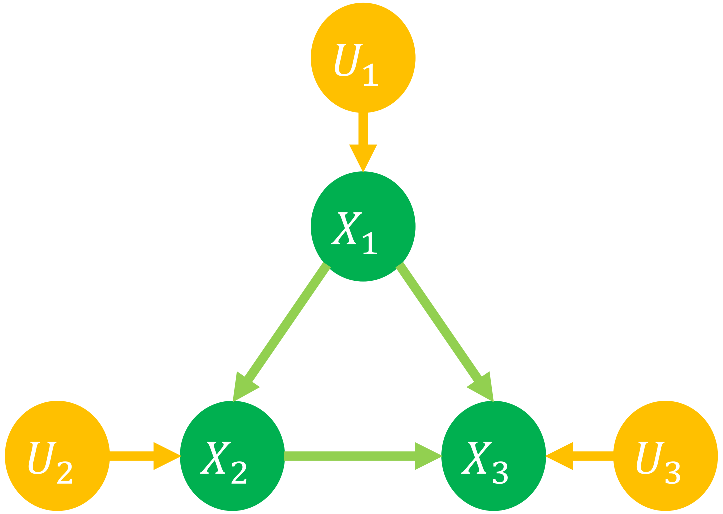
- Structural Causal Model (SCM) $\mathcal{M} = (\mathcal{U}, \mathcal{V}, f)$
  - ➢For each node in DAG $\mathcal{G}$, we have corresponding variable X_i $(\forall i \in [d])$
  - ➢Set of exogenous variables $\mathcal{U} = \{U_1, \cdots, U_d\}$
  - ➢Set of endogenous variables $\mathcal{V} = \{X_1, \cdots, X_d\}$
  - ➢Set of structural equations $f = \{f_1, \cdots, f_d\}$
    - $X_i = f_i(Pa(X_i), U_i)$ $(\forall i \in [d])$

Example of DAG and SCM

- DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
  - $\mathcal{V} = \{1, 2, 3\}$
  - $\mathcal{E} = \{(1,2), (1,3), (2,3)\}$



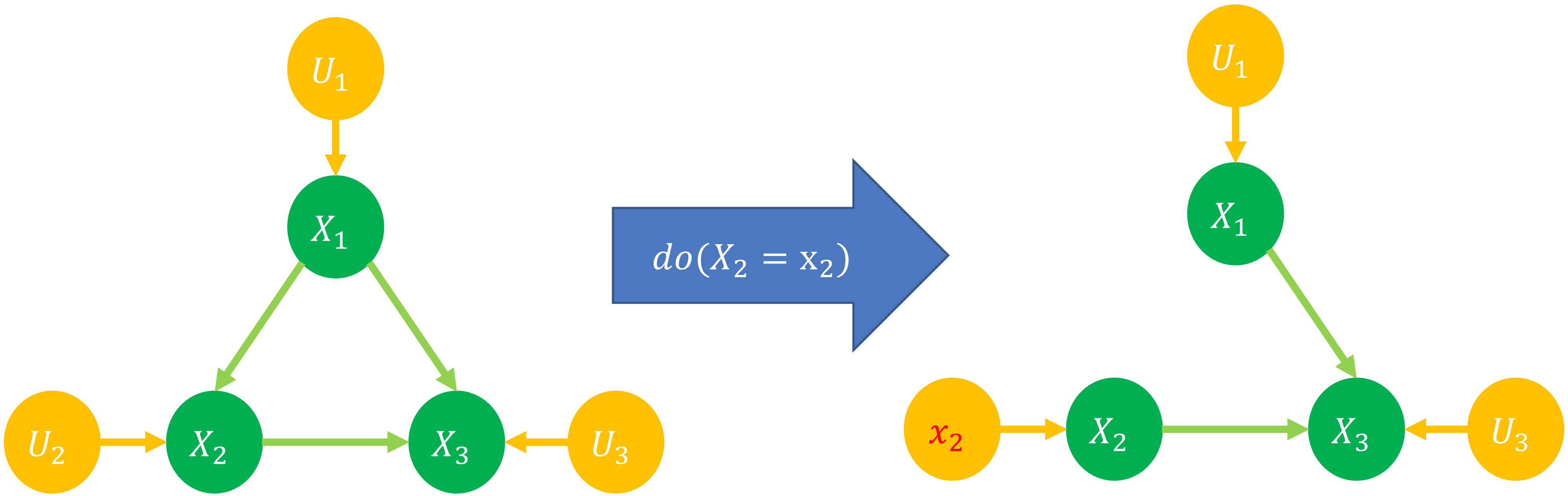
- SCM $\mathcal{M} = (\mathcal{U}, \mathcal{V}, f)$
  - $\mathcal{U} = \{U_1, U_2, U_3\}$
  - $\mathcal{V} = \{X_1, X_2, X_3\}$
  - $f = \{f_1, f_2, f_3\}$



$$X_1 = f_1(U_1) = U_1$$
$$X_2 = f_2(X_1, U_2) = 2X_1 + U_2$$
$$X_3 = f_3(X_1, X_2, U_3) = 3X_1 + 4X_2^3 + U_3$$

Definition of do operator

- $do(X_i = x_i)$: intervene the node X_i by setting it to x_i in SCM
  - ➢ Set the exogenous variable to the intervened value $U_i = x_i$
  - ➢ Delete all the edges into X_i from the endogenous variables



Definition of Causal Effect ATE and Goal

- Given observational data $X \in \mathbb{R}^{n \times d}$
  - n samples and d nodes

- Want to find the Average Treatment Effect (ATE) of $X_i = x_i$ versus $X_i = 0$ on X_j $(\forall i, j \in \mathcal{V}, i \neq j)$
  - $ATE(x_i, 0) := \mathbb{E}\left[x_j \middle| do(X_i = x_i)\right] - \mathbb{E}\left[x_j \middle| do(X_i = 0)\right]$
    $= \int_{x_j} x_j \, \nu\left(X_j = x_j \middle| do(X_i = x_i)\right) dx_j - \int_{x_j} x_j \, \nu\left(X_j = x_j \middle| do(X_i = 0)\right) dx_j$
  - $\nu\left(X_j = x_j \middle| do(X_i = x_i)\right)$ is the probability density function of X_j after the surgery on the graph by do operator $do(X_i = x_i)$
- Goal:
  - Sample $X_j | do(X_i = x_i)$ from the target distribution $\nu\left(X_j = x_j \middle| do(X_i = x_i)\right)$

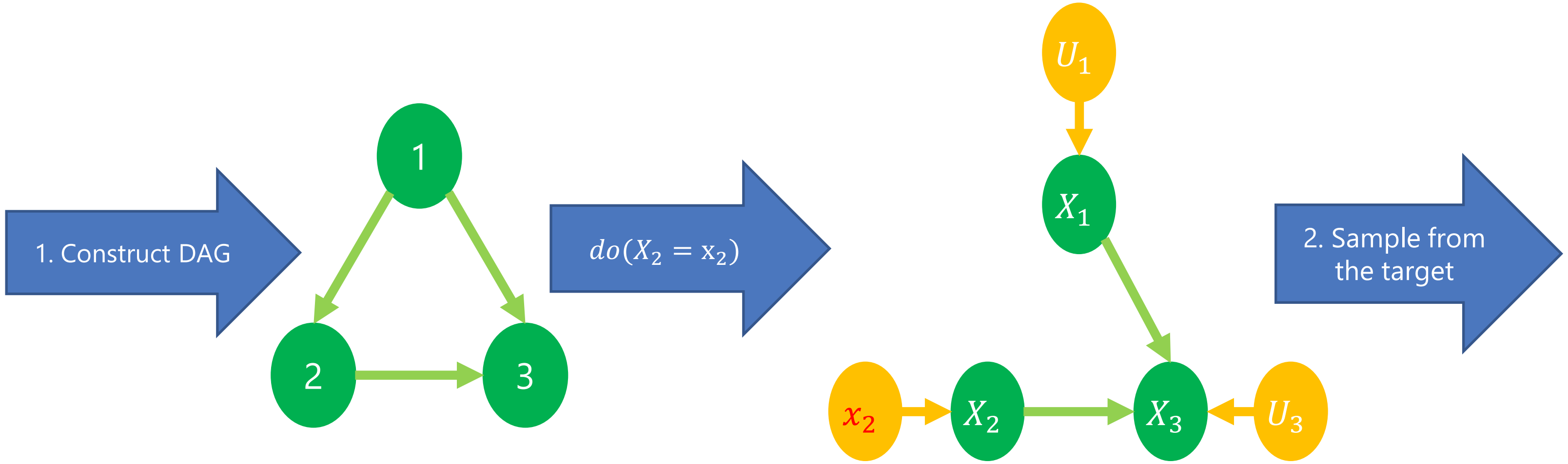
Literature Review

Two main problems

1. Given data $X \in \mathbb{R}^{n \times d}$, construct DAG

2. Given data $X \in \mathbb{R}^{n \times d}$ and DAG, sample from the target

X_1	X_2	X_3
2	4	4
6	8	9
0	9	8
2	28	82
91	42	49
49	0	3
19	32	93



(X_1, X_2, X_3)
$\sim \nu(X_1, X_2, X_3 | do(X_2 = x_2))$

Two main problems

1. Given data $X \in \mathbb{R}^{n \times d}$, how to construct DAG
   - SCORE (Rolland et al 2022): use score matching under Assumption 1 and 2
   - DiffAN (Sanchez et al 2022): use diffusion model to speed up under Assumption 1 and 2

2. Given data $X \in \mathbb{R}^{n \times d}$ and DAG, how to sample from the target distribution $\nu(X|do(X_i = x_i))$ where $X = (X_1, \cdots, X_d)$
   - DCM (Chao et al 2023): under Assumption 1

Assumption 1. (Causal Sufficiency): There is no unmeasured confounders
Assumption 2. (Additive noise model): structural equations have form of $X_i = f_i\big(Pa(X_i)\big) + U_i \;\; \forall i \in [d]$

1. Given data $X \in \mathbb{R}^{n \times d}$, how to construct DAG

- If we know the topological order π of nodes in DAG, then there exists a known algorithm to construct DAG by pruning edges (Buhlmann et al. 2014)

- Topological order π is a permutation of d nodes such that $\pi_i < \pi_j \iff j \in De(X_i)$ where $De(X_i)$ is the descendant nodes of i

- Problem boils down to how to construct topological order π given data $X \in \mathbb{R}^{n \times d}$

1. Given data $X \in \mathbb{R}^{n\times d}$, how to construct topological order π

- Lemma
  - Variance of the (j, j) element of Jacobian of the score function is 0 if and only if j is the leaf node
  - $Var_X\left[H_{j,j}(\log v(X))\right] = 0 \iff j$ is the leaf node

1. Given data $X \in \mathbb{R}^{n \times d}$, how to construct topological order π

- SCORE (Rolland et al 2022)
  - ➢ Estimate the Jacobian of the score function by score, find the leaf, delete the corresponding column of the data and iteratively continue the same procedure to construct topological order π
  - ➢ Drawback: need to estimate the Hessian of log density at each step
  - ➢ Algorithm complexity: $\mathcal{O}(dn^3)$

Algorithm 1 SCORE-matching causal order search

Input: Data matrix $X \in \mathbb{R}^{n \times d}$.
Initialize $\pi = []$, nodes $= \{1, \ldots, d\}$
for $k = 1, \ldots, d$ **do**
  Estimate the score function $s_{nodes} = \nabla \log p_{nodes}$ (for example using Algorithm 1).
  Estimate $V_j = \mathrm{Var}_{X_{nodes}} \left[\frac{\partial s_j(X)}{\partial x_j} \right]$.
  $l \leftarrow$ nodes$[\arg\min_j V_j]$
  $\pi \leftarrow [l, \pi]$
  nodes $\leftarrow$ nodes $- \{l\}$
  Remove l-th column of X
end for
Get the final DAG by pruning the full DAG associated with the topological order π.

---

1. Given data $X \in \mathbb{R}^{n\times d}$, how to construct topological order π

- DiffAN (Sanchez et al 2022):
  - Use the diffusion Probabilistic model (Ho et al 2020) to estimate the score function
    - $H_{i,j}(\log \nu(x)) \approx \nabla_{\mathrm{i,j}}\epsilon_\theta(x,t)$
  - No need to calculate the hessian of log density at each iteration by updating the score function by using the formula for the residue
    - $\Delta_l(x,t) \approx \nabla_l \epsilon_\theta(x,t) \cdot \frac{\epsilon_\theta(x,t)}{\nabla_{l,l}\epsilon_\theta(x,t)}$
  - Algorithm complexity: $\mathcal{O}(n+d^2)$

Algorithm 1: Topological Ordering with DiffAN

Input: $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, trained diffusion model $\boldsymbol{\epsilon}_\theta$, ordering batch size k
$\pi = [], \Delta_\pi = \mathbf{0}^{k\times d}, \boldsymbol{M}_\pi = \mathbf{1}^{k\times d}, score = \boldsymbol{\epsilon}_\theta$
while $\|\pi\| \neq d$ **do**

$\boldsymbol{B} \xleftarrow{k} \boldsymbol{X}$	// Randomly sample a batch of k elements
$\boldsymbol{B} \leftarrow \boldsymbol{B} \circ \boldsymbol{M}_\pi$	// Mask removed leaves
$\Delta_\pi = \mathrm{Get}\Delta_\pi(score, \boldsymbol{B})$	// Sum of Equation 8 over π
$score = score(-\pi) + \Delta_\pi$	// Update score with residue
$leaf = \mathrm{GetLeaf}(score, \boldsymbol{B})$	// Equation 9
$\pi = [leaf, \pi]$	// Append leaf to ordered list
$\boldsymbol{M}_{:,leaf} = \mathbf{0}$	// Set discovered leaf to zero

end
Output: Topological order π

2. Given data $X \in \mathbb{R}^{n \times d}$ and DAG, sample from the target $\nu(X|do(X_i = x_i))$

- DCM (Diffusion-based Causal Model) (Chao et al 2023)
  - Train the parameter of DDIM (Song et al 2021) at each node
  - For root node X_i, sample by empirical distribution E_i
  - For intervened node X_i, set it to the intervened value γ_i
  - For other nodes X_i, sample by the revere diffusion $Dec_i(Z_i, X_{Pa_i})$ using the trained parameter with the parent nodes X_{Pa_i} and corresponding exogenous nodes $Z_i \sim \mathcal{N}(0,1)$ as the input

Algorithm 1 DCM Training

Input: Distribution Q, scale factors $\{\alpha_t\}_{t=1}^T$, causal DAG $\mathcal{G}$ with node i represented by X_i

```
while not converged do
    Sample X^0 ~ Q
    for i = 1, ..., K do
        t ~ Unif[{1, ..., T}]
        ε ~ N(0, I)
        Update parameters of node i's diffusion model ε_θ^i, by minimizing the following loss (based on (1)) :
        ‖ε − ε_θ^i(√α_t X_i^0 + √(1 − α_t) ε, X_{pa_i}^0, t)‖_2^2
    end for
end while
```

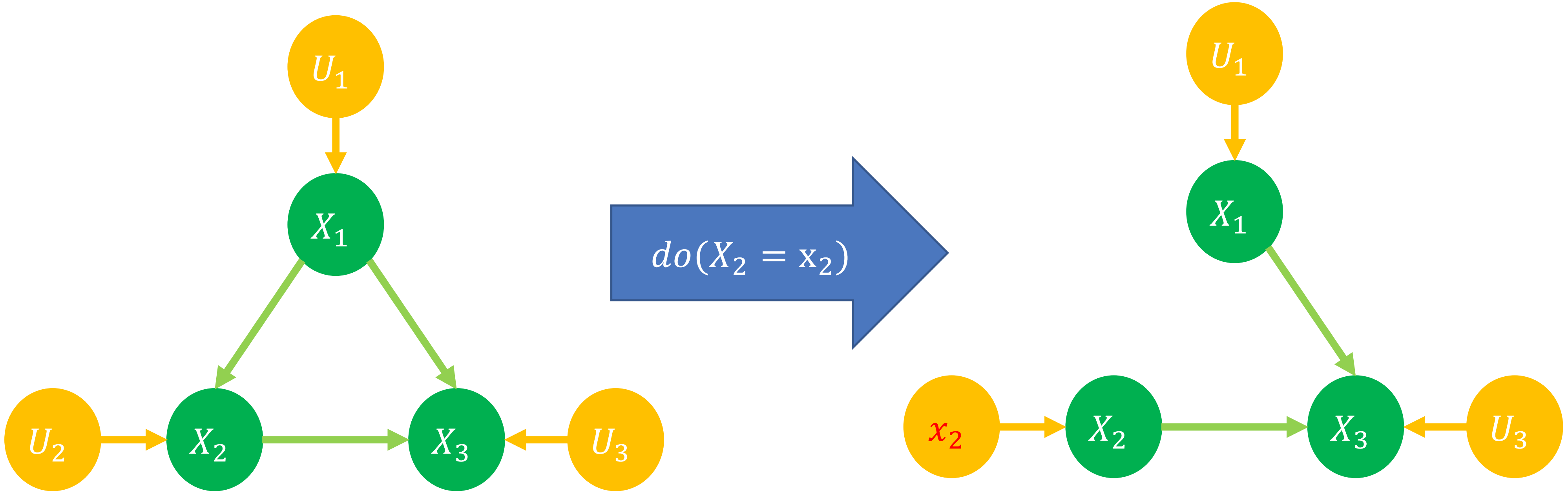
Algorithm 2 Observational/Interventional Sampling

Input: Intervention set $\mathcal{I}$ with values γ, optional noise $Z_i \sim \mathcal{N}(0, I_{d_i})$ for $i \in [K]$

```
for i = 1, ..., K do
    if i ∈ I then
        X̂_i = γ_i
    else if i is a root node then
        X̂_i ~ E_i, the empirical distribution
    else
        X̂_i ← Dec_i(Z_i, X̂_{pa_i})
    end if
end for
Return X̂
```

2. Given data $X \in \mathbb{R}^{n \times d}$ and DAG, sample from the target $\nu(X|do(X_i = x_i))$

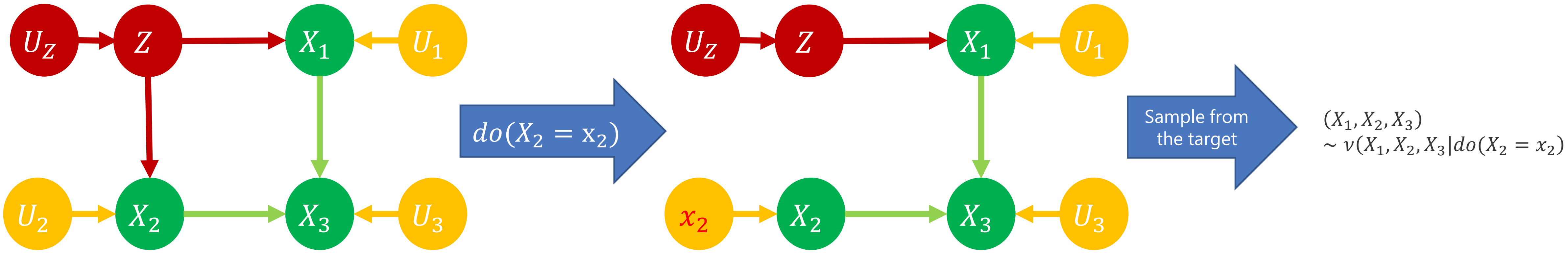
- Example of DCM (Chao et al 2023)
  - Want to sample from $\nu\left(X_1, X_2, X_3 | do(X_2 = x_2)\right)$
  - Root node $\mathrm{X}_1 \sim E_1$
  - Intervened node $X_2 = x_2$
  - Other nodes $X_3 = Dec_3(U_3, X_1, X_2)$



Result of New Algorithm

Q: How to find the causal effect where there exist unobserved confounders

- For the literature review, we assumed causal sufficiency (there is no unmeasured confounders)
- But in practice, we often cannot measure all confounders like Z



Q: How to find the causal effect where there exist unobserved confounders

- Definition of Backdoor criterion:
  - A set of variables $\mathcal{B}$ satisfies backdoor criterion with respect to (X, Y) in DAG $\mathcal{G}$ if
    1. No node in $\mathcal{B}$ is a descendant of X
    2. $\mathcal{B}$ blocks all paths between X and Y which contains an arrow into X
  - If there exist unmeasured confounders, then Backdoor criterion tells us which variable we should adjust for

Q: How to find the causal effect where there exist unobserved confounders

- BDCM (Backdoor Diffusion-based Causal Model)
  - Combine Backdoor criterion (Pearl 2016) and DCM (Chao et al 2023)
  - Backdoor DCM, for each node X_i, instead of having the parents X_{Pa_i} and corresponding exogenous nodes Z_i as the input of the decoder of the diffusion model, includes the nodes which meets the backdoor criterion $X_{\mathcal{B}_i}$ and the corresponding exogenous nodes Z_i and also includes the intervened node X_j if it is the child of the intervened node ($X_j \in X_{Pa_i}$).

Algorithm 5 BDCM Training

Input: target distribution ν, 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 γ_i
while not converge **do**
  Sample $X^0 \sim \nu$
  for $i = 1, \cdots, d$ **do**
    $t \sim \text{Unif}[\{1, \cdots, T\}]$
    $\epsilon \sim \mathcal{N}(0, 1)$
    Update the parameter of the node i's diffusion model ϵ_θ^i by minimization of the following loss function depending on the nodes.
    if $X_j \in X_{Pa_i}$ **then**

$$\left\|\epsilon - \epsilon_\theta^i\left(\sqrt{\alpha_t}X_i^0 + \sqrt{1-\alpha_t}\epsilon, X_{\mathcal{B}_i}^0, X_j, t\right)\right\|_2^2 \quad (8)$$

    else

$$\left\|\epsilon - \epsilon_\theta^i\left(\sqrt{\alpha_t}X_i^0 + \sqrt{1-\alpha_t}\epsilon, X_{\mathcal{B}_i}^0, t\right)\right\|_2^2 \quad (9)$$

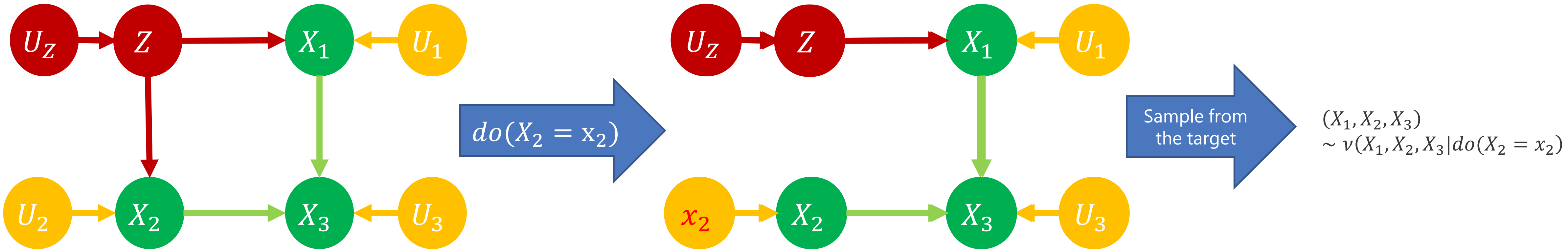
    end if
  end for
end while

Algorithm 6 BDCM Sampling

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

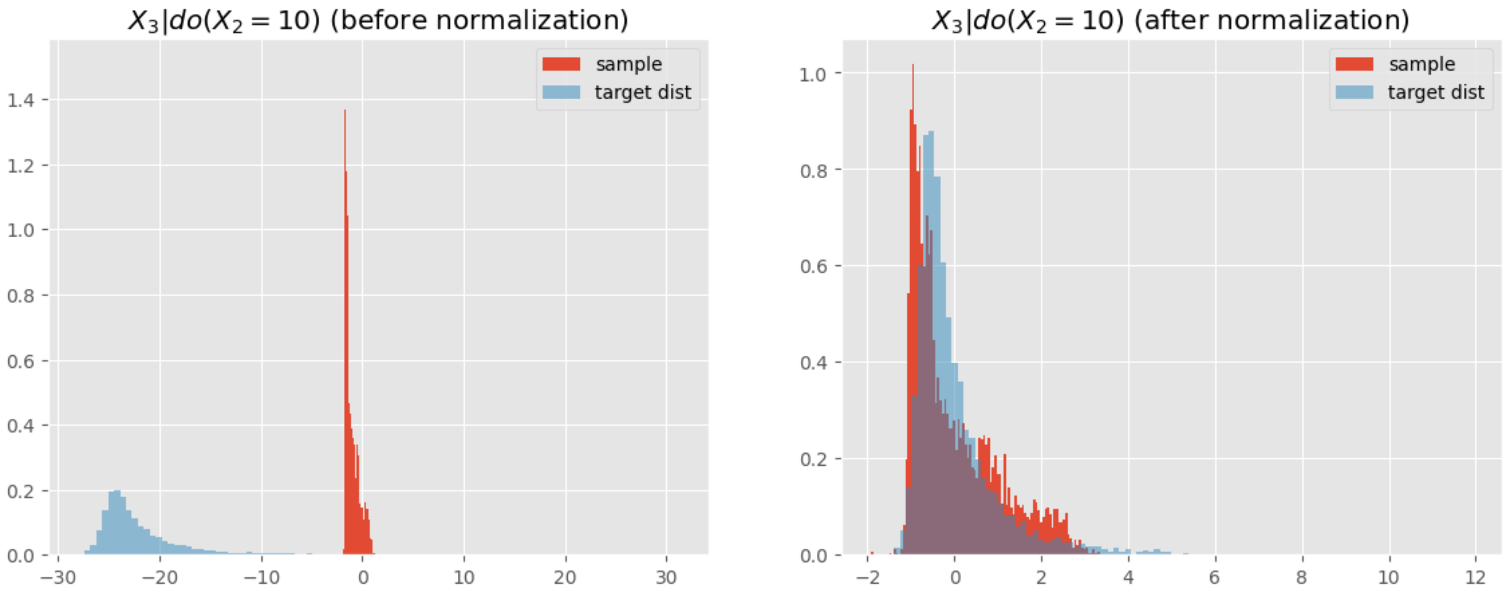
Q: How to find the causal effect where there exist unobserved confounders

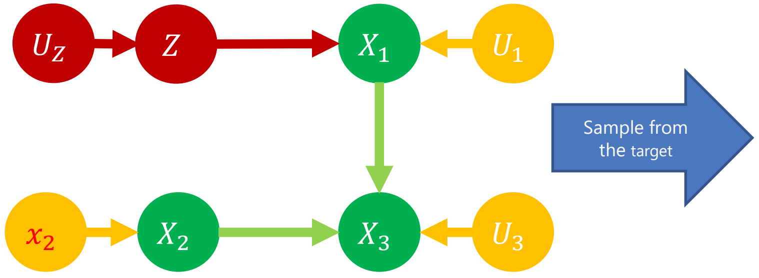
- Example of Backdoor DCM
  - ➢ Want to sample from $\nu\left(X_1, X_2, X_3 | do(X_2 = x_2)\right)$
  - ➢ Z is the unmeasured confounder
  - ➢ Root node $\mathrm{X}_1 \sim E_1$
  - ➢ Intervened node $X_2 = x_2$
  - ➢ Other nodes $X_3 = Dec_3(U_3, X_1, X_2)$



Q: How to find the causal effect where there exist unobserved confounders

- Experimental result





(X_1, X_2, X_3)
$\sim \nu(X_1, X_2, X_3|do(X_2 = x_2))$

Discussion and Future Work

Discussion and Future Work

- Contribution
  1. Combine and summarize the three recent literature on the usage of diffusion model to answer the causal effect
  2. Propose the new algorithm Backdoor DCM to sample from the target where there exist unobserved confounders
  3. Show that it works at least simple setting by experiment

- Limitations
  - Backdoor DCM works only when we normalize the target

- Future work
  - Prove the convergence guarantee of Backdoor DCM
  - Implement the comprehensive algorithm by Python
  - Compare DCM and Backoff DCM by the empirical analysis
  - Generalize more by using the Front-door criterion (Pearl 2016) as well, which is another criterion to adjust for the nodes where there exist unobserved confounders

References

- Guido W. Imbens and Donald B. Rubin. Causal Inference for Statistics, Social, and Biomedical Sciences. Cambridge University Press. 2015. https://doi.org/10.1017/CBO9781139025751
- Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. Causal Inference in Statistics: A Primer. 2016. Wiley.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. 2020. arXiv:2006.11239 [cs.LG]. https://doi.org/10.48550/arXiv.2006.11239.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. ICLR 2021. arXiv:2010.02502 [cs.LG]. https://doi.org/10.48550/arXiv.2010.02502.
- Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. Ann. Statist. 42(6): 2526-2556. 2014. DOI: 10.1214/14-AOS1260
- Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russel, Bernhard Schölkopf, Dominik Janzing, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. 2022. arXiv:2203.04413 [cs.LG]. https://doi.org/10.48550/arXiv.2203.04413.
- Pedro Sanchez, Xiao Liu, Alison Q O'Neil, and Sotirios A. Tsaftaris. Diffusion Models for Causal Discovery via Topological Ordering. 2022. ICLR 2023. arXiv:2210.06201 [cs.LG]. https://doi.org/10.48550/arXiv.2210.06201.
- Patrick Chao, Patrick Blöbaum, and Shiva Prasad Kasiviswanathan. Interventional and Counterfactual Inference with Diffusion Models. 2023. arXiv:2302.00860 [stat.ML]. https://doi.org/10.48550/arXiv.2302.00860.