Appendix for

"On Causal Discovery in the Presence of Deterministic Relations"

Appendix organization:

A1 RELATED WORKS

In this part, we will introduce more related works in causal discovery (Spirtes & Zhang, 2016). As we mentioned in the main paper, constraint-based and score-based methods are two primary categories in causal discovery. Constraint-based methods utilize the conditional independence test (CIT) to learn a skeleton of the directed acyclic graph (DAG), and then orient the edges upon the skeleton. Such methods contain Peter-Clark (PC) algorithm (Spirtes & Zhang, 2016) and Fast Causal Inference (FCI) algorithm (Spirtes, 2001). Some typical CIT methods include kernel-based independent conditional test (Zhang et al., 2012) and approximate kernel-based conditional independent test (Strobl et al., 2019).

Score-based methods normally use a score function and rely on a particular search strategy to look for the intended graph. The search strategy usually involve greedy search, exact search, or continuous optimization. The first continuous-optimization based method is NOTEARS (Zheng et al., 2018), which casts the Bayesian network structure learning task into a continuous constrained optimization problem with the least squares objective, using an algebraic characterization of directed acyclic graph (DAG). Subsequent work GOLEM (Ng et al., 2020) adopts a continuous unconstrained optimization formulation with a likelihood-based objective. NOTEARS is designed under the assumption of the linear relations between variables, therefore, another subsequent works have extended NOTEARS to handle nonlinear cases via deep neural networks, such as DAG-GNN (Yu et al., 2019) and DAG-NoCurl (Yu et al., 2021). Moreover, ENCO (Lippe et al., 2022) presents an efficient DAG discovery method for directed acyclic causal graphs utilizing both observational and interventional data. AVCI (Lorch et al., 2022) infers causal structure by performing amortized variational inference over an arbitrary data-generating distribution. These methods might suffer from various optimization issues, including convergence (Wei et al., 2020), sensitivity to data standardization (Reisach et al., 2021), and nonconvexity (Ng et al., 2023). Since they are only guaranteed to find a local optimum, therefore the quality of the solution can not be guaranteed, even in the asymptotic cases.

Besides the constrain-based and score-based methods, another major category of causal discovery methods is function causal model based methods. Those methods rely on the causal asymmetry property, such as the linear non-Gaussian model (LiNGAM) (Shimizu et al., 2006), the additive noise model (Hoyer et al., 2008), and the post-nonlinear causal model (Zhang & Hyvarinen, 2012). Apart from those methods, there are also some hybrid methods, such as neural conditional dependence (NCD) method, which reframes the GES algorithm to be more flexible than the standard score-based version and readily lends itself to the nonparametric setting with a general measure of conditional dependence.

A2 PROOFS

In this section, we provide the proofs of theorems and lemmas in the main paper, including Lemma 1, Theorem 3, Theorem 4, and Theorem 5.

A2.1 PROOF OF LEMMA 1

Proof: Assume a deterministic cluster $S = \{X\} \cup Y$, where X is any one variable in the DC, and Y is the set of the other deterministic variables in S.

By the definition of determinisic relation, we can have

$$f(X, Z) = 0, \tag{6}$$

In other words, we may also obtain $X = g(Z)$ or $Z = h(X)$ without a random noise term.

When we are doing conditional independent test to evaluate the null hypothesis $Y \perp\!\!\!\perp X|Z$, where Y is any one non-deterministic variable in the system, then a standard procedure is to conduct regression in RKHS (Zhang et al., 2012; Huang et al., 2018) in the following form

$$\begin{aligned} \phi_X &= F_1(Z) + u_1, \\ \phi_Y &= F_2(Z) + u_2, \end{aligned} \tag{7}$$

where ϕ, F_1 and F_2 are the nonlinear feature mapping in the RHKS.

Then the null hypothesis $Y \perp\!\!\!\perp X|Z$ holds true, if and only if the $\|\Sigma_u\|_{HS}^2 = 0$ where Σ_u is the variance matrix of two residues u_1 and u_2.

Given that X and Z are deterministically related, and also $X = g(Z)$. Therefore, the residue term after the kernel ridge regression will always be **0** (Perfectly representation! Note that X corresponds to ϕ_X and $g(Z)$ corresponds to $F_1(Z)$ in the Eq. 7).

In other words, $\|\Sigma_u\|_{HS}^2 = 0$ will always holds true. Then, the null hypothesis will also hold true: $Y \perp\!\!\!\perp X|Z$.

To summarize, we can conclude: $Y \perp\!\!\!\perp X|Z$ will always hold true, if X and Z deterministically related. To extend this result from one variable Y to arbitrary non-deterministic variable, and extend the result from one variable X to arbitrary deterministic variable. We can conclude that: Any deterministic variable, given the rest deterministic variables in DC, is always conditionally independent from any non-deterministic variable in NDC.

Proof ends.

A2.2 PROOF OF THEOREM 3

Proof: As suggested by the generalized score (Huang et al., 2018), with proper score functions and seach procedures, asymptotically the resulting Markov equivalence class has the same independence constraints as the data generative distribution.

(i) First of all, we would like to discuss the local consistency of generalized score.

For the regression problem one can define the effective dimension of the kernel space and the complexity of the regression function according to **?**. Then under mild conditions, the CV-likelihood score is locally consistent .

Lemma 6 *Suppose that the sample size of each test set* n_0 *satisfies*

$$n_0 \to \infty, \frac{n_0}{n} \to 0 \text{ as } n \to \infty,$$

and suppose that the regularization parameter λ *satisfies*

$$\lambda = O(n^{-\frac{b}{bc+1}}),$$

where n *is the total sample size,* b *is a parameter of the effective dimension of the kernel space with* $b > 1$*, and* c *indicates the complexity of the regression function with* $1 < c \leq 2$*.*

Lemma 7 *Assume that all conditions given in Lemma 6 hold. With the CV likelihood under the regression framework in RKHS as a score function and with the GES search procedure, it guarantees to find the Markov equivalence class which is consistent to the data generative distribution asymptotically.*

Lemma 7 ensures that, with proper score functions and seach procedures, asymptotically the resulting Markov equivalence class has the same independence constraints as the data generative distribution. For the complete proofs, please refer to the Appendix A5 of paper (Huang et al., 2018).

(ii) Then, We will provide the proof by contra-positive in both directions based on the consistency of the generalized score as shown above.

1) "If" direction:

Suppose that exact score-based search asymptotically outputs a DAG $\mathcal{H}$ (having the highest generalized score) that does not belong to the MEC of the true DAG $\mathcal{G}$. Since the generalized score is known to be consistent, $(\mathcal{H}, \mathbb{P})$ must satisfy the Markov assumption, because otherwise its generalized score is lower than that of the true DAG $\mathcal{G}$ and exact search would not have output $\mathcal{H}$. By assumption, the generalized score of $\mathcal{H}$ is higher than that of $\mathcal{G}$, which, by the consistency of generalized, implies that $|\mathcal{H}| \leq |\mathcal{G}|$, and therefore, $(\mathcal{G}, \mathbb{P})$ does not satisfy the SMR assumption.

2) "Only if" direction:
Suppose that $(\mathcal{G}, \mathbb{P})$ does not satisfy the SMR assumption. Then there exists a DAG $\mathcal{H}$ not in the MEC of $\mathcal{G}$ such that $|\mathcal{H}| \leq |\mathcal{G}|$, and $(\mathcal{H}, \mathbb{P})$ satisfies the Markov assumption. Without loss of generality, we choose $\mathcal{H}$ with the least number of edges. We first consider the case in which $|\mathcal{H}| < |\mathcal{G}|$. Since both $\mathcal{H}$ and $\mathcal{G}$ satisfy the Markov assumption, by the consistency of generalized, the generalized score of $\mathcal{H}$ is higher than that of $\mathcal{G}$, which implies that exact score-based search will not output any DAG from the MEC of $\mathcal{G}$. For the case with $|\mathcal{H}| = |\mathcal{G}|$, since they are both Markov with distribution $\mathbb{P}$, they have the same generalized score. Therefore, exact search will output a DAG that belongs to the MEC of either $\mathcal{H}$ or $\mathcal{G}$, and is not guaranteed to output a DAG from the MEC of the true DAG $\mathcal{G}$.

Proof ends.

A2.3 PROOF OF THEOREM 4

Proof: We will divide the whole proofs into two parts. For the first part, we aim to prove the "representation" theorem, and for the second part, we aim to further prove the "perfect representation" theorem.

(i) Representation:

Assume there is a MEC $\mathcal{M}$, which contains both directed edges and undirected edges. Let X be a random variable in $\mathcal{M}$ and Z be the set of all non-descendant neighbors including direct causes and undirected neighbors of X. Suppose the random variables X and Z are over measureable spaces $\mathcal{X}$ and $\mathcal{Z}$, respectively.

Without assuming a particular functional causal form, we usually exploit a regression framework in the RKHS, to encode

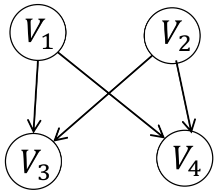


Figure A1: An example graph with deterministic relations: $V_3 = f(V_1, V_2)$, $V_4 = g(V_1, V_2)$.

general dependence relations between two random variables.
Define a RKHS $\mathcal{H}_{\mathcal{X}}$ on $\mathcal{X}$ with continuous feature mapping $\phi_{\mathcal{X}} : \mathcal{X} \rightarrow \mathcal{H}_{\mathcal{X}}$. Here, we consider

$$\phi_{\mathcal{X}}(X) = F(Z) + u, \tag{8}$$

where $F : \mathcal{Z} \rightarrow \mathcal{H}_{\mathcal{X}}$ and u represents the regression residue or noise. When applying the kernel ridge regression, we can obtain the estimated residue

$$\hat{u} = \varepsilon(\boldsymbol{K}_Z + \varepsilon I)^{-1}\phi(X), \tag{9}$$

where ε is a small positive regularization parameter for kernel ridge regression, and $\boldsymbol{K}_Z$ is the centralized kernel matrix of Z. To evaluate whether such a residue exists, one may consider Hilbert-Schmidt norm of the variance matrix

$$\Sigma_{\hat{u}} = \hat{u}^T \hat{u} = 0, \tag{10}$$

If the above equation holds true, then we may conclude that there is no noise term in the relationship between X and Z, in other words, X can be represented by Z (without extra noise term).

Vice versa.

(ii) Perfect representation:

Intuitively speaking, the perfect representation can be motivated by the example, as shown in Figure A1, where $V_3 = f(V_1, V_2)$, $V_4 = g(V_1, V_2)$.

When we are representing the variable V_3, there should be multiple ways

$$\begin{aligned} V_3 &= f(V_1, V_2) \\ &= h(V_1, V_2) + g(V_1, V_2) \\ &= h(V_1, V_2) + V_4. \end{aligned} \tag{11}$$

Accordingly, we can say that V_3 can be represented by $\{V_1, V_2\}$ or $\{V_1, V_2, V_4\}$. Clearly, the minimum deterministic cluster should be $\{V_1, V_2, V_3\}$ and $\{V_1, V_2, V_4\}$. In other words, $\{V_1, V_2, V_4\}$ can be redundant in the relationships to represent V_3. Therefore, we can conclude that: V_3 can be represented by $\{V_1, V_2, V_4\}$, but perfectly represented by $\{V_1, V_2\}$.

Here, if we cannot find a smaller subset of current set for representation, then we can say that this current set is a perfect representation of that variable. Mathematically, we have:

X can be perfectly represented by Z, if and only if

1) X can be represented by Z and,
2) X cannot be represented by any subset Z_s of Z, where $0 < |Z_s| < |Z|$.

Vice versa. With the concept of the perfect representation, we can easily detect all the MinDC in our graph.

Proof ends.

A2.4 PROOF OF THEOREM 5

Proof:

First of all, we will explain why we need the listed three assumptions. Then, we will explain why we need to have constraint on $|\,\mathrm{PA}_i\,| < |\,\mathrm{MinDC}\,| - 1$.

(i) As mentioned in our main paper, there are three phases of our proposed DGES. During the first phase, we need to run GES. To ensure the accuracy of output (particularly on the NDC part), we need the assumptions of Markov and non-deterministic faithfulness (See Assumption 1 and 3). Then

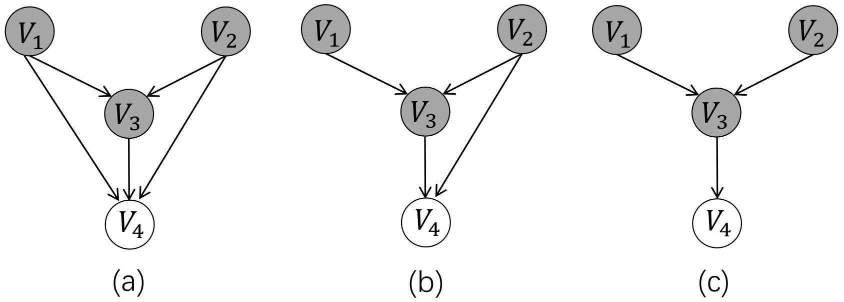


Figure A2: An example graph with determinisic relation where $V_3 = f(V_1, V_2)$. (a) A non-deterministic variable V_4 connects to $\{V_1, V_2, V_3\}$. (b) A non-deterministic variable V_4 connects to $\{V_2, V_3\}$. (c) A non-deterministic variable V_4 connects to $\{V_3\}$. Here among the three graphs, only the graph (c) can be partially identified.

in the third phase, we need to perform the exact search exclusively on the EDC, where the Sparest Markov Representation (SMR) assumption will be needed.

(ii) As for why we need to condition on $|\,\mathrm{PA}_i\,| < |\,\mathrm{MinDC}\,| - 1$, we can start with explaining why $|\,\mathrm{PA}_i\,| = |\,\mathrm{MinDC}\,|$ and $|\,\mathrm{PA}_i\,| = |\,\mathrm{MinDC}\,| - 1$ will fail the provided identifiability.

Let's take an example with four variables, where three of them are deterministically related, as shown in Figure A2. Here among the three graphs, only the graph (c) can be partially identified, and the graph (a) and (b) cannot achieve partial identifiability.

We further assume a linear functional causal model, then we can formulate the deterministic relationship as

$$aV_1 + bV_2 + cV_3 = 0, \tag{12}$$

where a, b, c are any linear coefficients. Based on the above formulation, the causal equation of variable V_4 in Figure A2(a) can be represented as

$$\begin{aligned} V_4 &= dV_1 + eV_2 + fV_3 + \epsilon \\ &= dV_1 + eV_2 + f\frac{1}{c}(aV_1 + bV_2) + \epsilon \\ &= dV_1 + e\frac{1}{b}(aV_1 + cV_3) + fV_3 + \epsilon \\ &= d\frac{1}{a}(bV_2 + cV_3) + eV_2 + fV_3 + \epsilon, \end{aligned} \tag{13}$$

where ϵ is the random noise injected into V_4. Clearly, the above four equations are all valid, in other words, V_4 can be possibly represented by different sets of variables, meaning that this case is not guaranteed to be identified.

Regarding the variable V_4 in Figure A2(b), the causal equation can be represented as

$$\begin{aligned} V_4 &= eV_2 + fV_3 + \epsilon \\ &= eV_2 + f\frac{1}{c}(aV_1 + bV_2) + \epsilon \\ &= e\frac{1}{b}(aV_1 + cV_3) + fV_3 + \epsilon. \end{aligned} \tag{14}$$

Again, the above three equations are all valid, in other words, V_4 can be possibly represented by different sets of variables, meaning that this case is also not guaranteed to be identified.

However, in Figure A2(c), things are different. The causal equation of variable V_4 can be represented as

$$\begin{aligned} V_4 &= fV_3 + \epsilon \\ &= f\frac{1}{c}(aV_1 + bV_2) + \epsilon. \end{aligned} \tag{15}$$

When the SMR assumption is satisfied, we can identify the only one case, which is $V_3 \to V_4$.

Now, we extend the three-variable case to the general linear case where there is a MinDC with the cardinality $|\,\mathrm{MinDC}\,|$. And we can easily conclude the true conditions to be: $|\,\mathrm{PA}_i\,| < |\,\mathrm{MinDC}\,| - 1$.

Furthermore, we extend the linear to nonlinear case, we can get conclude the true condition as shown above.

Proof ends.

A3 MORE DETAILS ABOUT THE EXPERIMENTS

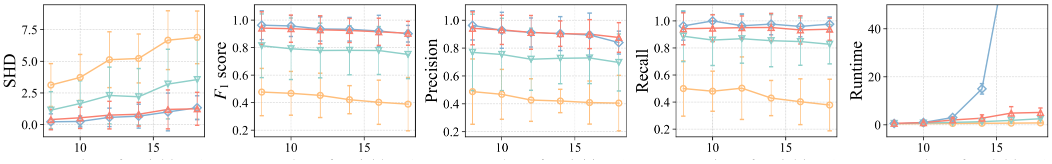
A3.1 IMPLEMENTATION DETAILS

We provide the implementation details of our method and other baseline methods for synthetic datasets.

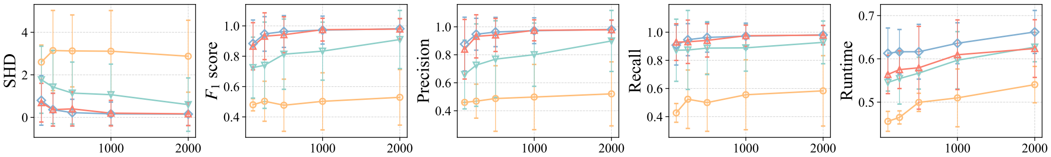
- DPC (Glymour, 2007): The method is an extension for traditional PC algorithm (Spirtes et al., 2000), the key idea is that: every time when we do the conditional independence test, we aim to remove the potential deterministic variables from the conditioning set so that the faithfulness will not be violated. Here we follow the paper, and use the covariance to measure the closeness of two variables. If the covariance between two variables are greater than 0.9, we then remove the variable from the conditioning set in conditional independence test. Meanwhile, for linear Gaussian model, we choose FisherZ test, while for nonlinear model we choose kernel-based test (Zhang et al., 2012), and the significance level is set to $\alpha = 0.05$ by default. We implement this method based on the Causal-learn package `https://github.com/py-why/causal-learn` (Zheng et al., 2023).
- GES (Chickering, 2002): This method is a classical score-based method with greedy search. Our implementation is based on the code from `https://github.com/juangamella/ges`. For linear Gaussian model, we use BIC score. And for general nonlinear model, we use generalized score with cross-validation likelihood (Huang et al., 2018). The penalty parameter for controlling the sparsity is set to 1.
- A* (Yuan & Malone, 2013): A* is one of the classical exact score-based methods. Actually, there are some heuristic algorithms proposed to accelerate the search procedure. Considering in our scenarios, we do not utilize any heuristic tricks for the experiments in order to ensure the accuracy of solutions. Our experiments are based on the implementations on the Causal-learn package `https://github.com/py-why/causal-learn` (Zheng et al., 2023).
- DGES (ours): The first phase of our method is to run GES, as introduced above. The exact search in the third phase we incorporate is the A* as mentioned above. During the second stage, when we aim to detect the deterministic clusters and checking whether a variable can be perfectly represented by some others, we set that if the term $\|\Sigma_u\|_{HS}^2 < 1e{-}3$, although theoretically the value should exactly be zero. Meanwhile, the regularization parameter for the kernel ridge regression is set to $1e{-}10$.

A3.2 EVALUATION ON TWO DCS

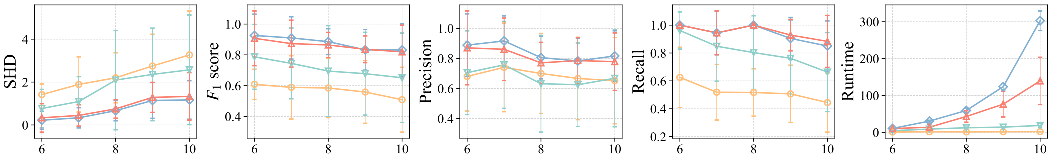
Figure 4 in the main paper presents the simulated results focused on graphs containing just a single deterministic constraint (DC). In contrast, Figure A3 in the Appendix offers insights into scenarios involving two DCs, even allowing for the possibility of overlapping variables. An evident



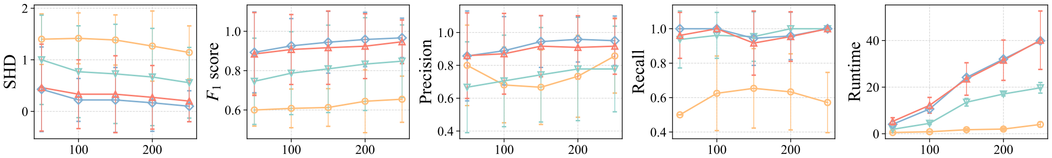
(a) Linear Gaussian model with varying number of variables d.



(b) Linear Gaussian model with varying number of samples n.



(c) Nonlinear model with varying number of variables d.



(d) Nonlinear model with varying number of samples n.

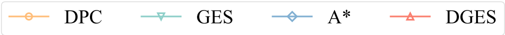


Figure A3: Simulated results on graph with two deterministic clusters. We evaluate different functional causal models on varying number of variables and samples, respectively. For each setting, we consider SHD (↓), F_1 score (↑), precision (↑), recall (↑) and runtime (↓) as evaluation criteria.

trend emerges: as the system incorporates more deterministic variables, the runtime of our Directed Graphical Exploration System (DGES) inevitably escalates. This phenomenon can be attributed to the increased number of deterministic variables demanding detection and inclusion in Phase 3, where an exact search is performed.

It is worth noting that as the number of variables in the system increases, the runtime of A* experiences a rapid surge. In stark contrast, DGES exhibits a more stable increase in runtime, demonstrating its efficiency and suitability for both linear and nonlinear models.

The outcomes gleaned from these experiments collectively indicate that DGES exhibits competitive performance compared to established baselines. Notably, the exact method A* and our proposed DGES consistently outperform other baseline methods like Greedy Equivalence Search (GES) and PC, across a spectrum of evaluation criteria and diverse settings. It is intriguing to note that in deterministic systems, the score-based method GES consistently outperforms the constraint-based method DPC. This observation suggests that score-based approaches maintain a comprehensive perspective on causal discovery, which appears to be less susceptible to the challenges posed by deterministic relationships, unlike constraint-based methods.