Report on *Dynamic Causal Bayesian Optimization*

Yutian ZHOU

1 Literature Review

Causality is a fundamental concept in scientific research that seeks to answer two critical questions: *why* and *what if*. Unlike traditional statistical methods, such as correlation analysis, which primarily focus on associations between variables, causal inference aims to uncover the underlying mechanisms of these relationships. This makes it indispensable for understanding and explaining the cause-and-effect dynamics in complex systems.

Pearl's causal graph model is a cornerstone of modern causal inference. It primarily uses Directed Acyclic Graphs (DAGs) to represent causal relationships. In a DAG, *nodes* denote variables (e.g., treatments, exposures, outcomes, or patient characteristics), while *edges* (arrows) represent direct causal relationships. DAGs provide a formal, intuitive framework to hypothesize and communicate causal assumptions, aiding research in domains like clinical epidemiology, social science, and economics. These graphs are particularly useful in addressing questions related to causation, mediation, and interaction (Greenland *et al.*, 1999; Glymour, 2006).

To construct a DAG, researchers define the causal variables of interest, including the exposure (or treatment) E and the outcome D, while also identifying confounders, mediators, and other variables that may affect the causal structure. Even unmeasured variables, which might influence E or D, should be represented in the DAG if they are hypothesized to exist. This ensures a comprehensive depiction of the assumed causal mechanisms.

A key feature of Pearl's framework is the **do-operator**, denoted as $\text{do}(X = x)$, which represents an external intervention on a variable X. The *do-calculus* connects observational distributions to interventional ones, enabling the estimation of causal effects without performing explicit experiments (Pearl, 1995). This approach is particularly valuable in fields where experiments are impractical, such as social science and economics.

The main tasks of causal inference can be categorized into three aspects:

- **Causal Discovery:** Identifying causal structures from observational or experimental data.
- **Causal Effect:** Quantifying the impact of an intervention on a target variable.
- **Counterfactual Inference:** Predicting potential outcomes in hypothetical scenarios

In the following sections, we will explore two key aspects of causal inference in more detail.

In recent years, causal inference has become a key research focus, emphasizing causality over correlation to better identify spurious relationships. Models that adopt a causal perspective tend to exhibit stronger generalization and transferability, making them applicable to a wide range of scenarios. Many studies have focused on developing more efficient causal inference algorithms, leveraging machine learning techniques to handle high-dimensional data, address sparse data and complex causal relationships, and identify and reduce the influence of unobserved or unmanipulable variables.

The article *Dynamic Causal Bayesian Optimization*(Aglietti *et al.*, 2021) focuses on causal decision making, I will introduce the related literature in Section 1.1. Causal decision making is part of casual effect research. Additionally, in Section 1.2, I will briefly introduce literaure on causal Discovery paper that I find very interesting, which is also part of the second section.

1.1 Causal Effect: Causal Decision Making

,

In the field of causal decision-making, several studies have integrated multi-armed bandits (MAB) and reinforcement learning (RL) with causal inference (Bareinboim *et al.*, (2015), Buesing *et al.*, (2018), Lee and Bareinboim (2018), Lee and Bareinboim (2019) and Lu *et al.*, (2018)). These works focus on scenarios where actions or arms correspond to interventions on causal graphs with

complex dependencies between decisions and rewards. For MAB, Lee and Lee and Bareinboim (2018) identified a set of possibly-optimal arms that an agent should play to maximize its expected reward. This work was later extended to causal graphs with non-manipulable variables (Lee and Bareinboim, 2019). For RL, Lu *et al.*, (2018) combines RL with causal inference to address confounders in observational data by estimating latent-variable models to account for hidden factors. This framework modifies standard RL algorithms, such as Actor-Critic, into their deconfounding variants and has demonstrated superior performance in confounded environments.

In contrast to these approaches, the current paper integrates causal inference with Bayesian Optimization(BO) and explores its application in dynamic settings. BO is a widely used heuristic for optimizing costly objective functions without explicit functional forms (Jones *et al.*, (1998) and Shahriari *et al.*, (2015)). However, traditional BO assumes independence between input variables, which can disrupt the causal dependencies in the data and potentially lead to suboptimal solutions. Intervening on a subset of variables, by contrast, can propagate effects through the causal graph and achieve better outcomes. Causal Bayesian Optimization(CBO), introduced by Aglietti *et al.*, (2020), addresses this limitation by incorporating causal dependencies between variables and integrating both observational and interventional data. This approach reduces uncertainty in causal effect estimation and optimizes decision-making processes, such as identifying the most effective interventions or treatments while minimizing the number of required experiments. Also, in CBO, actions do not affect the environment directly but rather adjust the distribution after intervention.

Building on the foundation of CBO, this paper extends the framework to dynamic settings, enabling more sophisticated optimization by accounting for temporal dependencies in causal structures and improving decision-making processes in sequential scenarios.

A recent extension of this line of work is introduced in Sussex *et al.*, (2023). ACBO generalizes CBO by considering scenarios where external agents or events also intervene on the system. This generalization is particularly relevant for non-stationary environments influenced by external factors such as weather changes, market forces, or adversarial behaviors. In ACBO, the downstream reward is impacted not only by the agent's interventions but also by potentially adversarial interventions on certain nodes in the causal graph, which may only be observed retrospectively.

1.2 Casual Discovery

Causal discovery primarily involves two key tasks: learning the topological structure of causal relationships (e.g., Choo *et al.*, (2022), Lorch *et al.*, (2022) and Huang *et al.*, (2022)) and extracting causal relationships from relevant data using algorithmic methods. This process is particularly important in the data analysis of recommendation systems, especially for tasks such as causal feature analysis, effect evaluation, and interpretability studies (e.g., Jalaldoust *et al.*, (2022) and Günther *et al.*, (2022)).

Traditional causal discovery methods often assume the absence of latent variables, a simplification that rarely aligns with real-world scenarios. To address this, recent research has developed methods that explicitly account for latent variables. Broadly, these approaches focus on either identifying causal relationships among observed variables in the presence of latent variables or detecting latent variables and reconstructing their causal structures. For instance, Li *et al.*, (2023) introduces techniques to recover the full causal structure involving both observed and latent variables under weaker assumptions, and further establishes the identifiability of linear latent variable models. For the perspective of gragh structure, T.-Z. Wang and Zhou (2021) proposes a method called ACIC (Active target effect Identification with latent Confounding). This approach learns the necessary local structure of the causal graph using a partially mixed ancestral graph, which can efficiently represent causal relationships in the presence of latent confounders. By combining observational data with a few active interventions, ACIC identifies the causal effects of intervened variables on the response variable under conditions where only the response variable is observable. The method leverages a graphical characterization that enables efficient estimation of causal effects using a generalized back-door criterion.

Several frameworks have advanced the field by tackling the challenges posed by latent variables. Dong *et al.*, (2023) leverages rank information from the covariance matrix of measured variables, enabling it to handle hidden variables causally linked to nearly any part of the network. This framework also establishes theoretical conditions for identifying certain latent structures. Similarly, Jin *et al.*, (2023) examines systems where both measured and latent variables form a partially observed linear causal structure. The study concludes that, aided by higher-order statistics, the causal graph is nearly fully identifiable if each latent group contains a sufficient number of latent or measured variables. Another important advancement in causal discovery is the development of causal sensitivity analysis, which ensures robust causal conclusions in the presence of unobserved

confounders. Frauen *et al.*, (2023) proposes a generalized framework for sensitivity analysis, offering mathematical guarantees for causal reasoning even when unobserved confounding factors are present.

Algorithmic methods have practical significance, enabling better understanding of feature relationships, intervention effects, and model interpretability. For instance, Jalaldoust *et al.*, (2022) investigates causal structures in event data, which is often used in recommendation contexts. Similarly, Günther *et al.*, (2022) addresses the challenges of heteroskedasticity in causal inference, a common issue in real-world datasets. An additional significant contribution to causal discovery is the method proposed in Y. Wang *et al.*, (2024). This study highlights the limitations of relying solely on observational data, which often falls short due to assumptions such as the Markov property and faithfulness. Active interventions are necessary to recover structural causal models, but these can be prohibitively expensive in real-world scenarios. To address these challenges, the authors propose a Bayesian optimization-based method inspired by Bayes factors, which aims to maximize the probability of obtaining decisive and correct evidence.

2 Solution to Part 1

2.1 Exploration set

Overall, the calculation of the *Exploration Set* requires knowledge of the structural equation model(SEM), i.e., the specific relationships between variables, as well as the values of exogenous variables. Alternatively, it requires at least the expected values of outputs under different interventions. The general case implementation can be found in the program ***ES.py*** mentioned in the GitHub repository.

As mentioned in the paper, the selection of the exploration set can refer to Lee and Bareinboim (2018), which introduces definitions of MIS(Definition 3.1 below) and POMIS. MIS can be used as exploration set.

Definition 3.1 (Lee and Bareinboim (2018)). Minimal Intervention set (MIS). Given $\langle \mathcal{G}, \mathbf{Y}, \mathbf{X}, \mathbf{C} \rangle$*, a set of variables* $\mathbf{X}_s \in \mathcal{P}(\mathbf{X})$ *is said to be a mis if there is no* $\mathbf{X}'_s \subset \mathbf{X}_s$ *such that* $\mathbb{E}\left[Y \mid do\left(\mathbf{X}_s = \mathbf{x}_s\right)\right] = \mathbb{E}\left[Y \mid do\left(\mathbf{X}'_s = \mathbf{x}'_s\right)\right]$.

However, the simplified conclusion about how to find the MIS(Proposition 1 below) in this paper cannot be directly applied in our case, as it is only valid in situations where there is no non-manipulable variables. In our scenario, we need to account for the presence of non-manipulable variables.

Proposition 1 (Lee and Bareinboim (2018)). A set of variables $\mathbf{X} \subseteq \mathbf{V} \backslash \{Y\}$ *is a minimal intervention set for G with respect to Y if and only if* $\mathbf{X} \subseteq an\left(Y\right)_{G_{\mathbf{X}}}$.

Proposed Solutions:

i For the general case, we can use the function ***compute mis*** in ***ES.py*** in the GitHub repository, which is written based on the definitions 3.1 above. The *input* includes the graph structure(graph, non-manipulable, manipulable, target), the structural equation model, and the intervention values ans number of samples to be tested. The *output* is the MIS set, which can be used as the exploration set.

ii For special cases, we can temporarily treat all variables as manipulable variables, assuming there are no non-manipulable variables. And calculate the corresponding set $\hat{MIS}$. When $\hat{MIS}$ does not contain non-manipulable variables, $\hat{MIS}$ is the MIS we need which is the exploration set, using the description in the Lee and Bareinboim (2019): *denote* $\mathbb{M}^{\mathbb{N}}_{\mathcal{G},Y}$ *as set of MISs given* $\langle \mathcal{G}, Y, \mathbf{N} \rangle$ *where* $\mathbf{N}$ *is the nonmanipulation set and we omit* $\mathbf{N}$ *if* $\mathbf{N} = \emptyset$*. Then* $\mathbb{M}^{\mathbb{N}}_{\mathcal{G},Y} = \{\mathbf{W} \in \mathbb{M}_{\mathcal{G},Y} \mid \mathbf{W} \cap \mathbf{N} = \emptyset\}$. The MIS calculation for the case without non-manipulable variables can be computed using the theorem mentioned in Lee and Bareinboim (2018) and the code is in `https://github.com/sanghack81/SCMMAB-NIPS2018`. However, this special case is of limited applicability.

Additional Notes:

Regarding the definitions and theorems in the structure article, I believe that the theorem and the definition of MIS do not necessarily align perfectly. This is because the article does not specify the definition of ancestors of variables. The expected value of the output variable under intervention does not fully represent the entire distribution. If the theorem defines ancestors as the commonly understood "all nodes reachable by tracing back along directed edges", rather than being derived from the perspective of output expectations (which is part of the MIS definition), the two may not be completely equivalent.

For example, consider the following scenario:

Example: Take a causal graph: X to Y and Z to Y, where X and Z are independent binary variables $\{0, 1\}$. Assume that the generation mechanism for Y is as follows: Y = X XOR Z (XOR represents the exclusive OR operation).

According to the MIS Definition 3.1 (Lee and Bareinboim (2018)) above, X, Z is not an MIS. However, according to the Proposition 1 (Lee and Bareinboim (2018)) on , $\{X, Z\}$ is an MIS.

2.2 Key ideas of this paper

The main idea of the paper is to separate the variables dependent on the current period from those dependent on previous periods (as stated in Assumptions 1(2) in Aglietti *et al.*, (2021)). This separation ensures that Equation (6) in Appendix B in Aglietti *et al.*, (2021) holds, enabling the dynamic problem to be divided into two parts: one part can be treated as a static CBO problem, which does not involve interventions, while the other part represents the temporal relationship linked to the optimization results of all previous interventions, including the optimal interventions from earlier periods.

Essentially, as described in Definition 2.1 in Aglietti *et al.*, (2020), the parent variable of the output corresponds to the causal intrinsic dimensionality of output Y, represented by the expectation $\mathbb{E}_{P(Y|\text{do}(\mathbf{X}=\mathbf{x}))}[Y]$. Therefore, the process fundamentally involves dimensionality reduction and decoupling.

More specifically, the variable out put Y depends on a set of variables $Pa(Y_t)$. Initially, the function for Y can be expressed as $Y = f_{Y_t}(Pa(Y_t)) + \epsilon$. If this function can be decomposed into the sum of two subfunctions, such as: $f^{Y}_{Y_t}(Y_t^{PT})$ and $f^{NY}_{Y_t}(Y_t^{PNT})$,

- **Dimensionality Reduction**: The high-dimensional function f is decomposed into two lower-dimensional functions, $f^{Y}_{Y_t}$ and $f^{NY}_{Y_t}$. Each sub-function has a lower input dimensionality compared to the original function f_{Y_t}.
- **Decoupling**: The dependency of Y on $Pa(Y_t)$ is decoupled into two independent influences. This allows us to analyze and understand these two influences independently, without considering their interactions.
- **The linear additive of expectation**: From a mathematical perspective, since we need to compute the expectation of output $\mathbb{E}[Y]$, this decomposition into two components, using the property $E[f_{Y_t}] = E[f^{Y}_{Y_t}] + E[f^{NY}_{Y_t}]$. Notably, this additive form allows the expectation to be split without requiring independence between the two sets of variables, significantly simplifying the calculations.

To achieve dimensionality reduction, decoupling, and the use of the linearity of expectation in this problem, the authors separate the output of previous periods (as part of the parent variable) from other variables. This allows the problem to be divided into two parts: a recursive component related to the optimization results of previous interventions, and another component that can be treated as a single-period CBO problem (since it does not involve interventions).

Additionally, Assumptions 1 and 3 further simplify the computations.

2.3 Acquisition function

The acquisition function is problematic. The current formulation of Casual EI seeks to find interventions that maximize $y_{s,t}$. However, in Equation 1 in the Aglietti *et al.*, (2021), it requires minimizing the expected output. Therefore, the positions of $y_{s,t}$ and y_t^* should be swapped in the acquisition function, which should be as follows:

$$\text{EI}_{s,t}(\mathbf{x}) = \mathbb{E}_{p(y_{s,t})}\left[\max\left(y_t^\star - y_{s,t}, 0\right)\right] / \text{cost}\left(\mathbf{X}_{s,t}, \mathbf{x}_{s,t}\right)$$

2.4 Other errors or questionable issues

In addition to the acquisition function, there are some other minor errors in the paper. For example:

- On Page 2, in Definition 2, the term $\mathbf{C}_{0,t}$ is defined as $\mathbf{C}_{0:t} = \mathbf{C}_t \cup \mathbf{C}_{0:t-1}$, but in the Figure 2 on Page 4, it is defined as $\mathbf{C}_{0:t} = \mathbf{C}_t \cup \mathbf{C}_{0:t-1} \cup \mathbf{C} = \mathbf{Y}_{0:t-1} \cup \mathbf{X}_{0:t-1}$, which is not clearly defined.
- Misuse of notation.

  - In the last sentence in the likelihood part on Page 7, the I following *sigma* represents the identity matrix. However, it could easily be confused with the I in the for interventional data set $D^I_{s,t}$.
  - For instance, M_t in Definition 2 and the bold M_t in Section 3.2 refer to different things. M_t in Definition 2 is SCM and the bold M_t in Section 3.2 is MIS set.

- Other typos:
  - In Assumption 1(2), the formula should be $f_{Y_t}(Pa(Y_t)) = f^Y_{Y_t}(Y^{PT}_t) + f^{NY}_{Y_t}(Y^{PNT}_t)$. It miss the subscript of Y in function subscript
  - In Appendix D, the second line of the fifth equation should be corrected to $X'_{s,t} = X_{s,t} \setminus S_{s,t}$. It use the wrong subscript of X

2.5 Not clear part and other discussion

There are also some potentially important but insufficiently discussed aspects

- Since the paper assumes a known SEM (Structural Equation Model), this is difficult to achieve in real-world scenarios, as data rarely satisfies this condition. Causal relationships in the data are often not fully known. Additionally, Assumption 1 (and 3) requires the absence of unobserved variables, which is difficult to satisfy in practice.

- The interventional data assumed in the paper is generated by simply adding a perturbation term epsilon to $f_{s,t}$, but in practice, it is also challenging to collect data after interventions in reality.

- Based on my understanding, comparing with RL, the main drawback of DCBO is that it focuses only on the current-period optimum, whereas reinforcement learning learns backward (from future to past), which is the all periods optimization. Ont the other hand, the advantage lies in actions do not change the environment but only affect the interventional data. And the target is to seek the optimal action rather than a full strategy, which is more stable in practice. This distinction could be further explored and discussed.

- The balance between exploration and exploitation, *i.e.*, the value for H.

2.6 Reproduction of code

The original implementation of DCBO was carried out using GPy (GPy, since 2012). Recreating this code with TensorFlow Probability (TFP) presents challenges due to significant differences between the two libraries. The primary difficulty stems from GPy's use of an instantiated `GPRegression()` object for both training and inference, whereas TFP employs `GaussianProcess()` for model training and `GaussianProcessRegressionModel()` for making predictions. To bridge this gap, a `TFPModelWrapper()` class has been developed. This wrapper saves the observed data and the trained kernel as attributes, and implements `optimize()` and `predict()` methods to interface with the respective TFP classes.

- **Not mentioned part:** Simply optimizing the kernel parameters can lead to excessively high variance values when using the RBF kernel. Therefore, imposing a prior over the hyperparameters is necessary to stabilize training. In the implementation, a Gamma distribution with shape parameter $\alpha = 2$ and rate parameter $\beta = 0.5$ is employed as a prior, reflecting our belief about the appropriate scale of the kernel's variance.

- **Discussion according to the reproduction:** The implementation reveals the development path from BO, to CBO, and finally to DCBO. Firstly, BO utilizes a simple GP based on an RBF kernel to regress the predictions' mean and variance. Building on this, CBO enhances the BO model by incorporating interventions. It updates both observational and interventional data, thereby creating mean and variance functions to refine the prior. Furthermore, DCBO accounts for the evolution of data and variables over time. This is achieved by updating transition and emission functions, facilitating online inference.

3 How Casual Inference Matters in Finance

On one hand, causal analysis is beneficial for improving the explainability of financial models. However, it is often challenging to directly analyze causal effects in finance due to the limited size of datasets and the difficulty in observing causal relationships. Discussions around causal discovery are necessary, especially when latent variables are involved. Additionally, the dynamic setting discussed in the paper aligns well with the needs of the financial domain. Beyond static models, it is crucial to process and transmit information effectively in dynamic systems over time.

Specifically, accurate causal relationship prediction can significantly enhance price discovery in financial markets. Furthermore, a well-designed causal analysis framework can be applied in financial risk management to identify and remove confounding factors, improving the reliability and robustness of financial models.

An interesting paper, Tang *et al.*, (2020), offers insights that could be highly relevant to finance. The paper addresses the challenge of long-tailed distributions in deep learning, where models tend to favor majority (head) classes and perform poorly on minority (tail) classes. Existing solutions, such as re-sampling or re-weighting, often lack theoretical support and lead to overfitting or underfitting. To tackle this, the authors introduce a causal inference framework that analyzes the dual effects of momentum, identifying its "bad effect" as a confounder introducing bias toward head classes and its "good effect" as a mediator that improves feature learning and head-class predictions. They propose a novel method to retain the beneficial effects of momentum while removing its bias through causal intervention during training and counterfactual reasoning during inference. This ensures robust feature learning and unbiased predictions for both head and tail classes.

This study uses causal inference to address training issues caused by data imbalance. In the financial domain, data scarcity and imbalance are common problems that often lead to overfitting and poor out-of-sample performance. This paper helps identify spurious causal relationships and reduces the impact of data imbalance. It is particularly significant for modeling long-tail risks in finance, such as rare *Black Swan* events, by leveraging causal inference to improve the prediction of infrequent risk events.

References

Aglietti, Virginia *et al.*, (2020). "Causal bayesian optimization", *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 3155–3164.

Aglietti, Virginia *et al.*, (2021). "Dynamic causal Bayesian optimization", *Advances in Neural Information Processing Systems*, Vol. 34, pp. 10549–10560.

Bareinboim, Elias, Forney, Andrew, and Pearl, Judea (2015). "Bandits with unobserved confounders: A causal approach", *Advances in Neural Information Processing Systems*, Vol. 28.

Buesing, Lars *et al.*, (2018). "Woulda, coulda, shoulda: Counterfactually-guided policy search", *arXiv preprint arXiv:1811.06272*,

Choo, Davin, Shiragur, Kirankumar, and Bhattacharyya, Arnab (2022). "Verification and search algorithms for causal DAGs", *Advances in Neural Information Processing Systems*, Vol. 35, pp. 12787–12799.

Dong, Xinshuai *et al.*, (2023). "A versatile causal discovery framework to allow causally-related hidden variables", *arXiv preprint arXiv:2312.11001*,

Frauen, Dennis *et al.*, (2023). "A neural framework for generalized causal sensitivity analysis", *arXiv preprint arXiv:2311.16026*,

Glymour, M Maria (2006). "Using causal diagrams to understand common problems in social epidemiology", *Methods in social epidemiology*, pp. 393–428.

GPy (since 2012). *GPy: A Gaussian process framework in python*, `http://github.com/SheffieldML/GPy`.

Greenland, Sander, Pearl, Judea, and Robins, James M (1999). "Causal diagrams for epidemiologic research", *Epidemiology*, Vol. 10 No. 1, pp. 37–48.

Günther, Wiebke *et al.*, (2022). "Conditional independence testing with heteroskedastic data and applications to causal discovery", *Advances in Neural Information Processing Systems*, Vol. 35, pp. 16191–16202.

Huang, Biwei *et al.*, (2022). "Latent hierarchical causal structure discovery with rank constraints", *Advances in neural information processing systems*, Vol. 35, pp. 5549–5561.

Jalaldoust, Amirkasra, Hlaváčková-Schindler, Kateřina, and Plant, Claudia (2022). "Causal discovery in Hawkes processes by minimum description length", *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. No. 6, pp. 6978–6987.

Jin, Songyao *et al.*, (2023). "Structural estimation of partially observed linear non-gaussian acyclic model: A practical approach with identifiability", *The Twelfth International Conference on Learning Representations*.

Jones, Donald R, Schonlau, Matthias, and Welch, William J (1998). "Efficient global optimization of expensive black-box functions", *Journal of Global optimization*, Vol. 13, pp. 455–492.

Lee, Sanghack and Bareinboim, Elias (2018). "Structural causal bandits: Where to intervene?", *Advances in neural information processing systems*, Vol. 31.

— (2019). "Structural causal bandits with non-manipulable variables", *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. No. 01, pp. 4164–4172.

Li, Xiu-Chuan, Zhang, Kun, and Liu, Tongliang (2023). "Causal Structure Recovery with Latent Variables under Milder Distributional and Graphical Assumptions", *The Twelfth International Conference on Learning Representations*.

Lorch, Lars *et al.*, (2022). "Amortized inference for causal structure learning", *Advances in Neural Information Processing Systems*, Vol. 35, pp. 13104–13118.

Lu, Chaochao, Schölkopf, Bernhard, and Hernández-Lobato, José Miguel (2018). "Deconfounding reinforcement learning in observational settings", *arXiv preprint arXiv:1812.10576*,

Pearl, Judea (1995). "Causal diagrams for empirical research", *Biometrika*, Vol. 82 No. 4, pp. 669–688.

Shahriari, Bobak *et al.*, (2015). "Taking the human out of the loop: A review of Bayesian optimization", *Proceedings of the IEEE*, Vol. 104 No. 1, pp. 148–175.

Sussex, Scott *et al.*, (2023). "Adversarial Causal Bayesian Optimization", *The Twelfth International Conference on Learning Representations*.

Tang, Kaihua, Huang, Jianqiang, and Zhang, Hanwang (2020). "Long-tailed classification by keeping the good and removing the bad momentum causal effect", *Advances in neural information processing systems*, Vol. 33, pp. 1513–1524.

Wang, Tian-Zuo and Zhou, Zhi-Hua (2021). "Actively identifying causal effects with latent variables given only response variable observable", *Advances in Neural Information Processing Systems*, Vol. 34, pp. 15007–15018.

Wang, Yuxuan *et al.*, (2024). "Bayesian Intervention Optimization for Causal Discovery", *arXiv preprint arXiv:2406.10917*,