ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10445823

GRAPH IDENTIFICATION AND UPPER CONFIDENCE EVALUATION FOR CAUSAL BANDITS WITH LINEAR MODELS

Chen Peng, Di Zhang† *and Urbashi Mitra*

Electrical & Computer Engineering, † Industrial Systems Engineering, University of Southern California

ABSTRACT

In this paper, the causal bandit problem is investigated, in which the objective is to select an optimal sequence of interventions on nodes in a graph. By exploiting the causal relationships between the nodes whose signals contribute to the reward, interventions are optimized. First, a method to learn the directed acyclic graph is proposed that strongly reduces sample complexity relative to the prior art and adopts a novel edge detection method based on mutual information by learning sub-graphs. It is assumed that the graph is governed by linear structural equations; it is further assumed that the distribution of interventions is unknown. Under the assumption of Gaussian exogenous inputs and minimum-mean squared error weight estimation, a new uncertainty bound tailored to the causal bandit problem is derived. This uncertainty bound drives an upper confidence bound based intervention selection to optimize the reward. Numerical results compare the new methodology to existing schemes and show a substantial performance improvement.

Index Terms— Causal bandit, linear structure equation model, causal graph identification, upper confidence bound, mutual information

1. INTRODUCTION

The multi-armed bandit (MAB) is a useful model for sequential decision-making problems with applications to clinical trials [1], recommendation systems [2], financial portfolio design [3], *etc*. In MAB problems, an agent selects an arm in each step and observes corresponding outcomes to maximize the long-term cumulative reward. In the classic setting, the stochastic rewards generated by different arms are assumed to be statistically independent. To model realistic scenarios with dependence, *causal* bandits are considered [4]. The causal structure can be exploited to improve decision-making.

Cause-effect relationships can be represented by Bayesian networks, in the form of directed acyclic graphs (DAGs). DAGs can encode the causal relationship among factors that contribute to the reward [4]. We interpret the arms as different interventions on the nodes of a DAG and the reward as the stochastic outcome of a certain node. The objective is to maximize the cumulative reward by selecting a sequence of interventions.

Existing literature on causal bandit can be categorized based on their assumptions about the causal graph topology and the probability distribution of the interventions. While many works assume topology or DAG knowledge [4–8], it is not known in practice. The setting with no knowledge of topology and interventional distribution has been investigated recently [9, 10], where algorithms are proposed with improved regret guarantees over non-causal schemes. However, these prior works are based on the hard intervention model, where the causal relations between a node and its parents are completely cut off. Herein, we consider soft interventions.

To solve causal bandit problems, the major challenges are graph identification and the exploration-exploitation balance. For graph identification, existing methods mainly fall into two categories, independence-based and score-based [11]. Notice that solving causal bandit problems solely by generic graph identification is inefficient because optimal intervention selection is not equivalent to graph identification. Regarding the exploration-exploitation balance, the upper confidence bound (UCB) is a widely adopted method that combines current estimates with future potential. The classic UCB scheme uses the number of visits as a general uncertainty measure [12], while in causal bandits, uncertainty can be better quantified by bounding the variance of problem-specific estimators [8–10].

In this paper, we propose the Causal Sub-graph UCB (CS-UCB) scheme that assumes no knowledge of the causal graph topology and interventional distributions. The algorithm learns the *critical* causal structure (defined in the sequel), and utilizes causal knowledge for decision-making, in an alternating manner. The main contributions of this paper are:

1. We first propose learning sub-graphs versus the entire graph, which dramatically reduces compute and sample complexities.
2. For the sub-graph learning problem, we propose an edge-weighted mutual information measure for edge detection. This framework makes efficient use of limited data to learn the critical part of the causal structure,

This work is funded in part by one or more of the following grants: NSF CCF-1817200, ARO W911NF1910269, DOE DE-SC0021417, Swedish Research Council 2018-04359, NSF CCF-2008927, NSF CCF-2200221, ONR 503400-78050, ONR N00014-15-1-2550, NSF CCF 2200221 and USC + Amazon Center on Secure and Trusted Machine Learning.

such that only the crucial edges are learned accurately.
3. An uncertainty bound tailored to the causal bandit framework is derived and used to drive an UCB strategy to resolve the exploration-exploitation dilemma.
4. Numerical results indicate that the proposed algorithm identifies the optimal intervention much faster than standard MAB schemes by exploiting the causal structure. Moreover, compared to strategies that only focus on graph identification, the proposed algorithm has low sample complexity and learns the causal structure with a limited loss of the long-term reward.

2. SYSTEM MODEL

2.1. The Causal Graph Model with Soft Intervention

The observational causal structure is represented by a DAG, $(\mathcal{V}, \boldsymbol{B})$, where $\mathcal{V}$ is the set of N nodes and $\boldsymbol{B}$ is the edge-weight matrix. We consider soft interventions, defined as

$$\boldsymbol{a} = (a_1, \ldots, a_N)^T \in \{0,1\}^N \equiv \mathcal{A}, \tag{1}$$

where a_i represents whether node i is intervened (1) or not (0). Different from hard interventions, soft interventions do not cut off causal relationships between the intervened node and its parents, but change the upcoming edges to the node.

Further, we denote the interventional weight matrix by $\boldsymbol{B}'$ so that the post-intervention matrix $\boldsymbol{B_a}$ can be constructed as

$$[\boldsymbol{B_a}]_i = \mathbf{1}(a_i = 1)\,[\boldsymbol{B}']_i + \mathbf{1}(a_i = 0)\,[\boldsymbol{B}]_i\,, \tag{2}$$

where $\mathbf{1}(\cdot)$ is the indicator function and $[\cdot]_i$ represents the i-th column of a matrix. Denote the set of parents of node i by $\mathrm{pa}(i, a_i)$ and the set of ancestors by $\mathrm{an}(i, a_i)$. The i-th column of the post-intervention weight matrix determines $\mathrm{pa}(i, a_i)$ and how these parents influence node i.

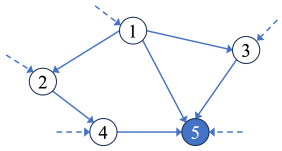
With intervention, the vector of stochastic values associated with the nodes is represented by $X = (X_1, \ldots, X_N)^T$. The causal relationship among nodes is described by a linear structure equation model (LinSEM),

$$X = (\boldsymbol{B_a})^T X + \boldsymbol{\epsilon}, \tag{3}$$

where $\boldsymbol{\epsilon}$ is a vector of Gaussian noise/exogenous variables, independent of X. We assume $\boldsymbol{\epsilon}$ has independent elements with known mean vector $\boldsymbol{\nu}$ and unknown covariance.

2.2. The Causal Bandit Model

In the MAB framework, an agent performs a sequence of actions to maximize cumulative reward over a finite horizon T. We consider node N as the reward node in the causal graph model, which generates stochastic rewards in each time step. An example causal graph is given in Fig. 1, where the value of the solid node is considered as the reward and the effects of exogenous variables are represented by dashed arrows.



	Parent	Ancestor
Node 1	∅	∅
…	…	…
Node 5	{1,3,4}	{1,2,3,4}

Fig. 1. A causal graph ($N = 5$) with the reward node 5.

To compute the expected reward under intervention $\boldsymbol{a}$, we recognize that in LinSEM, there exists a causal flow between every ancestor-descendant pair. Thus each X_i can be written as a linear combination of exogenous variables in $\boldsymbol{\epsilon}$, weighted by the causal flow. Define the flow-weight matrix as

$$\boldsymbol{C_a} \doteq (\boldsymbol{I} - \boldsymbol{B_a})^{-1}\,, \tag{4}$$

where the (i, j)-th entry represents the net flow weight from node i to j. In this way, we rewrite (3) as

$$X = (\boldsymbol{I} - \boldsymbol{B_a})^{-T}\,\boldsymbol{\epsilon} = (\boldsymbol{C_a})^T\,\boldsymbol{\epsilon}, \tag{5}$$

where $\boldsymbol{I}$ denotes the identity matrix. The expectation of X under intervention $\boldsymbol{a} \in \mathcal{A}$ is formulated as

$$\boldsymbol{\mu_a} \doteq \mathbb{E}\left[(\boldsymbol{I} - \boldsymbol{B_a})^{-T}\,\boldsymbol{\epsilon}\right] = (\boldsymbol{I} - \boldsymbol{B_a})^{-T}\,\boldsymbol{\nu}. \tag{6}$$

Thus with the knowledge of post-intervention weight matrices, the optimal intervention can be obtained as

$$\boldsymbol{a}^* \doteq \arg\max_{\boldsymbol{a} \in \mathcal{A}}\ [\boldsymbol{\mu_a}]_N\,. \tag{7}$$

In each time step t, the agent selects an intervention $\boldsymbol{a}^t$, observes X^t and collects reward X_N^t. The randomness of the observation comes from the exogenous variables $\boldsymbol{\epsilon}^t$, which are independent of the intervention. The objective is to maximize the expected cumulative reward, $\sum_t [\boldsymbol{\mu}_{\boldsymbol{a}}^t]_N$.

3. THE CS-UCB ALGORITHM

In this section, we introduce two major components of the proposed CS-UCB algorithm: sub-graph learning and uncertainty bound based decision-making.

3.1. Causal Sub-graph Learning

Causal graph identification is equivalent to determining the causal relationships between every pair of nodes in the graph, represented by directed edges. To estimate the causal graph, we adopt a *score-based* method. Specifically, given observations up to step t, $\boldsymbol{X}^{1:t} \doteq (X^1, \ldots, X^t)$ of dimension $N \times t$, we rate the ability of different graph structures to fit the data.

However, finding the exact structure is difficult in practice because the number of DAGs grows super-exponentially with the number of nodes. The issue becomes even more serious for causal bandits, because the agent only has t observations by which to score 2^N possible distinct graphs. To

alleviate computational complexity, we notice that although there are 2^N post-intervention distributions characterized by $\boldsymbol{B}_{\boldsymbol{a}}$, those weight matrices are composed of columns of $\boldsymbol{B}$ and $\boldsymbol{B}'$. Therefore, instead of identifying 2^N graphs, we can identify causal relationships induced by columns of $\boldsymbol{B}$ and $\boldsymbol{B}'$. Since the complete edge set is uniquely decomposable into sub-graph edge sets, we claim that identifying the complete graph is equivalent to identifying all the sub-graphs.

As a foundation of causal reasoning, the principle of *independent mechanisms* [11] states that the causal variables and the mechanism producing the effect variable are independent. Thus, testing the independence of residuals has been investigated for causal inference [13–15]. Herein, we consider the minimum mean-square error (MMSE) estimation, with the estimator and residual defined as

$$\hat{X}_i^t(a_i) \doteq \mathbb{E}\left[X_i \middle| \boldsymbol{X}^{1:t}_{\hat{\mathrm{pa}}(i,a_i)}\right], \quad R_i^t(a_i) \doteq \hat{X}_i^t(a_i) - X_i, \tag{8}$$

where $\hat{\mathrm{pa}}(i, a_i)$ represents the estimated parent set. Also, we denote the estimated edge-weight and flow-weight matrices at step t by $\hat{\boldsymbol{B}}_{\boldsymbol{a}}^t$ and $\hat{\boldsymbol{C}}_{\boldsymbol{a}}^t$. With the estimated mechanism producing X_i, the principle of independent mechanisms implies that the residual should be independent from any causal covariate. Therefore, the dependence between R_i^t and X_j provides information about the causal relation between node i and j.

As a general dependence measure, mutual information, denoted by $I(\cdot)$, is widely utilized for causal inference (see *e.g.* [11]). However, for a multivariate model with limited data, pure mutual information based methods suffer from stochastic errors. The next proposition[1] illustrates this issue.

Proposition 1. With the signal model defined in Section 2.1, the following inequalities hold, $\forall \boldsymbol{a}$,

$$I(R_i^t(a_i); X_j) \leq I\Big(\sum_{l \notin \mathrm{an}(i,a_i)} \big[\hat{\boldsymbol{C}}_{\boldsymbol{a}}^t\big]_{li} \epsilon_l - \epsilon_i; \sum_{l \notin \mathrm{an}(i,a_i)} [\boldsymbol{C}_{\boldsymbol{a}}]_{lj}\, \epsilon_l\Big) + I\Big(\sum_{k \in \mathrm{an}(i,a_i)} \big[\hat{\boldsymbol{C}}_{\boldsymbol{a}}^t - \boldsymbol{C}_{\boldsymbol{a}}\big]_{ki} \epsilon_k; \sum_{k \in \mathrm{an}(i,a_i)} [\boldsymbol{C}_{\boldsymbol{a}}]_{kj}\, \epsilon_k\Big), \tag{9}$$

$$I\Big(\sum_{l \notin \mathrm{an}(i,a_i)} \big[\hat{\boldsymbol{C}}_{\boldsymbol{a}}^t\big]_{li} \epsilon_l - \epsilon_i; \sum_{l \notin \mathrm{an}(i,a_i)} [\boldsymbol{C}_{\boldsymbol{a}}]_{lj}\, \epsilon_l\Big) \leq \log\Big(1 + \frac{\sigma_{E1}}{\sigma_{E2}} \big|\big[\hat{\boldsymbol{B}}_{\boldsymbol{a}}^t\big]_{ji}\big|\Big) - \frac{1}{2}\log(1 - \rho_E^2), \tag{10}$$

where σ_{E1}^2, σ_{E2}^2 are the variance of the following variables

$$E(X_j, i) \doteq \sum_{l \notin \mathrm{an}(i,a_i)} [\boldsymbol{C}_{\boldsymbol{a}}]_{lj}\, \epsilon_l, \tag{11}$$

$$E(R_i^t(a_i), \backslash j) \doteq \sum_{l \notin \mathrm{an}(i,a_i)} \big[\hat{\boldsymbol{C}}_{\boldsymbol{a}}^t\big]_{li} \epsilon_l - \epsilon_i - \big[\hat{\boldsymbol{B}}_{\boldsymbol{a}}^t\big]_{ji} E(X_j, i), \tag{12}$$

and ρ_E stands for the correlation between them.

[1] See https://github.com/Chen-Peng-98/Causal-Bandit-Supplementary.git for the proofs.

In (9), the first part is induced by causal error while the second part is induced by intrinsic error. Observe that for a small estimated edge weight, the causal error induced mutual information could be hard to detect, especially when the intrinsic error induced mutual information is not negligible. Since the goal is to detect causal errors, we propose an edge-weighted measure,

$$I_{\mathrm{w}}(R_i^t(a_i); X_j) \doteq I(R_i^t(a_i); X_j) - \log\big|\big[\hat{\boldsymbol{B}}_{\boldsymbol{a}}^t\big]_{ji}\big|, \tag{13}$$

that can be used as a criterion to reject incorrect edges from sub-graphs. To estimate empirical mutual information, the k-nearest neighbour distance based method is employed [16].

Since a complete graph is required to evaluate interventions, we make edge rejecting decisions by considering sub-graphs jointly. The remaining problem is to determine the number of edges to reject. In general, any residual based test for causal identification suffers from type I error [11], thus the threshold should be chosen carefully. In terms of finding the optimal intervention, rejecting an actual edge is worse than accepting a nonexistent edge. Thus we consider the estimated observational graph as a *critical* failure if

$$\exists\, [\boldsymbol{B}]_{ij} \neq 0, \quad \big[\hat{\boldsymbol{B}}^t\big]_{ij} = 0, \tag{14}$$

or $\hat{\boldsymbol{B}}^t$ does not represent a DAG. The same criterion is used to evaluate the estimated interventional graph.

To minimize the critical failure rate, an edge is only rejected when necessary. Specifically, the edge-weighted mutual information is calculated for each (i, j) pair and the one with the largest I_{w} is rejected, until the complete graph becomes a DAG. Once the estimated edge set is determined, we construct the weight matrices by MMSE estimation.

3.2. Uncertainty Bound

Since uncertainty exists about the accuracy of the estimated weights, exploration is necessary. Interventions should be evaluated by both how close the estimated rewards are to being maximal and the uncertainties in those estimates. *Assuming no causal error*, we derive an uncertainty bound on the estimation error of the expected reward, so that the potential of an intervention is taken into account. Define weight error matrices and expectation error vectors as

$$\Delta \boldsymbol{B}_{\boldsymbol{a}}^t \doteq \hat{\boldsymbol{B}}_{\boldsymbol{a}}^t - \boldsymbol{B}_{\boldsymbol{a}}, \quad \Delta \boldsymbol{\mu}_{\boldsymbol{a}}^t \doteq \hat{\boldsymbol{\mu}}_{\boldsymbol{a}}^t - \boldsymbol{\mu}_{\boldsymbol{a}}, \tag{15}$$

where $\hat{\boldsymbol{\mu}}_{\boldsymbol{a}}^t$ is the estimated mean of X. Further, denote the covariance of the i-th weight error vector by $\boldsymbol{\Sigma}_{\boldsymbol{a}}^t(i)$. The following theorem[1] provides a concentration inequality for the error of the estimated reward.

Theorem 1. Under the assumption of no causal error, MMSE estimation, and the signal model in Section 2.1, we have

$$\mathbb{P}\Big\{\big|\big[\Delta \boldsymbol{\mu}_{\boldsymbol{a}}^t\big]_N\big| \geq U(\boldsymbol{X}^{1:t}, \boldsymbol{a}, \delta)\Big\} \leq \delta, \tag{16}$$

where the error upper-bound $U(\boldsymbol{X}^{1:t}, \boldsymbol{a}, \delta)$ is defined as

$$U(\boldsymbol{X}^{1:t}, \boldsymbol{a}, \delta) \doteq 2(N^2+2N)^{\frac{1}{4}} \left\| \left[(\boldsymbol{I} - \hat{\boldsymbol{B}}_{\boldsymbol{a}}^t)^{-1} \right]_N \right\|_2 \cdot \|\boldsymbol{\mu}_{\boldsymbol{a}}\|_2 \sqrt{\ln\left(\frac{2N}{\delta}\right) \sum_{i=1}^{N} \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{a}}^t(i))}. \quad (17)$$

Based on the uncertainty bound, the proposed algorithm selects intervention in each time step as

$$\boldsymbol{a}^{t+1} = \arg\max_{\boldsymbol{a} \in \mathcal{A}} \left\{ \left[(\boldsymbol{I} - \hat{\boldsymbol{B}}_{\boldsymbol{a}}^t)^{-1} \right]_N^T \boldsymbol{\nu} + \alpha U(\boldsymbol{X}^{1:t}, \boldsymbol{a}, \delta) \right\}, \quad (18)$$

where α is a parameter that controls the exploration level.

4. NUMERICAL RESULTS

In this section, we numerically evaluate the performance of the proposed CS-UCB algorithm, with LinSEM and soft intervention. We focus on graphs with size $N = 8$ where the nonzero elements of $\boldsymbol{B}$ and $\boldsymbol{B}'$ are randomly generated following the uniform distribution on $[-2, 2]$. The distribution of each exogenous variable ϵ_i is set to $\mathcal{N}(1, 1)$. The horizon length is set as $T = 1000$ and the Monte Carlo run for each set of parameters is repeated 100 times.

For comparison, first we consider the vanilla UCB algorithm, which selects interventions according to

$$\boldsymbol{a}^{t+1} = \arg\max_{\boldsymbol{a}} \left\{ \frac{\sum_\tau \mathbf{1}(\boldsymbol{a}^\tau = \boldsymbol{a}) X_N^\tau}{N_t(\boldsymbol{a})} + \alpha' \sqrt{\frac{\ln t}{N_t(\boldsymbol{a})}} \right\}, \quad (19)$$

where $N_t(\boldsymbol{a})$ denotes the number of visits and α' is the parameter controlling the exploration level. Notice that the vanilla UCB algorithm does not exploit causal structure and its sample complexity scales as 2^N. Another comparison scheme we consider is a penalized Likelihood-based causal graph identification algorithm, called GOLEM [17]. Since the GOLEM algorithm identifies the whole graph, intervention is required to be the all-zeros vector for learning $\boldsymbol{B}$ and all-ones for learning $\boldsymbol{B}'$. In order to adapt the GOLEM algorithm to causal bandits, we divide the horizon into two parts. The first part is dedicated to graph identification while the second part is for earning rewards, with interventions selected according to

$$\boldsymbol{a}^{t+1} = \arg\max_{\boldsymbol{a}} \left\{ \left(\boldsymbol{I} - \tilde{\boldsymbol{B}}_{\boldsymbol{a}}^t \right)^{-T} \boldsymbol{\nu} \right\}, \quad (20)$$

where $\tilde{\boldsymbol{B}}_{\boldsymbol{a}}^t$ denotes the weight matrix learned by GOLEM.

Figure 2 plots the cumulative reward as a function of time for three different algorithms. Also, we plot the percentage of selecting the optimal intervention in Fig. 3. The vanilla UCB algorithm gains some information about the optimal intervention after exploring every intervention with $2^N = 256$ time steps. By the end of the horizon, the vanilla UCB selects the optimal intervention 69% of the time. The GOLEM-MAB algorithm tries to identify the causal graph in the first 400 steps, and exploits this knowledge to achieve an optimal intervention selection ratio of 73%. The proposed CS-UCB algorithm gains causal knowledge and utilizes it in an alternating manner, to achieve an optimal intervention selection ratio of more than 70% after the first 100 steps. In terms of cumulative rewards, the CS-UCB algorithm achieves a 27.3% improvement compared with the vanilla UCB algorithm and a 49.9% improvement compared with the GOLEM-MAB algorithm.

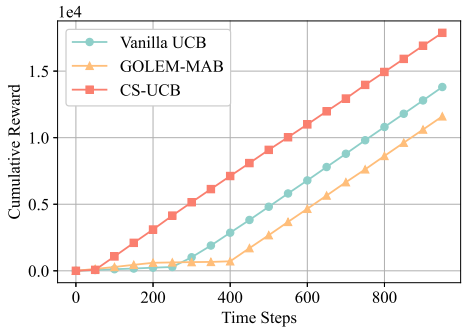


Fig. 2. Cumulative reward as a function of time steps.

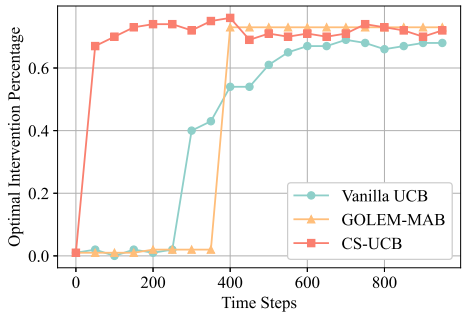


Fig. 3. The percentage of selecting the optimal intervention as a function of time steps.

5. CONCLUSIONS

In this paper, we investigated the causal bandit problem without prior knowledge of the causal graph topology and the interventional distribution. The sub-graph learning-based causal identification approach is proposed, which inherently makes efficient use of the limited data to learn the critical causal structure. Moreover, we analyze and propose an uncertainty bound to balance exploration with exploitation. Numerical results show that the proposed algorithm is able to identify the optimal intervention much faster than existing methods and achieves larger cumulative reward by exploiting the causal structure effectively.

6. REFERENCES

[1] Siqi Liu, Kay Choong See, Kee Yuan Ngiam, Leo Anthony Celi, Xingzhi Sun, and Mengling Feng, "Reinforcement learning for clinical decision support in critical care: comprehensive review," *Journal of medical Internet research*, vol. 22, no. 7, pp. e18477, 2020.

[2] Qian Zhou, XiaoFang Zhang, Jin Xu, and Bin Liang, "Large-scale bandit approaches for recommender systems," in *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part I 24*. Springer, 2017, pp. 811–821.

[3] Weiwei Shen, Jun Wang, Yu-Gang Jiang, and Hongyuan Zha, "Portfolio choices with orthogonal bandit learning," in *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[4] Finnian Lattimore, Tor Lattimore, and Mark D Reid, "Causal bandits: Learning good interventions via causal inference," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[5] Yangyi Lu, Amirhossein Meisami, Ambuj Tewari, and William Yan, "Regret analysis of bandit problems with causal background knowledge," in *Conference on Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 141–150.

[6] Vineet Nair, Vishakha Patil, and Gaurav Sinha, "Budgeted and non-budgeted causal bandits," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 2017–2025.

[7] Aurghya Maiti, Vineet Nair, and Gaurav Sinha, "A causal bandit approach to learning good atomic interventions in presence of unobserved confounders," in *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

[8] Burak Varici, Karthikeyan Shanmugam, Prasanna Sattigeri, and Ali Tajer, "Causal bandits for linear structural equation models," *arXiv preprint arXiv:2208.12764*, 2022.

[9] Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari, "Causal bandits with unknown graph structure," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24817–24828, 2021.

[10] Arnoud De Kroon, Joris Mooij, and Danielle Belgrave, "Causal bandits without prior knowledge using separating sets," in *Conference on Causal Learning and Reasoning*. PMLR, 2022, pp. 407–427.

[11] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf, *Elements of causal inference: foundations and learning algorithms*, The MIT Press, 2017.

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

[13] Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf, "Distinguishing cause from effect using observational data: methods and benchmarks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1103–1204, 2016.

[14] AmirEmad Ghassami, Negar Kiyavash, Biwei Huang, and Kun Zhang, "Multi-domain causal structure learning in linear systems," *Advances in neural information processing systems*, vol. 31, 2018.

[15] Hang Wu and May D Wang, "An information theoretic learning for causal direction identification," in *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 2020, pp. 287–294.

[16] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, pp. 066138, 2004.

[17] 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*, vol. 33, pp. 17943–17954, 2020.