KCRL: A Prior Knowledge Based Causal Discovery Framework with Reinforcement Learning

Uzma Hasan UZMAHASAN@UMBC.EDU
Causal AI Lab, Department of Information Systems
University of Maryland, Baltimore County
Baltimore, Maryland, USA

Md Osman Gani MOGANI@UMBC.EDU
Causal AI Lab, Department of Information Systems
University of Maryland, Baltimore County
Baltimore, Maryland, USA

Abstract

Causal discovery is an important problem in many sciences that enables us to estimate causal relationships from observational data. Particularly, in the healthcare domain, it can guide practitioners in making informed clinical decisions. Several causal discovery approaches have been developed over the last few decades. The success of these approaches mostly rely on a large number of data samples. In practice, however, an infinite amount of data is never available. Fortunately, often we have some prior knowledge available from the problem domain. Particularly, in healthcare settings, we often have some prior knowledge such as expert opinions, prior RCTs, literature evidence, and systematic reviews about the clinical problem. This prior information can be utilized in a systematic way to address the data scarcity problem. However, most of the existing causal discovery approaches lack a systematic way to incorporate prior knowledge during the search process. Recent advances in reinforcement learning techniques can be explored to use prior knowledge as constraints by penalizing the agent for their violations. Therefore, in this work, we propose a framework KCRL [1] that utilizes the existing knowledge as a constraint to penalize the search process during causal discovery. This utilization of existing information during causal discovery reduces the graph search space and enables a faster convergence to the optimal causal mechanism. We evaluated our framework on benchmark synthetic and real datasets as well as on a real-life healthcare application. We also compared its performance with several baseline causal discovery methods. The experimental findings show that penalizing the search process for constraint violation yields better performance compared to existing approaches that do not utilize prior knowledge.

1. Introduction

Causal discovery (CD) or causal structure search is an important problem in many sciences that identifies the causal relationships between variables (Glymour et al. (2019)). The outcome of causal discovery is a causal graph, particularly a directed acyclic graph (DAG) (Williams et al. (2018)) where directed edges represent the cause and effect relationships with an arrow from the *cause* to the *effect* (Pearl (2009)). Over the last few decades,

1. Code URL: https://github.com/UzmaHasan/KCRL

CD has been studied widely to develop methods that try to capture the underlying causal story from observational data (Glymour et al. (2019), Vowels et al. (2021)). Also, with the ever-increasing application of artificial intelligence (AI) in healthcare, the need to build causality-driven AI systems has received increased attention from clinical researchers and policymakers (Yang et al. (2013), Hernán et al. (2019)). Knowing the causal structure plays a significant role in unraveling the data generating mechanism and thereby, facilitates informed decision-making. For example, to understand what factors truly influence a clinical outcome, we need to look beyond mere correlations and consider the causal relationships between them. Thus, causal discovery plays a significant role in clinical decision support systems to help researchers understand the causal structure and leverage it to investigate optimum treatment policy.

The gold standard to identify causal relationships is to perform randomized control trials (RCTs) (Hariton and Locascio (2018)). However, RCTs are increasingly considered time-consuming, costly, and often infeasible due to ethical and efficacy concerns (Cai et al. (2021)). For example, to study how the consumption of cocaine has long-term effects on brain health would require conducting an experiment forcing a group of people to consume cocaine for a while and observing the outcome at the end. This type of experimental scenario is risky, unethical as well as time-consuming. Hence, researchers are often left to experiment with non-manipulative observational data (OD) (Vowels et al. (2021)) collected using methods such as human observation, clinical observation, open-ended surveys, and various devices. Over the years, the adoption of computerized systems has increased in healthcare sector. Systems such as electronic health records (EHRs) are a great source of observational data (Pacaci et al. (2018)). Researchers have increasingly focused on developing methods to discover causal structures from purely observational data (Spirtes et al. (2000), Pearl (2009)). These methods rely on different assumptions such as causal sufficiency and causal faithfulness to infer causal relations (Spirtes et al. (2000)).

Although there exist several approaches for causal discovery, these methods suffer from the lack of a large number of data samples required for reliable and efficient structure search. Most of these methods can find true causal graphs in the presence of an infinite amount of data and appropriate model conditions (Ng et al. (2019)). However, the requirement of a vast amount of data is impractical for real-life problems. On a brighter note, often there exists some prior knowledge about the causal relationships between the observables which can be a valuable resource to address the data scarcity problem (Sinha and Ramsey (2021)). For example, in healthcare research, we often have some prior knowledge such as expert opinions, findings from prior RCTs, systematic reviews, and domain knowledge (Murad et al. (2016), Burns et al. (2011), Bhargava and Bhargava (2007)) about the clinical problem. Considering prior knowledge is important as it can be used to raise the statistical power of CD algorithms to infer causal relationships (Anjum et al. (2009)). Even the consideration of a small set of causal evidence can help to significantly increase the performance of causal discovery (Borboudakis and Tsamardinos (2012)). As per Sinha and Ramsey (2021), background knowledge, such as specifying one variable as the cause of another, can further refine the set of DAGs that enhances model efficiency (Castelo and Siebes (2000)) as well as overcome data limitation.

However, to the best of our knowledge, none of the existing CD approaches have a systematic way to incorporate the available knowledge (such as literature evidence, experts'

opinion, and prior RCTs) during the causal discovery process. Hence, this research aims to propose a framework for the systematic inclusion of existing knowledge during the CD process. Our framework utilizes the existing evidence as a constraint to reward or penalize the search process. It uses a reinforcement learning (RL) based search which is a promising and emerging technique among the various recent advances in CD approaches (Vowels et al. (2021)). To summarize, we aim to impose the prior knowledge as constraints to penalize an RL agent during the structure search process. The prior knowledge constraints are the presence or absence of causal edges. The agent will be guided for any violation of the imposed constraints. This will significantly improve performance by reducing the search space and also, enabling a faster convergence to the optimal structure. It will also facilitate the discovery of a greater number of true causal edges. We believe the proposed approach *KCRL* will add a new dimension by providing a systematic way to impose existing knowledge as constraints during the causal discovery process. We demonstrate the effectiveness of the proposed method on benchmark synthetic clinical datasets namely LUCAS (Lucas et al. (2004)) and ASIA (Lauritzen and Spiegelhalter (1988)). Both datasets are related to the diagnosis, prevention, and cure of lung cancer. We further experimented on a benchmark real dataset named the SACHS (Sachs et al. (2005)) that represents a protein signaling network. Moreover, we validated the performance of our approach on a clinical application that measures the effect of oxygen therapy intervention in intensive care unit (ICU) patients (Gani et al. (2020). Our contributions are summarized below:

- We propose a novel causal discovery framework that can be used in any domain including healthcare where some prior knowledge such as experts' opinion, domain knowledge, literature evidence, prior RCTs and experimental evidence are available. Our framework is novel as it incorporates the prior knowledge as constraints during the causal search process and also, penalizes the search process for constraint violations.

- The components of our framework such as the search process and reward mechanism are modular. It means these can be adopted from different existing implementations of these modules and can be varied based on the experimental context.

- We formulate a stopping criterion for our algorithm. It based on the experimental findings and depends on the performance metrics. When these metrics exceed a certain threshold or when their values become constant with increasing epochs, the stopping point is reached.

- We validate the effectiveness of our proposed framework by conducting experiments on benchmark synthetic and real datasets with available ground truth graphs. The experimental results demonstrate that our approach outperforms the baseline approaches over different evaluation metrics in all the experiments.

- We also demonstrate its validation on a real-life clinical problem related to oxygen therapy in ICU. The clinical application entails a timely and important healthcare problem applicable in a variety of disease conditions including severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) patients.

Generalizable Insights about Machine Learning in the Context of Healthcare

Learning causal structures from observational data is a significant problem in machine learning. Knowing the causal graph can unravel the data generating mechanism which can have an important impact on clinical decision making. It can help clinicians understand the disease mechanism better and explore their causes. However, often the existing algorithms fail to discover the true causal relations due to lack of a large amount of observational data. On the contrary, fortunately, we often have some prior knowledge or experimental evidence available. In healthcare domain, we have previous findings on the medical problems in different forms such as prior RCTs, domain knowledge, expert opinion and evidence from literature. These resources are not utilized by the existing approaches during the discovery process. The actual power of the algorithm and data can be harnessed only when some form of prior knowledge is used. The utilization of the existing evidence can improve the performance of the discovery process as well as provide a solution to the data limitation problem. Our framework provides a systematic way to utilize the prior findings during the structure search process. This will improve the discovery of causal graphs which are important to estimate treatment effects to help healthcare researchers make better and informed decisions. It will also optimize the search process to enable an early convergence to the optimal structure. Our approach can be used in a wide variety of healthcare problems whenever there exists some prior knowledge.

2. Related Work

The cost ineffectiveness or ethical concerns of randomized experiments led to the consideration of causal discovery from purely observational studies with greater importance (Cai et al. (2021)). A considerable amount of studies have been developed over the past few decades to learn causal structure from observational data. A summarized discussion of these approaches is given below.

Among the various approaches to find the underlying causal graph, constraint-based methods such as PC and FCI (Harris and Drton (2013), Spirtes et al. (2000)) perform conditional independence tests to check dependency between the variables. On the other hand, score-based methods generally optimize a score function (Yu et al. (2019)) and rely on local heuristics to perform the search. The most widely used score-based method Greedy Equivalence Search (GES) (Chickering (2002)) finds causal structures by greedily adding or deleting edges until it reaches a local maximum. Though GES is expected to work well with infinite data and appropriate model conditions, it gives much less satisfactory results with a finite amount of the data (Ng et al. (2019)). Hybrid methods such as Max-Min Hill-Climbing (MMHC) (Tsamardinos et al. (2006)) combine conditional dependency tests with a score-based approach. Another category of causal discovery approaches is based on Functional Causal Models (FCMs) (Glymour et al. (2019)). ANM (Hoyer et al. (2008)) is such a function-based approach for nonlinear causal discovery with additive noise models. Among other noteworthy function-based approaches, DirectLiNGAM (Shimizu et al. (2011)) which is a direct learning algorithm and ICALiNGAM (Shimizu et al. (2006)) which is an ICA-based learning algorithm are noteworthy both of which are linear non-Gaussian acyclic models (LiNGAM). Recently, there have been an increasing number of methods which leverage the advantages of continuous optimization to seek the structure

from data which resulted in the union of causal discovery and deep learning methods (Vowels et al. (2021)). These approaches formulate the combinatorial graph-search problem into a continuous optimization problem. DAGs with NOTEARS (Zheng et al. (2018)), first formulated the combinatorial graph search problem as a continuous optimization problem. DAG-GNN (Yu et al. (2019)) extends NOTEARS by incorporating a deep generative model known as Variational AutoEncoder (VAE) with neural network functions and providing a variant of acyclicity constraint suitable for deep learning methods. Another method GraN-DAG (Lachapelle et al. (2019)) also extends NOTEARS to identify nonlinear relationships between variables using neural networks. A more efficient version of NOTEARS is GOLEM (Ng et al. (2020)) that can reduce the number of optimization iterations. It is basically a likelihood-based structure learning method with continuous unconstrained optimization. CGNN (Goudet et al. (2018)) combines continuous optimization, neural networks and a hill-climbing search algorithm to optimize and learn the structure of a DAG. MaskedNN (Ng et al. (2019)) reformulates the Structural Equation Model (SEM) in an augmented form with a binary graph adjacency matrix. CAREFL (Khemakhem et al. (2021)) combines causal discovery with normalizing flows, a deep learning framework. DEAR (Shen et al. (2020)) combines a Variational AutoEncoder (VAE) with an adversarial loss in order to infer a latent space with causal structure. RL-BIC (Zhu et al. (2019)) uses reinforcement learning (RL) to find the causal graph with the best score. In CORL, Wang et al. (2021) formulated the ordering search problem as a multi-step Markov decision process and used RL to optimize the proposed model based on the reward mechanisms designed for each ordering.

Causal discovery with prior knowledge Over the years, studies have been conducted to include existing evidence in the causal search process and see how this inclusion can benefit the discovery process. Mahony et al. (2001) proposed an approach for encoding prior knowledge into learning algorithms by imposing a Riemannian geometry on parameter space. Borboudakis et al. (2011) presented a constraint-based approach to incorporate prior knowledge in causal models by orienting the unoriented edges of a partially oriented ancestral graph (PAG) based on the available causal edges. Flores et al. (2011) considers expert elicited pairwise relations between two variables (pairwise priors) where each edge has a frequency associated with it. The model presented by Mansinghka et al. (2012) assumes that variables come in one or more classes, and the prior probability of an edge existing between two variables is a function only of their classes. Borboudakis and Tsamardinos (2012) presents algorithms for incorporating path constraints to Partial DAGs (PDAGs) and PAGs which use chronological backtracking search, with forward checking and a pruning rule. Xu et al. (2015) introduces three types of prior knowledge given by domain experts such as existence of parent node, absence of parent node, and distribution knowledge of nodes and edges into the Markov chain Monte Carlo (MCMC) algorithm. Oyen et al. (2016) presents a novel Bayesian network (BN) discovery algorithm for learning a DAG via statistical inference of conditional dependencies from observed data with an informative prior on the partial ordering of variables. Another method PKCL (Wang et al. (2020)) leverages the Markov Blanket (MB) sets learned in the local stage to learn the global BN structure in which prior knowledge is incorporated to guide the global learning phase. Sinha and Ramsey (2021) presents an approach named Kg2Causal that uses the knowledge graph

(KG) derived edges to guide the data-driven inference of a causal Bayesian network. It generates 100 random DAGs, optimizes them using Tabu search (Gendreau (2003)), extract edges present in the KG and incorporate them as a prior.

Despite the progress in causal discovery and the incorporation of the prior knowledge, there is a lack of systematic approach to reward or penalize the search algorithm for supporting or violating existing evidence. However, such an approach can help to improve the performance of the causal discovery process and also, incorporate the evidence well.

3. Methodology

Our framework KCRL is inspired by the recent advances in continuous optimization strategies for structure search. KCRL imposes prior knowledge as constraints and uses a reward-penalty mechanism to guide the search process. In this section, at first we provide a brief introduction to the importance of prior knowledge and how prior knowledge can be used as a constraint during causal discovery in RL-based search. Then, in the following subsections, we discuss our proposed framework in details as well as its implementation and complexity analysis.

Significance of prior knowledge We often have some prior knowledge available for causal modeling in many applications. For example, in medicine, most of the time we have prior knowledge about etiology, symptoms, and treatment of underlying diseases or conditions in biomedical literature or knowledge bases (Sinha and Ramsey (2021)). This gives us an ample opportunity to leverage prior knowledge to address data scarcity. A number of works highlight the importance of inclusion of background knowledge in causal structure learning. A recent work by Andrews et al. (2020) shows how the FCI algorithm is sound and complete with the incorporation of tired background knowledge. They also mention that incorporation of the available evidence while causal discovery not only causes identification of the additional causal relationships but also helps to identify unresolved causal relationships. As per Borboudakis and Tsamardinos (2012), even a few causal constraints can orient a significant number of edges. Xu et al. (2015) mentions that an algorithm can adopt prior knowledge with even very small confidence with an inherent probability and have an influence on the search process to find a better structure. In a Bayesian network discovery approach, Oyen et al. (2016) mentioned that, without expert knowledge about the ordering of the variables, many of the edges learned are the reverse of the true edges. Thus, a hybrid approach of combining the observational data with expert-backed knowledge in the form of a constraint or prior probability reduces both the search space and biases in the search, hence improving the learning efficiency (O'Donnell et al. (2006)). From the above discussion, we can see how important it is to utilize the existing evidence during CD and even a small amount of prior knowledge can have significant effect upon the causal discovery process.

Prior knowledge in RL based search Even though the use of reinforcement learning for causal discovery has not yet been fully explored, this technique has significant opportunities in finding causal structures if utilized strategically (Vowels et al. (2021)). Recently, RL has achieved promising results in causal discovery from observational data (Wang et al. (2021)). RL solves a problem by trial and error, and employs an iterative reward feedback

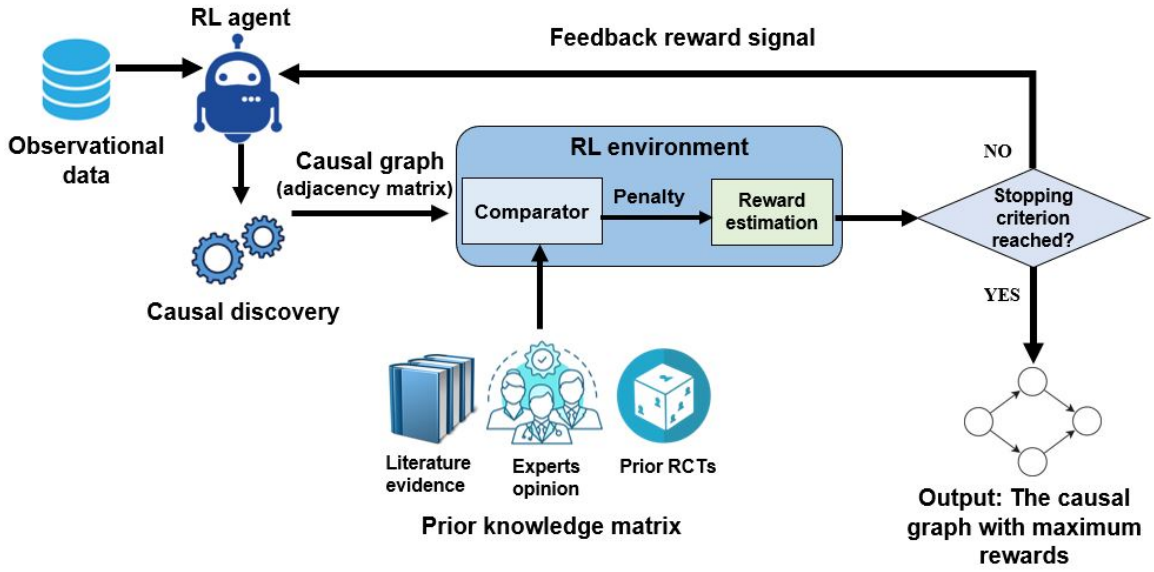


Figure 1: The proposed framework (KCRL)

(positive or negative) after taking actions (Sutton and Barto (2018)). Constraints such as existing evidence, acyclity, etc. can be imposed to guide an RL agent towards an optimized policy. A feedback can be sent by rewarding or punishing the agent for following or violating the constraints, respectively. The positive or negative rewards not only ensures the enforcement of the constraints, but also ensures formalising a better search strategy for discovering the causal graph. A better feedback or reward mechanism is formed by penalizing the search process for each incorrectly identified edge or edge orientation w.r.t. the prior information. This will alert an RL agent regarding its graph formation strategy.

3.1. Proposed Framework

Our proposed framework (Figure 1) for causal discovery shows how the existing evidence (known causal edges) is used to guide the search process. Here, prior information of two types can be used. If there exists a directed edge from one node to another node, then it is denoted by 1 and if there exists no causal edge between two nodes, then it is represented with 0. At the beginning, the observational data is provided as input to the RL agent. Its task is to convert data to a causal graph (precisely an adjacency matrix of the graph). Data to adjacency matrix conversion can be done via any existing encoder-decoder architecture such as the Transformer model (Vaswani et al. (2017)), Neural Tensor Network (NTN) model (Socher et al. (2013)), etc. Next, the generated adjacency matrix is passed through a comparator. The comparator consists of information regarding the existing edges and their orientation. Task of the comparator is to compare the edges in the produced graph by the agent with the subset of the true edges and determine the total number of mismatched edges p. For each mismatched edge, it increases the penalty by one. After the comparison, it multiplies the total number of mismatched edges p with a penalty weight β and this

product is the final penalty P_k for the generated graph.

$$S_{BIC} = nd\, log((\sum_{i=1}^{d} RSS_i)/(nd)) + \#(edges)logn \tag{1}$$

$$S_M(\mathcal{G}_h; D) = \sum_{i=1}^{m} S_M(X_i, PA_i^{\mathcal{G}_h}) \tag{2}$$

The graph is then scored using a score function such as the Bayesian Information Criterion (BIC) (Equation 1 by Zhu et al. (2019)) where, n denotes the total number of samples in the dataset and d denotes the number of nodes in the graph; the Marginal Likelihood (ML) (Equation 2 by Huang et al. (2018)), etc. The score function scores a model by determining how well the model fits the dataset and also, how complex it is (Bishop (2012)). After that, an acyclicity enforcing function such as Equation 3 by Zheng et al. (2018), Equation 4 by Yu et al. (2019), etc. is used to determine if there is any cycle in the graph or not. The goal is to ensure the acyclicity property of a DAG. Here, $\circ$ is the Hadamard product and e^A is the matrix exponential of A.

$$h(W) = tr(e^{W \circ W}) - d = 0 \tag{3}$$

$$tr[(I + \alpha A \circ A)^m] - m = 0 \tag{4}$$

The penalty P_k w.r.t evidence is then summed up with the score S and the acyclicity penalty P_a to compute the total reward R (Equation 5). Here, both the penalties for violation of acyclicity and prior knowledge constraints are negative in the reward function (Equation 5). Finally, the reward R (Equation 5) is fed back to the RL agent as a summary of its performance in the present state. This feedback assists an RL agent to determine whether its search strategy is correct or needs to be updated. The entire process ends when the stopping criterion S_c is reached (details in subsection 3.2). S_c depends on the performance metrics P_m which is also discussed in subsection 3.2. The output is the best rewarded causal graph among all the generated graphs. The step by step process of our framework is in Algorithm 1.

$$R(\mathcal{G}) = S(\mathcal{G}) + P_a + P_k \tag{5}$$

3.2. Stopping Criterion

Based on the experimental observations, we formulated some criteria which if satisfied, will be sufficient to stop the discovery process. There can be a number of conditions on which the stopping criterion S_C can rely on. During the structure discovery process, if over a period of time there is no change in the value of the performance metrics P_m (False Discovery Rate (FDR), True Positive Rate (TPR) and Structural Hamming Distance (SHD)), then we can consider to stop the discovery process (details of the metrics are in Section 5). As well as, if the majority of the metrics reach a certain threshold T and the minimum number of iterations I_{min} required for the graph discovery is reached, then the process can be stopped. The threshold T can be domain specific. For example, in a clinical scenario, a metric P_m

Algorithm 1 Causal discovery with prior knowledge constraints

Input: An observational dataset D; prior knowledge/evidence set E
Output: A causal graph with the maximum reward R
for $i = 0, 1, 2, ...$; **do**
  Generate graph adjacency matrix W from D
  for *each evidence* e *in* E **do**
    if $e \notin W$ **then**
      Increment p:= p+1 {where p represents the mismatched edges}
    end
  end
  Compute Penalty P_k:= β * p {where β = penalty weight and p = the mismatched edges}
  Compute Score S:= S_{BIC} and penalty for acyclicity P_a: = $h(W)$
  Compute Reward R:= $S + P_a + P_k$
  Feedback the Reward R
  Compute performance metrics P_m
  if *Stopping criterion* S_c *reached* **then**
    break
  end
end

may be required to have a threshold value greater than or equal to 98% whereas, in another domain the requirement can be much lower. Again it might be the case that the required threshold for a metric is reached within 1000 iterations. But, 1000 iterations might not be sufficient enough for discovering the optimal casual structure. So, there exists a trade off between the number of iterations and the threshold. It is therefore necessary to have a minimum number of iterations before deciding to stop the search process. This number can be decided by monitoring the model's performance. From the extensive experiments that we have conducted, it is found that after reaching 10K iterations, there's no more change in the performance metrics with the increment in the iterations. Hence, an iteration value of 10K could be sufficient to find the causal graphs for the types of networks used in our experiments (see Section 5). While for larger networks, a slightly higher number of iterations could be sufficient. Equation 6 represents the stopping criteria S_c where i is the iteration number and n = the total number of iterations the model runs.

$$S_c = \lim_{i \to n} \Delta P_m = 0 \parallel P_m = T \wedge I_{min} \tag{6}$$

3.3. Framework Implementation

To implement our framework, we choose the different components in our model from the following existing works. A Self-Attention based encoder-decoder architecture (Vaswani et al. (2017)) have been used to convert data to an adjacency matrix. The encoder is fed with input data distributions and the decoder takes the encoded information from the encoder to output an adjacency matrix of the graph. For the search strategy, the RL approach in Zhu et al. (2019) has been adopted in which they have used an Actor-Critic RL algorithm for graph search and used a BIC (Neath and Cavanaugh (2012)) function

(Equation 1) to score the generated graphs. To check for cycles in the graphs, the acyclity enforcing function by Zheng et al. (2018) has been used. To use prior information during the causal discovery, we implemented a comparator that compares the adjacency matrix produced by the decoder with the existing edge information stored in a prior knowledge matrix. Our framework uses the Equation 7 to compute reward where its goal is to discover the causal graph with the maximum reward. In this equation, the first term calculates the score of the graph, the second term computes the penalty for acyclicity violation and the last term represents the penalty for violating the prior knowledge where, β represents the penalty weight (details in Section 5 and Appendix D).

$$R = -[S_{BIC} + h(W) + \beta p] \tag{7}$$

3.4. Computational Complexity

Our approach is summarized in Algorithm 1. For the generation of adjacency matrix, it uses a Self-Attention based encoder-decoder (Vaswani et al. (2017)) which has a $\mathcal{O}(n^2 d)$ complexity per layer. Computational cost for the score (Equation 1) is $\mathcal{O}(dn^2)$ which is similar to RL-BIC2 (Zhu et al. (2019)). To check for acyclicity, it requires the computation of matrix exponential with $\mathcal{O}(d^3)$ cost per iteration. It is same as NOTEARS (Zheng et al. (2018)) which adopts the proximal quasi-Newton algorithm (Zhong et al. (2014)) to reduce the number of iterations ($\mathcal{O}(d^3)$) needed to converge. While the cost for P_k computation is constant. However, it can be seen that computation of the rewards is most expensive due to a large cost of the acyclicity function. This can be minimized by using a different technique to compute acyclicity such as the one in Equation 4.

4. Cohort

In this section, we describe the cohort for the Oxygen-therapy (OT) dataset (Gani et al. (2020)) that we used for experimentation. It is a real-life clinical application related to the causal effect estimation of oxygen therapy in the ICU patients. It closely followed the study guideline and selection criteria used in a pilot RCT (Panwar et al. (2016)). For cohort selection, it considers a large ICU database called MIMIC-III (Johnson et al. (2016)) that contains data routinely collected from patients in the US. The database comprises a total of 53,432 ICU stays for adult patients among which 26,026 patients (between the age group 18 to 100 years old) who received invasive mechanical ventilation (IMV) were considered. Among these patients, 4,062 patients received IMV for at least 168 hours or above. Out of these 4,062 patients, 3,812 patients received liberal oxygenation ($SPO_2 \geq 96\%$) and 250 patients received conservative oxygenation ($88 \leq SPO_2 \leq 95$ %). In total, the OT dataset contains 26 variables that involves patient demographics, ventilator settings, and oxygenation parameters. Oxygenation parameters and ventilator settings were collected every 4 hours for at least 7 days. The ground truth graph (see Appendix A) has 62 causal edges. For experimentation with our framework, we used 16 edges as prior knowledge (details in Section 5). These edges were selected based on literature evidence described in Gani et al. (2020).

5. Experiments and Results

In this section, we present a comprehensive set of experiments to demonstrate the effectiveness of our proposed approach KCRL. We conduct experiments on benchmark synthetic and real datasets as well as on an important real-world clinical application. We compared the performance of KCRL against several baselines including recent gradient based CD methods such as NOTEARS (Zheng et al. (2018)), GOLEM (Ng et al. (2020)), GraN-DAG (Lachapelle et al. (2019)), RL-BIC2 (Zhu et al. (2019)), and also a couple of function based CD approaches namely ANM (Hoyer et al. (2008)), ICA-LiNGAM (Shimizu et al. (2006)) and DirectLiNGAM (Shimizu et al. (2011)). We discuss these methods and their implementation details in Appendix B.

Study design We briefly discuss our experimental settings here. To use existing evidence during the causal discovery process, we consider a subset of the true causal edges (precisely 25% of the total edges) as prior knowledge. We considered multiple datasets for experiments which differ in terms of network size and edge densities. To *preserve uniformity* in the amount of evidence considered, we used 25% of the true causal edges for all datasets. For the penalty weight β, we select the initial value as $\beta = 0.1$ with an upper limit 1. After every 1000 epochs, β is increased by 0.1. We select the initial and the incremental values for β based on a grid search. For the ease of clarity, we want to specify how prior knowledge works for our approach. KCRL *does not include the prior edges directly into the causal graph*. Rather it compares the estimated causal edges with the true causal edges (in evidence set) and penalizes the agent for any mismatch in the edges (see subsection 3.1). The agent is never informed directly which are the edges that are mismatched. Rather, it only receives a cumulative penalty score for all mismatched edges. Furthermore, we want to clarify that none of the baseline methods except DirectLiNGAM utilize prior knowledge at all for the causal discovery. Although DirectLiNGAM mentions that if some prior knowledge on a part of the structure is available, then the number of causal orders and connection strengths to be estimated gets smaller. However, it explicitly does not use prior knowledge for the structure search. Also, to the best of our knowledge, the other prior knowledge based methods discussed in Section 2 do not have any reproducible code implementations available to use in the comparative analysis. Nevertheless, we believe a comparison with the baseline approaches help us to understand how the absence or presence of prior knowledge during the search impacts the performance of causal discovery.

Evaluation criteria The estimated graphs were evaluated using three metrics commonly used to evaluate causal graphs: the false discovery rate (FDR), the true positive rate (TPR), and the structural hamming distance (SHD). All these metrics signify a measure of accuracy or the performance of the discovery process. FDR computes the fraction of the estimated edges that are false, which gives a numerical estimate of how enriched the discoveries are compared to the ground-truth (Benjamini and Hochberg (1995)). SHD computes the number of edge insertions, deletions or flips required to transform the estimated graph to the true causal graph (Norouzi et al. (2012)). Hence, lower the SHD and FDR, closer is the estimated graph to the true causal graph. TPR measures the model accuracy (van Ravenzwaaij and Ioannidis (2019)) by computing the probability of the estimated true edges (Wang et al. (2021)). Thus, higher the TPR, better the performance of causal discovery methods.

5.1. Results on Synthetic Data

The performance of KCRL and the baseline methods on the benchmark synthetic datasets (LUCAS and ASIA) are shown in Table 1. Here, we present the performance metrics to compare our method against the baselines. In the following paragraphs, we discuss the experimental results in details.

LUCAS dataset The Lung Cancer Simple (LUCAS) (Lucas et al. (2004)) dataset contains toy data artificially generated by Causal Bayesian networks with binary variables. It has a 12-node causal graph. The data generative model of the LUCAS dataset is a Markov process. The ground-truth graph has 12 causal edges (see Appendix A). With sample size $n = 2000$, we conduct experiment where we considered 3 true causal edges (25% of the total edges) as prior knowledge. We can see in Table 1 for LUCAS dataset that KCRL achieves the best results in terms of FDR (0.43) and SHD (8). In case of true positives, GOLEM performs slightly better with a TPR 0.45 compared to KCRL (0.36). NOTEARS, DirectLiNGAM and GOLEM are on par with KCRL and have close results w.r.t all the metrics. ANM performs very poorly on this dataset. It has 0 findings of the true positive edges and the worst FDR. This could be due to the reason that ANM is specialized to model nonlinear causal relationships only. Overall, KCRL seems to perform better than others with the best results in two out of the three metrics.

Table 1: Comparison of evaluation metrics.

	LUCAS			ASIA		
Method	**FDR**	**TPR**	**SHD**	**FDR**	**TPR**	**SHD**
KCRL	**0.43**	0.36	**8**	0.25	**0.75**	**3**
NOTEARS	**0.43**	0.33	11	0.83	0.13	12
GOLEM	0.5	**0.45**	9	0.75	0.25	11
Gran-DAG	0.5	0.09	10	**0**	0.13	7
RL-BIC2	0.67	0.36	11	0.55	0.63	7
ANM	1	0	18	0.75	0.25	12
ICA-LiNGAM	0.67	0.18	10	0.6	0.25	7
DirectLiNGAM	0.5	0.36	**8**	**0**	0.5	4

ASIA dataset We further experimented on another synthetic dataset named ASIA (Lauritzen and Spiegelhalter (1988)), also known as the Lung Cancer dataset. This is a small toy network that models lung cancer in patients from Asia. Precisely, it describes different lung diseases (tuberculosis, lung cancer or bronchitis) and their relations to smoking and patients visits to Asia. This is commonly used as a benchmark dataset for graphical models. It is a small network with 8 nodes and 8 edges (see Appendix A). For experimentation, with sample size $n = 1000$, we used 2 ground-truth edges (25% evidence). From the empirical results reported in Table 1, it can be observed that KCRL achieved the best TPR (0.75) and SHD (3) compared to the baselines. It has the second best FDR (0.25) which is lower than the other methods except for Gran-DAG and DirectLiNGAM that have zero false discoveries. ANM and GOLEM perform quite poorly with similar FDRs and TPRs and a slightly varied SHD. NOTEARS is the worst performing method w.r.t all the metrics. RL-

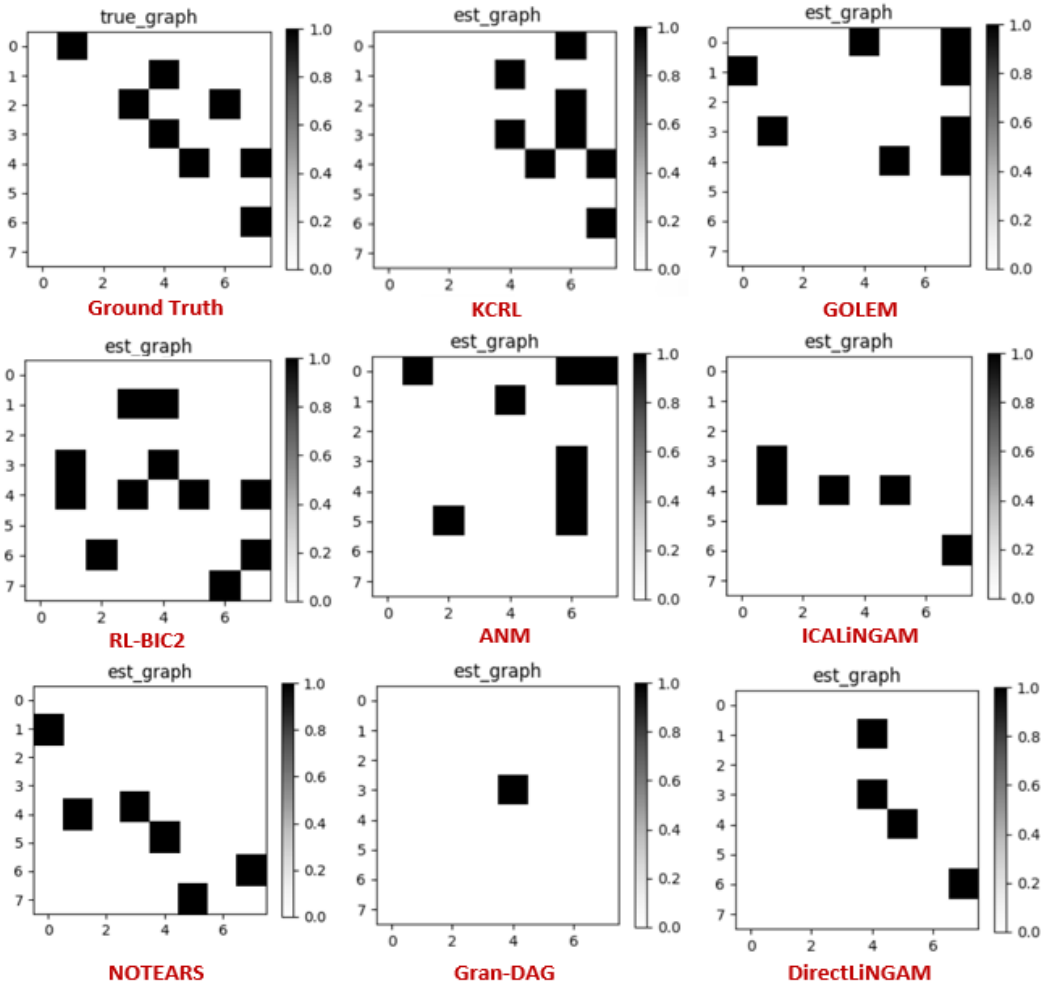


Figure 2: Estimated graphs of all the approaches for the ASIA dataset.

BIC2 seems to have an average performance compared to all the approaches. The estimated graphs compared against the ground-truth are shown in Figure 2.

5.2. Results on Real Data

SACHS dataset We consider a well known real dataset, SACHS (Sachs et al. (2005)). It is a widely used dataset for benchmarking causal discovery methods. It contains the expression levels of protein and phospholipid which are used to discover the implicit protein signal network. It has 11 nodes and 17 edges as represented in the ground-truth graph (see Appendix A). Although it has both observational and interventional data, we consider the observational data with $n = 853$ samples as our approach only relies on observed data. KCRL achieves the best results in terms of the TPR (0.35) and SHD (11). NOTEARS has almost similar results as KCRL with slightly higher SHD and an identical TPR. However, it has a higher FDR. GraN-DAG has the best FDR 0.25 but lacks significantly in TPR.

RL-BIC2 and ICA-LiNGAM both perform quite similarly with equal values of TPR and SHD. GOLEM has the second lowest TPR and the highest FDR resulting in SHD 24. Although DirectLiNGAM has a moderate FDR and SHD, however, it does not perform well in terms of true positives and has the lowest TPR (0.12). ANM failed to discover any causal relationships and produced an empty graph for this dataset. Hence we did not report the metrics for ANM. We think that ANM focuses more on nonlinear causal discovery with additive noise models and thus, fails to identify the causal relationships in SACHS.

Table 2: Comparison of evaluation metrics.

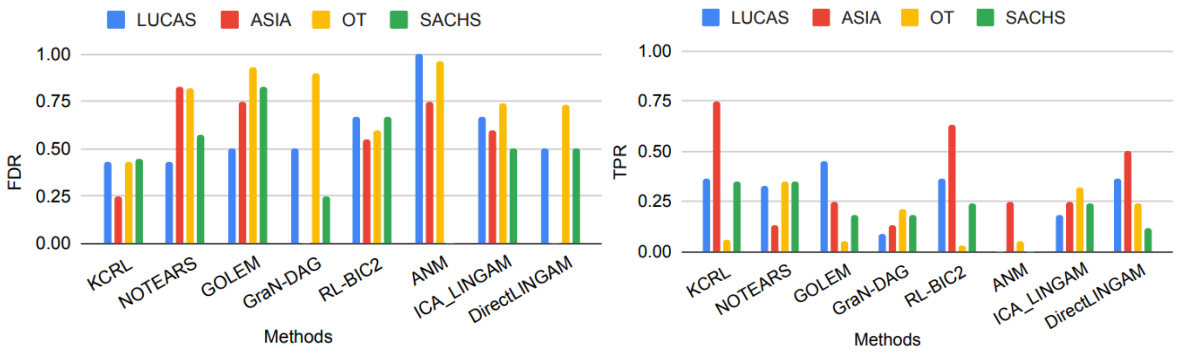
	SACHS			Oxygen-therapy		
Method	**FDR**	**TPR**	**SHD**	**FDR**	**TPR**	**SHD**
KCRL	0.45	**0.35**	**11**	**0.43**	0.06	**59**
NOTEARS	0.57	**0.35**	12	0.82	**0.35**	124
GOLEM	0.83	0.18	24	0.93	0.05	93
Gran-DAG	**0.25**	0.18	14	0.9	0.21	149
RL-BIC2	0.67	0.24	14	0.6	0.03	63
ANM	N/A	N/A	N/A	0.96	0.05	125
ICALiNGAM	0.5	0.24	14	0.74	0.32	84
DirectLiNGAM	0.5	0.12	15	0.73	0.24	76

5.3. Results on Real-world Clinical Application

Oxygen-therapy dataset We further evaluated our framework on a real-world clinical application, the Oxygen-therapy (OT) dataset (Gani et al. (2020)), that involves a timely and important clinical research problem - the effect of oxygen therapy intervention in ICU. It has observational EHR data collected during routine health care from hospitals. It contains 26 continuous and discrete variables. The study investigates the effect of liberal versus conservative oxygen therapy on mortality during the mechanical ventilation in the ICU (Panwar et al. (2016), Girardis et al. (2016)). The outcome of this study is useful in a variety of disease conditions, including severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) patients in the ICU. The dataset has a total of 26 nodes and 62 edges. For experimentation, we used $n = 1000$ samples and 16 edges (25% true edges) which were selected based on literature evidence. The performance metrics for this dataset are presented in Table 2. KCRL outperformed others in two out of the three evaluation metrics (FDR and SHD). Although, NOTEARS has the best TPR, however, its performance lacks heavily in the other two metrics. Overall, GOLEM and ANM perform the least. This could be due to the dense nature of the graph. GraN-DAG does not perform well either, as it requires the computation of a large number of optimization parameters for this dataset. ICALiNGAM and DirectLiNGAM perform moderately compared to others since they are not good at modelling nonlinear relationships.

5.4. Result Analysis

KCRL outperformed the baseline approaches in SHD (Figure 4) for *all four datasets*. This signifies its reliability in estimating the *minimal-distant* true causal graph. Also, w.r.t.



(a) (b)

Figure 3: (a) FDRs and (b) TPRs of all the approaches for the four datasets. KCRL has the *best FDR (lowest) in case of two datasets (LUCAS and OT)* and the second best FDR in case of the others. It has the *best TPR (highest) for the ASIA and SACHS dataset* and the second best TPR for LUCAS dataset.

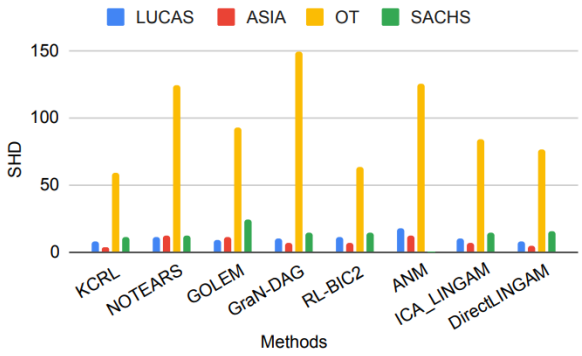


Figure 4: Estimated SHDs of all the approaches for the four datasets. KCRL *outperformed all the other methods* in terms of this metric as it has the *best SHDs (lowest) for all the datasets* compared to others.

FDR (Figure 3 (a)), it has the best results for *LUCAS and OT* datasets and, the best TPR (Figure 3 (b)), for *ASIA* and *SACHS* datasets. Relatively, KCRL has lower false discoveries in all experiments. This is important since false discoveries can have significant impact in healthcare domain. Overall, KCRL is ahead in performance w.r.t *two out of the three metrics* in all of the experiments that validate the effectiveness of our proposed framework. To summarize, our experimental findings show that existing causal discovery approaches that do not consider any prior knowledge lacks behind in performance compared to KCRL. The main reason behind the performance difference of KCRL and other methods is the influence of prior knowledge during the causal discovery process. We believe prior knowledge guides the RL agent towards the structure search by penalizing it for the mis-

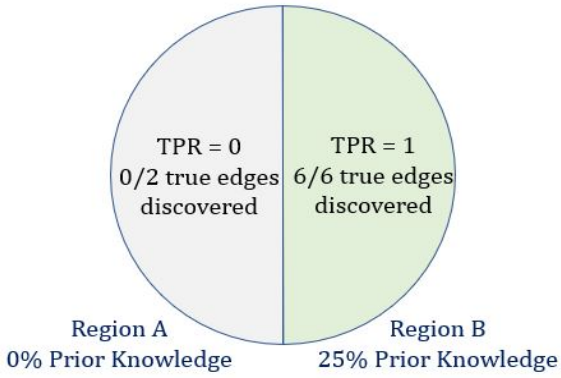


Figure 5: Discovery between regions of state-space (ASIA dataset)

matched edges. This guidance is helpful for the agent to accelerate the search process, re-think its search strategy and also, reduce the search space to some extent.

Sufficiency of prior knowledge To maintain uniformity in experimental settings, we used 25% prior knowledge throughout all the experiments. For the ASIA dataset, there are 8 ground-truth edges in the causal graph from which 2 edges (25% prior knowledge) were used as constraints to the search process. KCRL could discover 75% of the total number of true edges in case of this dataset. While, for the LUCAS (12 ground-truth edges) and SACHS (11 ground-truth edges) datasets, 36% and 35% of the true edges were discovered, respectively. Thus, we observe that the amount of prior knowledge sufficient to recover the true underlying causal graph can vary due the network sizes and the available dataset. Different datasets with networks of variable sizes may require different amounts of prior knowledge to discover the entire true graph. However, we find that leveraging any amount of available prior knowledge during causal discovery accelerates the discovery process and also, improves the accuracy.

Discovery between regions of state-space We performed an analysis to see how the discovery of the learned structure varies between regions of the state-space based on the availability of prior knowledge. For the ASIA dataset which has 8 true edges, we randomly divided the state-space into two regions (see Figure 5). One of the regions has no prior knowledge available (Region A) while the other has 25% prior knowledge (Region B). We found that in the region that has prior knowledge, our method achieved a TPR of 1 i.e. it could discover 6 out of the 6 true edges. While in the region without prior knowledge, it could not discover any true edges. This suggests that the discovery is usually better in regions with more prior knowledge or where some prior knowledge is available.

6. Discussion

We propose a novel approach to causal discovery that utilizes prior knowledge as constraints during structure search and penalizes the search process for violation of these constraints. This technique of penalizing for constraint violation guides and restricts the search process and improves the computational efficiency of the causal discovery. Essentially, it provides a

novel systematic way to utilize prior knowledge during CD. Our approach can also address the data shortage problem suffered by most of the existing causal discovery approaches. It can be used in any domain that has some form of prior information available. Particularly, for healthcare domain where often some prior information such as clinical domain knowledge, findings from prior RCTs, evidence from medical literature and expert opinion are available, this approach will help in efficient causal discovery through a systematic inclusion of evidence. Also, whenever RCTs are infeasible, this approach can be used to develop a causal graph and consequently, perform virtual RCTs using observational data. The experimental results on multiple benchmark datasets and a clinical application show that penalizing the search process for constraint (prior knowledge) violation can improve the performance of the CD as well as enables faster convergence to the optimal structure. We evaluated our approach on datasets with both small (8 or 12 or 17 edges graphs) and large (62 edges graph) edge densities. We also compared our approach with multiple baselines and observed that it outperforms the baselines in different metrics across all four datasets. Moreover, our framework is modular and its major components such as the score function, acyclicity function and search process can be adopted from other existing approaches based on the underlying assumptions for different application domains.

Limitations The limitations of this work are as follows. We were unable to compare our approach with some of the existing prior knowledge based methods due to the unavailability of reproducible codes online. Additionally, a more robust approach is needed for the selection of an optimal value for the penalty weight β. This can be a future work to find an efficient way to compute the penalty weight. Another limitation is that when we increase the edge densities in experiments, KCRL seems to suffer in performance w.r.t. TPR compared to its performance for sparse graphs. We believe this can be addressed using other robust score function for dense graphs. However, many real-world applications usually have smaller networks with sparse graphs such as the SACHS (Sachs et al. (2005)), and CANCER (Korb and Nicholson (2010)) datasets. Moreover, in this study we considered prior knowledge that is 100% accurate and trustworthy. Further investigation on how errors or biased prior knowledge can affect the search process and as a result, induce changes in the causal graph can be done. As a future work, we plan to test this framework for different amount of prior information and see how the performance varies. Also, this framework can be further evaluated by adopting the modular components from other existing implementations as a future work. Furthermore, how to incorporate prior knowledge from studies that may have different endpoints and inclusion/exclusion criteria can be investigated.

Acknowledgments

We would like to express our sincere gratitude to anonymous reviewers and the area chair for their insightful reviews that helped to improve this study. This study was supported in parts under grants from the National Science Foundation (NSF Award # 2118285), and UMBC Strategic Awards for Research Transitions (START). The content of this work does not necessarily represent the policy of NSF and you should not assume endorsement by the Federal Government.

References

Bryan Andrews, Peter Spirtes, and Gregory F Cooper. On the completeness of causal discovery in the presence of latent confounding with tiered background knowledge. In *International Conference on Artificial Intelligence and Statistics*, pages 4002–4011. PMLR, 2020.

Shahzia Anjum, Arnaud Doucet, and Chris C Holmes. A boosting approach to structure learning of graphs with and without prior knowledge. *Bioinformatics*, 25(22):2929–2936, 2009.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Kamlesh Bhargava and Deepa Bhargava. Evidence based health care a scientific approach to health care. *Sultan Qaboos University Medical Journal [SQUMJ]*, 7(2):105–107, 2007.

Christopher M Bishop. Pattern recognition and machine learning, 2006. , 60(1):78–78, 2012.

Giorgos Borboudakis and Ioannis Tsamardinos. Incorporating causal prior knowledge as path-constraints in bayesian networks and maximal ancestral graphs. *arXiv preprint arXiv:1206.6390*, 2012.

Giorgos Borboudakis, Sofia Triantafilou, Vincenzo Lagani, and Ioannis Tsamardinos. A constraint-based approach to incorporate prior knowledge in causal models. In *ESANN*. Citeseer, 2011.

Patricia B Burns, Rod J Rohrich, and Kevin C Chung. The levels of evidence and their role in evidence-based medicine. *Plastic and reconstructive surgery*, 128(1):305, 2011.

Ruichu Cai, Weilin Chen, Jie Qiao, and Zhifeng Hao. On the role of entropy-based loss for learning causal structures with continuous optimization. *arXiv preprint arXiv:2106.02835*, 2021.

Robert Castelo and Arno Siebes. Priors on network structures. biasing the search for bayesian networks. *International Journal of Approximate Reasoning*, 24(1):39–57, 2000.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

M Julia Flores, Ann E Nicholson, Andrew Brunskill, Kevin B Korb, and Steven Mascaro. Incorporating expert knowledge when learning bayesian network structure: a medical case study. *Artificial intelligence in medicine*, 53(3):181–204, 2011.

Md Osman Gani, Shravan Kethireddy, Marvi Bikak, Paul Griffin, and Mohammad Adibuzzaman. Structural causal model with expert augmented knowledge to estimate the effect of oxygen therapy on mortality in the icu. *arXiv preprint arXiv:2010.14774*, 2020.

Michel Gendreau. An introduction to tabu search. In *Handbook of metaheuristics*, pages 37–54. Springer, 2003.

Massimo Girardis, Stefano Busani, Elisa Damiani, Abele Donati, Laura Rinaldi, Andrea Marudi, Andrea Morelli, Massimo Antonelli, and Mervyn Singer. Effect of Conservative vs Conventional Oxygen Therapy on Mortality Among Patients in an Intensive Care Unit: The Oxygen-ICU Randomized Clinical Trial. *JAMA*, 316(15):1583–1589, 10 2016. ISSN 0098-7484. doi: 10.1001/jama.2016.11993. URL https://doi.org/10.1001/jama.2016.11993.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michele Sebag. Learning functional causal models with generative neural networks. In *Explainable and interpretable models in computer vision and machine learning*, pages 39–80. Springer, 2018.

Eduardo Hariton and Joseph J Locascio. Randomised controlled trials—the gold standard for effectiveness research. *BJOG: an international journal of obstetrics and gynaecology*, 125(13):1716, 2018.

Naftali Harris and Mathias Drton. Pc algorithm for nonparanormal graphical models. *Journal of Machine Learning Research*, 14(11), 2013.

Miguel A Hernán, John Hsu, and Brian Healy. A second chance to get causal inference right: a classification of data science tasks. *Chance*, 32(1):42–49, 2019.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1551–1560, 2018.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Ilyes Khemakhem, Ricardo Monti, Robert Leech, and Aapo Hyvarinen. Causal autoregressive flows. In *International Conference on Artificial Intelligence and Statistics*, pages 3520–3528. PMLR, 2021.

Kevin B Korb and Ann E Nicholson. *Bayesian artificial intelligence*. CRC press, 2010.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. *arXiv preprint arXiv:1906.02226*, 2019.

Steffen L Lauritzen and David J Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194, 1988.

Peter JF Lucas, Linda C Van der Gaag, and Ameen Abu-Hanna. Bayesian networks in biomedicine and health-care. *Artificial intelligence in medicine*, 30(3):201–214, 2004.

David G Luenberger and Yinyu Ye. Penalty and barrier methods. In *Linear and Nonlinear Programming*, pages 397–428. Springer, 2016.

Robert Mahony, Robert Williamson, et al. Prior knowledge and preferential structures in gradient descent learning algorithms. 2001.

Vikash Mansinghka, Charles Kemp, Thomas Griffiths, and Joshua Tenenbaum. Structured priors for structure learning. *arXiv preprint arXiv:1206.6852*, 2012.

M Hassan Murad, Noor Asi, Mouaz Alsawas, and Fares Alahdab. New evidence pyramid. *BMJ Evidence-Based Medicine*, 21(4):125–127, 2016.

Andrew A Neath and Joseph E Cavanaugh. The bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):199–203, 2012.

Ignavier Ng, Zhuangyan Fang, Shengyu Zhu, Zhitang Chen, and Jun Wang. Masked gradient-based causal structure learning. *arXiv preprint arXiv:1910.08527*, 2019.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. *Advances in Neural Information Processing Systems*, 33:17943–17954, 2020.

Mohammad Norouzi, David J Fleet, and Russ R Salakhutdinov. Hamming distance metric learning. In *Advances in neural information processing systems*, pages 1061–1069, 2012.

Diane Oyen, Blake Anderson, and Christine Anderson-Cook. Bayesian networks with prior knowledge for malware phylogenetics. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Rodney T O'Donnell, Ann E Nicholson, Bin Han, Kevin B Korb, Md Jahangir Alam, and Lucas R Hope. Causal discovery with prior information. In *Australasian Joint Conference on Artificial Intelligence*, pages 1162–1167. Springer, 2006.

Anil Pacaci, Suat Gonul, A Anil Sinaci, Mustafa Yuksel, and Gokce B Laleci Erturkmen. A semantic transformation methodology for the secondary use of observational healthcare data in postmarketing safety studies. *Frontiers in pharmacology*, 9:435, 2018.

Rakshit Panwar, Miranda Hardie, Rinaldo Bellomo, Loïc Barrot, Glenn M Eastwood, Paul J Young, Gilles Capellier, Peter WJ Harrigan, and Michael Bailey. Conservative versus liberal oxygenation targets for mechanically ventilated patients. a pilot multicenter randomized controlled trial. *American journal of respiratory and critical care medicine*, 193 (1):43–51, 2016.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Disentangled generative causal representation learning. *arXiv preprint arXiv:2010.02637*, 2020.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *The Journal of Machine Learning Research*, 12:1225–1248, 2011.

Meghamala Sinha and Stephen A Ramsey. Using a general prior knowledge graph to improve data-driven causal network learning. In *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*, 2021.

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. In *Advances in neural information processing systems*, pages 926–934, 2013.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

Don van Ravenzwaaij and John PA Ioannidis. True and false positive rates for different criteria of evaluating statistical evidence from clinical trials. *BMC medical research methodology*, 19(1):1–10, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like dags? a survey on structure learning and causal discovery. *arXiv preprint arXiv:2103.02582*, 2021.

Wei Wang, Gangqiang Hu, Bo Yuan, Shandong Ye, Chao Chen, YaYun Cui, Xi Zhang, and Liting Qian. Prior-knowledge-driven local causal structure learning and its application on causal discovery between type 2 diabetes and bone mineral density. *IEEE Access*, 8: 108798–108810, 2020.

Xiaoqiang Wang, Yali Du, Shengyu Zhu, Liangjun Ke, Zhitang Chen, Jianye Hao, and Jun Wang. Ordering-based causal discovery with reinforcement learning. *arXiv preprint arXiv:2105.06631*, 2021.

Thomas C Williams, Cathrine C Bach, Niels B Matthiesen, Tine B Henriksen, and Luigi Gagliardi. Directed acyclic graphs: a tool for causal studies in paediatrics. *Pediatric research*, 84(4):487–493, 2018.

Jun-Gang Xu, Yue Zhao, Jian Chen, and Chao Han. A structure learning algorithm for bayesian network using prior knowledge. *Journal of Computer Science and Technology*, 30(4):713–724, 2015.

Jing Yang, Ning An, Gil Alterovitz, Lian Li, and Aiguo Wang. Causal discovery based on healthcare information. In *2013 IEEE International Conference on Bioinformatics and Biomedicine*, pages 71–73. IEEE, 2013.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pages 7154–7163. PMLR, 2019.

Keli Zhang, Shengyu Zhu, Marcus Kalander, Ignavier Ng, Junjian Ye, Zhitang Chen, and Lujia Pan. gcastle: A python toolbox for causal discovery, 2021.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *arXiv preprint arXiv:1803.01422*, 2018.

Kai Zhong, Ian En-Hsu Yen, Inderjit S Dhillon, and Pradeep K Ravikumar. Proximal quasi-newton for computationally intensive l1-regularized m-estimators. *Advances in Neural Information Processing Systems*, 27, 2014.

Shengyu Zhu, Ignavier Ng, and Zhitang Chen. Causal discovery with reinforcement learning. *arXiv preprint arXiv:1906.04477*, 2019.

Appendix A. Ground-truth graphs

LUCAS The ground truth graph of LUCAS dataset: http://www.causality.inf.ethz.ch/data/LUCAS.html

ASIA The ground truth graph of ASIA dataset: https://www.bnlearn.com/bnrepository/discrete-small.html#asia

SACHS The ground truth graph of SACHS dataset: https://www.bnlearn.com/bnrepository/discrete-small.html#sachs

Oxygen-therapy The ground truth graph of the Oxygen-therapy dataset can be found in the following paper: https://arxiv.org/abs/2010.14774

Appendix B. Baselines

The implementation of the baselines that we considered for our experiments have been adopted from the repository gCastle (Zhang et al. (2021)) which is a toolchain that contains several baseline causal discovery algorithms. The link to the repository: https://github.com/huawei-noah/trustworthyAI/tree/master/gcastle. A brief discussion of the baselines is given below:

- NOTEARS (Zheng et al. (2018)) formulates the causal structure learning problem as a purely continuous optimization problem and provides a novel characterization of acyclicity that allows for a smooth, global search, as opposed to a combinatorial, local search. It uses the acyclicity function and a weighted adjacency matrix with the least squares loss to find the causal graph.

- GOLEM (Ng et al. (2020)) formulates a likelihood-based score function for causal discovery with continuous unconstrained optimization that studies the asymptotic role of the sparsity and DAG constraints for learning DAG models in the linear Gaussian and non-Gaussian cases. It is a more efficient version of NOTEARS since it can reduce the number of optimization iterations.

- GraN-DAG (Lachapelle et al. (2019)) is a score-based approach that extends NOTEARS to deal with non-linear causal relationships using neural networks (NNs) and formulates a novel characterization of acyclicity for NNs. It works well in case of non-linear Gaussian additive noise models.

- RL-BIC2 (Zhu et al. (2019)) is score-based causal discovery approach that uses reinforcement learning (RL) to search for the DAG with the optimal score. They use a predefined score-function (BIC score) and an acyclicity constraint from NOTEARS to formulate reward for the generated graphs.

- ANM (Hoyer et al. (2008)) proposes a the linear–non-Gaussian causal discovery framework that can be generalized to address nonlinear functional dependencies with additive noise models. It states that nonlinearities in the data-generating process are rather useful, as they typically provide information on the underlying causal mechanism.

- ICALiNGAM (Shimizu et al. (2006)) uses independent component analysis to discover the causal structure of continuous-valued data. It assumes that the data generating process is linear, there are no unobserved confounders, and the noises are non-Gaussian distributions of non-zero variances. It usually performs well on LiNGAM datasets and does guarantee performance in case of linear Gaussian datasets.

- DirectLiNGAM (Shimizu et al. (2011)) proposes a direct method to estimate a causal ordering of variables on LiNGAM datasets by successively reducing each independent component from given data in the model. The total steps to complete the process is equal to the number of the variables in the model and it assumes the sample size to be infinite.

Appendix C. Code

Code implementation for KCRL and experimental data is available in the following link: https://github.com/UzmaHasan/KCRL

Appendix D. Penalty weight/parameter

Generally the concept of penalizing attempts to approximate a constrained optimization problem with an unconstrained one and then apply standard search techniques to obtain solutions. The approximation is accomplished by adding a term to the objective function that states a high cost for violation of the constraints (Luenberger and Ye (2016)). If the value of the penalty parameter is made suitably large, the penalty term will exact such a heavy cost for any constraint violation that the minimization of the objective function will yield a feasible solution. However, how much large depends on the particular model. Since, it is almost always impossible to tell how large the parameter must be to provide a solution to the problem. Hence we need to initialize the penalty weight with a relatively small value. This will assure that no steep valleys are present in the initial optimization of the target function. Subsequently, the goal is to solve a sequence of unconstrained problems with strictly increasing values of the penalty weight chosen so that the solution to each new problem is close to the previous one. In our case the initial value of the penalty weight β was chosen as 0.1 with an upper limit 1. After each 1000 epochs, its value increases by 0.1.