ABCD-Strategy: Budgeted Experimental Design for Targeted Causal Structure Discovery

Raj Agrawal MIT
Chandler Squires MIT
Karren Yang MIT
Karthikeyan Shanmugam MIT-IBM Watson AI Lab, IBM Research NY
Caroline Uhler MIT

Abstract

Determining the causal structure of a set of variables is critical for both scientific inquiry and decision-making. However, this is often challenging in practice due to limited *interventional* data. Given that randomized experiments are usually expensive to perform, we propose a general framework and theory based on optimal Bayesian experimental design to select experiments for *targeted causal discovery*. That is, we assume the experimenter is interested in learning some function of the unknown graph (e.g., all descendants of a target node) subject to design constraints such as limits on the number of samples and rounds of experimentation. While it is in general computationally intractable to select an optimal experimental design strategy, we provide a tractable implementation with provable guarantees on both approximation and optimization quality based on submodularity. We evaluate the efficacy of our proposed method on both synthetic and real datasets, thereby demonstrating that our method realizes considerable performance gains over baseline strategies such as random sampling.

1 Introduction

Determining the causal structure of a set of variables is a fundamental task in causal inference, with widespread applications not only in artificial intelligence but also in scientific domains such as biology and economics (Friedman et al., 2000; Pearl, 2003; Robins et al., 2000; Spirtes et al., 2000). One of the most common ways of representing causal structure is through a *directed acyclic graph* (DAG), where a directed edge between two variables in the DAG represents a direct causal effect and a directed path indicates an indirect causal effect (Spirtes et al., 2000).

Causal structure learning is intrinsically hard, since a DAG is generally only identifiable up to its *Markov equivalence class* (MEC) (Verma and Pearl, 1991; Andersson et al., 1997). Identifiability can be improved by performing *interventions* (Hauser and Bühlmann, 2012; Yang et al., 2018), and several algorithms have been proposed for structure learning from a combination of observational and interventional data (Wang et al., 2017; Hauser and Bühlmann, 2012; Yang et al., 2018). Since experiments tend to be costly in practice, a natural question is how principled *experimental design* (i.e., selection of intervention targets) can be leveraged to maximize the performance of these algorithms under budget constraints.

Seminal works by Tong and Koller (2001) and Murphy (2001) showed that experimental design can improve structure recovery in causal DAG models. However, these methods assume a basic framework in which experiments are performed one sample at a time. In practice, experimenters often perform a batch of interventions and collect samples over multiple rounds of experiments; and they must also factor in budget and feasibility constraints, such as on the number of unique interventions that can be performed in a single experiment, the number of experimental rounds, and the total number of samples to be collected. In genomics, for instance, genome editing technologies have enabled the collection of batches of large-scale interventional gene expression data (Dixit et al., 2016). An imminent problem is understanding how to optimally select a batch of interventions and allocate samples across these interventions, over multiple experimental rounds in a computationally tractable manner.

Since the initial works by Tong and Koller (2001) and Murphy (2001), there have been a number of new experimental design methods under budget con-

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

straints (Hauser and Bühlmann, 2014; Ghassami et al., 2018; Ness et al., 2018). These methods suffer from two drawbacks: (1) poor computational scaling (cf. Ness et al., 2018) or (2) strong assumptions including the availability of infinite observational and/or interventional data from each experiment (cf. Hauser and Bühlmann, 2014; Ghassami et al., 2018). Since it is difficult to learn the correct MEC in a limited sample setting, it is desirable to use interventional samples not only to improve identifiability but also to help distinguish between observational MECs.

Generalizing the frameworks in (Tong and Koller, 2001; Murphy, 2001; Cho et al., 2016; Hauser and Bühlmann, 2014; Ness et al., 2018), we assume the experimenter is interested in learning some function $f(G)$ of the unknown graph G. Returning to gene regulation, one might set $f(G)$ to indicate whether some gene X is downstream of some gene Y, i.e. if X is a *descendant* of Y in G. Using targeted experimental design, all statistical power is placed in learning the target function rather than being agnostic to recovering all features in the graph. In addition, we also explicitly take into account that only finitely many samples are allowed in each round, and work under various budget constraints such as a limit on the number of rounds of experimentation.

We start by reviewing causal DAGs in Section 2 and then propose an entropy-based score function that generalizes the one by Tong and Koller (2001) and Murphy (2001) in Section 3. Since optimizing this score function is in general computationally intractable, we propose our *ABCD-Strategy* consisting of approximations via *weighted importance sampling* and greedy optimization in Section 4. We also provide guarantees for this algorithm based on *submodularity*. Further, in contrast to earlier score functions, we show that our proposed score function is provably *consistent*. Finally, in Section 5 we demonstrate the empirical gains of the proposed method over random sampling on both synthetic and real datasets.

2 Preliminaries

Causal DAGs: Let $G = ([p], A)$ be a *directed acyclic graph* (DAG) with vertices $[p] := \{1, \dots, p\}$ and directed edges A, where $(i, j) \in A$ represents the arrow $i \to j$. A *linear causal model* is specified by a DAG G and a corresponding set of edge weights $\theta \in \mathbb{R}^{|A|}$. Each node i in G is associated with a random variable X_i. Under the *Markov Assumption*, each variable X_i is conditionally independent of its nondescendants given its parents, which implies that the joint distribution factors as $\prod_{i=1}^{p} \mathbb{P}\big(X_i \mid \mathsf{Pa}_G(X_i)\big)$, where $\mathsf{Pa}_G(X_i)$ denotes the parents of node X_i (Spirtes et al., 2000, Chapter 4). This factorization implies a set of *conditional independence* (CI) relations; the *Markov equivalence class* (MEC) of a DAG G consists of all DAGs that share the same CI relations (Lauritzen, 1996, Chapter 3). The *essential graph* $\mathrm{Ess}(G)$ is a partially oriented graph that uniquely represents the MEC of a DAG by placing directed arrows on edges consistent across the equivalence class and leaves the other edges undirected (Andersson et al., 1997).

Learning with Interventions: Let *intervention* $I \subseteq [p]$ be a set of intervention targets. Intervening on I removes the incoming edges to the random variables $X_I := (X_i)_{i \in I}$ in G and sets the joint distribution of X_I to a new interventional distribution $\mathbb{P}^I$. The resulting *mutilated graph* is denoted by G^I. A typical choice of $\mathbb{P}^I$ is the product distribution $\prod_{i \in I} f_i(X_i)$, where each $f_i(X_i)$ is the probability density function for the intervention at X_i. We denote by $\mathcal{I}^* := \{I_1, \cdots, I_K\}$ the set of all $K \in \mathbb{N}$ allowed interventions and by $\mathcal{I} \subseteq \mathcal{I}^*$ the subset of selected interventions. An intervention $I = \emptyset$ indicates observational data. We assume that $\mathcal{I}^*$ is a *conservative family* of interventions, i.e., for any $i \in [p]$, there exists some $I_j \in \mathcal{I}^*$ such that $i \notin I_j$ (Hauser and Bühlmann, 2012). Given a conservative family of targets $\mathcal{I}$, two DAGs G_1 and G_2 are *$\mathcal{I}$-Markov equivalent* if they are observationally Markov equivalent and for all $I \in \mathcal{I}$, G_1^I and G_2^I have the same skeleta (Hauser and Bühlmann, 2012, 2015). The set of *$\mathcal{I}$-Markov equivalent* DAGs can be represented by the *$\mathcal{I}$-essential graph* $\mathrm{Ess}^{\mathcal{I}}(G)$, a partially directed graph with at least as many directed arrows as $\mathrm{Ess}(G)$ (Hauser and Bühlmann, 2012, Theorem 10).

Bayesian Inference over DAGs: In various applications, the goal is to recover a function $f(G)$ of the underlying causal DAG G given a mix of n independent observational and interventional samples $D = \{(X_{mi}, I^{(m)}) : I^{(m)} \in \mathcal{I}^*, m \in [n], i \in [p]\}$. For example, we might ask whether an undirected edge (i, j) is in A, or we might wish to discover which nodes are the parents of a node i. We can encode our prior structural knowledge about the underlying DAG through a *prior* $\mathbb{P}(G)$. The *likelihood* $\mathbb{P}(D \mid G)$ is obtained by marginalizing out θ:

$$\begin{aligned}\mathbb{P}(D \mid G) &= \int_\theta \mathbb{P}(D, \theta \mid G)\, d\theta \\ &= \int_\theta \mathbb{P}(D \mid \theta, G)\mathbb{P}(\theta \mid G)\, d\theta\end{aligned}$$

and can be computed in closed-form for certain distributions (Geiger and Heckerman, 1999; Kuipers et al., 2014). Applying Bayes' Theorem yields the *posterior distribution* $\mathbb{P}(G \mid D) \propto \mathbb{P}(D \mid G)\mathbb{P}(G)$, which describes the state of knowledge about G after observing the data D. Given the posterior, we can then compute

$\mathbb{E}_{\mathbb{P}(G|D)} f(G)$, the posterior mean of some target function $f(G)$. Note that when f is an indicator function, this quantity is a posterior probability.

3 Optimal Bayesian Experimental Design

Our goal is to learn some feature $f(G)$ of the unknown graph through experimental design under budget constraints such as limited number of experimental rounds. In principle, this question can be answered using *optimal Bayesian experimental design*, namely by selecting the experiment that maximizes the expected value of some *utility function* U, where the expectation is with respect to hypothetical data generated according to our current beliefs (Chaloner and Verdinelli, 1995). Here, the expected utility function U is a function defined on multisets of $\mathcal{I}^*$:

Definition 3.1. The *expected utility* $U^f(\xi; D)$ of a multiset of interventions $\xi \in \mathbb{Z}^{\mathcal{I}^*}$ for learning a function $f(G)$ given currently collected data D is given by

$$\begin{aligned} U^f(\xi; D) &= \mathbb{E}_{y\sim\mathbb{P}(y|D,\xi)}\, U^f(y, \xi; D) \\ &= \mathbb{E}_{G,\theta|D}\mathbb{E}_{y|G,\theta,\xi}\, U^f(y, \xi; D), \quad y \in \mathbb{R}^{|\xi|}, \end{aligned} \tag{1}$$

where $U^f(y, \xi; D) \in \mathbb{R}$ is a function measuring the utility of observing additional samples y from a proposed design ξ and $|\xi| \coloneqq \sum_{I\in\mathcal{I}^*} |\#$ times I in $\xi|$. The *optimal Bayesian design* ξ^* under a set of design constraints C is given by

$$\xi^* \in \underset{\xi\in\mathbb{Z}^{\mathcal{I}^*}\cap C}{\arg\max}\, U^f(\xi; D). \tag{2}$$

We denote samples collected from such an optimal strategy ξ^* by D_{ξ^*}

In Definition 3.1, y is distributed according to our current beliefs $\mathbb{P}(y \mid D, \xi) = \mathbb{E}_{G,\theta|D}\left[\mathbb{P}(y \mid G, \theta, \xi)\right]$, a *mixture distribution* over (G, θ), and the utility function $U^f(y, \xi; D)$ is averaged over this distribution. There are many potential choices for $U^f(y, \xi; D)$, a popular one being *mutual information*. Tong and Koller (2001), Cho et al. (2016) and Murphy (2001) propose optimizing mutual information for the problem of recovering the full graph. More precisely, they consider the problem where $f(G) = G$ in the active learning setting, where the experimenter can adaptively collect one sample at a time. We here extend their framework to general functions $f(G)$ and the *batched setting*, where multiple samples are collected at once and the total number of batches is fixed by the experimenter. Hence, U^f must be defined on multisets instead of elements of $\mathcal{I}^*$ since multiple samples (i.e., interventions of the same type) may be collected in each batch. Note that the difficulty in solving Eq. (2) stems from the constraint set C, which renders this optimization problem combinatorial.

Recently, Ness et al. (2018) proposed a Bayesian experimental design method to work in the batched setting. The authors proposed a utility function based on the expected number of additional edges that could be oriented by performing a particular intervention given the observational MEC. This function is similar to the one proposed by Hauser and Bühlmann (2014) and Ghassami et al. (2018), in which interventions are chosen that fully identify the causal network given the MEC. Unfortunately, the algorithm in Ness et al. (2018) has *factorial* dependence on the size of the batch; in addition, we prove in Appendix A.3 that their proposed utility function is, in general, not *consistent*; see Definition 3.3 for a definition of consistency.

We therefore follow the approach taken by Tong and Koller (2001) and Murphy (2001) and consider the utility function $U^f(y, \xi; D)$ to be given by mutual information. Maximizing the mutual information is equivalent to picking the set of interventions that leads to the greatest expected decrease in entropy of $f(G)$. The mutual information utility function is given by

$$U^f_{\text{M.I.}}(y, \xi; D) \coloneqq H(f \mid D) - H(f \mid D, y = y, \xi), \tag{3}$$

where the *entropy* $H(f \mid D)$ equals

$$\begin{aligned} &\sum_{e:f(G)=e} -\mathbb{P}(f(G) = e \mid D) \log \mathbb{P}(f(G) = e \mid D), \\ &\text{and} \quad \mathbb{P}(f(G) = e \mid D) = \mathbb{E}_{\mathbb{P}(G|D)} \mathbb{1}(f(G) = e), \\ &\mathbb{P}(G \mid D) \propto \int_\theta \mathbb{P}(D \mid G, \theta)\mathbb{P}(\theta \mid G)\mathbb{P}(G). \end{aligned}$$

To better understand the behavior of $U^f_{\text{M.I.}}$, we prove the following proposition, which highlights the behavior of $U^f_{\text{M.I.}}$ in the limit of infinite samples per intervention; this is the setting studied by Hauser and Bühlmann (2014) and Ghassami et al. (2018).

Proposition 3.2. *Suppose that the Markov equivalence class $\mathcal{G}$ of G^* is known and the goal is to identify the underlying true DAG G^*. Furthermore, assume a uniform prior over $\mathcal{G}$, infinite samples per intervention $I \in \mathcal{I}$, and at most K unique interventions per batch as in Ghassami et al. (2018). Then, $U_{M.I.}$ selects the interventions*

$$\mathcal{I}_{M.I.} \in \underset{|\mathcal{I}|\le K}{\arg\min}\, \frac{1}{|\mathcal{G}|} \sum_{G\in\mathcal{G}} \log_2 |Ess^{\mathcal{I}}(G)|$$

where $|Ess^{\mathcal{I}}(G)| \coloneqq |\{G' \in \mathcal{G} : G' \in Ess^{\mathcal{I}}(G)\}|$.

This result (proof in Appendix) shows that in the limiting case, mutual information selects interventions

that lead to the finest expected log $\mathcal{I}$-MEC sizes. This limiting behavior of mutual information parallels what graph-based score functions do, such as the ones considered by Hauser and Bühlmann (2014), Ghassami et al. (2018) and Ness et al. (2018), that invoke the *Meek Rules* (Verma and Pearl, 1992) to select interventions that orient the most number of edges in the $\mathcal{I}$-essential graphs (in expectation).

A score function based on mutual information is particularly appealing since it not only has desirable properties in the infinite sample setting, but also does not require the MEC to be known, naturally handling the case of finite sample sizes. In particular, a score function based solely on Meek rules will not pick the same intervention twice by definition, since repeating the same intervention does not improve identifiability. As a result, adapting graph-based score functions in the finite sample regime requires first constructing an intervention set and then allocating samples instead of jointly picking and allocating samples. Mutual information, on the other hand, can pick the same intervention twice; for example if a particular intervention is very informative, selecting it twice and allocating more samples to it might lead to a greater expected decrease in entropy than a new intervention.

3.1 Budget Constraints

So far we have not specified the constraint set C in Eq. (2). To this end, we assume that the experimenter has a total of N samples to allocate across B batches. While one could try to optimize the partition of N samples across batches, in this work we study the simpler case where each batch b, $1 \leq b \leq B$, receives a pre-specified amount of samples N_b with $\sum_b N_b = N$. For simplifying notation we assume throughout that $N_b = \frac{N}{B}$. We leave the study of adaptive batch sizes N_b for future work. The constraint set then equals,

$$C_{N,b} \coloneqq \{\xi \in \mathbb{Z}^{\mathcal{I}^*} : |\xi| = N_b\}, \tag{4}$$

where the subscripts on C emphasize the dependence on N and b. Then, the optimal design in batch b is obtained by solving the following combinatorial optimization problem:

$$\xi_b^* \in \underset{\xi \in \mathbb{Z}^{\mathcal{I}^*} \cap C_{N,b}}{\arg\max} \; U(\xi; D_{b-1}), \tag{5}$$

where $D_{b-1} \coloneqq [D_{\xi_1^*}, \cdots, D_{\xi_{b-1}^*}]$ is all the data collected at the start of batch b and $U(\xi; D_{b-1})$ could, for example, be the mutual information defined in Eq. (3). Notice that while a particular form of $U(\xi; D_{b-1})$ is provided in Definition 3.1, $U(\xi; D_{b-1})$ need not necessarily be a Bayesian utility function to fit within the framework of Eq. (5).

We now define a natural notion of consistency for any experimental design method that can be cast as an optimization routine in the form of Eq. (5). Since the consistency of a utility function should not depend on a specific constraint set such as $C_{N,b}$, Definition 3.3 assumes the constraint set is arbitrary and set by the practitioner.

Definition 3.3. Suppose $f(G)$ is identifiable in $\text{Ess}^{\mathcal{I}^*}(G^*)$, where G^* is the true unknown DAG. Let $C_{N,b}$, $1 \leq b \leq B$ denote the constraints in batch b. A utility function $U(\xi)$ is *budgeted batch consistent* for learning a target feature $f(G)$ if

$$\mathbb{P}\left(f(G) \mid D_B\right) \xrightarrow{\mu^* \text{ a.s.}} \mathbb{1}(f(G) = f(G^*)),$$

as $N, B \to \infty$, where μ^* is the law determined by the true unknown causal DAG (G^*, θ^*)

Theorem 3.4. *$U^f_{M.I.}$ is budgeted batch consistent for* single-node interventions*, i.e., when* $\mathcal{I}^* = \{\{1\}, \cdots, \{p\}\}$.

Remark 1. While Theorem 3.4 may not be surprising (proof in the Appendix), we found that various utility functions that seem natural and have been proposed in earlier work are not consistent in the budgeted setting. In particular, in Appendix A.3 we show that the utility function proposed by Ness et al. (2018) is not consistent for single-node interventions. The main issue is that there are DAGs and constraint sets for which the same interventions keep getting selected, instead of selecting new interventions to fully identify $f(G)$.

4 Tractable Algorithm

While Section 3 provides a general framework for targeted experimental design, there are several computational challenges that we have not yet addressed. The first challenge is computing $U^f_{\text{M.I.}}(\xi; D)$. This objective function requires summing over an exponential number of DAGs and marginalizing out the edge weights θ. In this section, we discuss how to approximate $U^f_{\text{M.I.}}(\xi; D)$ by sampling graphs (either through MCMC or the *DAG-bootstrap* Friedman et al. (1999)) and using the maximum likelihood estimator of θ for each graph. Taken together, these approximations not only allow the mutual information score to be computed tractably but also lead to desirable optimization properties. In particular, we prove in Theorem 4.1 that our approximate utility function is submodular. This property enables optimizing the approximate objective in a sequential greedy fashion with provable guarantees on optimization quality.

4.1 Expectation over (G, θ)

A serious problem from a computational perspective is the expectation over (G, θ) in Definition 3.1. Since

the number of DAGs grows *superexponentially* with p, enumerating all possible DAGs is intractable. Instead, in each batch b, we propose to sample T graphs according to the posterior $\mathbb{P}(G \mid D_{b-1})$. This can be done using a variety of different *Markov chain Monte-Carlo* (MCMC) samplers; see for example Heckerman et al. (1997); Ellis and Wong (2008); Friedman and Koller (2003); Niinimaki et al. (2016); Kuipers and Moffa (2017); Madigan and York (1995); Grzegorczyk and Husmeier (2008); Agrawal et al. (2018). An alternative that is often faster but still achieves good performance, is approximating the posterior via a high-probability candidate set of T DAGs $\hat{\mathcal{G}}_T$ (Heckerman et al., 1997; Friedman et al., 1999). While there are many ways to build up this set, a popular approach is through the *nonparametric DAG bootstrap* (Friedman et al., 1999). The main idea is to subsample the data (with replacement) T times and fit a DAG learning algorithm to each of the generated datasets to construct $\hat{\mathcal{G}}_T$. Each $G \in \hat{\mathcal{G}}_T$ can then be weighted according to the ratio of unnormalized posterior probabilities,

$$w_{G,D} \coloneqq \frac{\mathbb{P}(G)\mathbb{P}(D \mid G)}{\sum_{G \in \hat{\mathcal{G}}_T} \mathbb{P}(G)\mathbb{P}(D \mid G)} \tag{6}$$

to form an approximate posterior $\hat{\mathbb{P}}(G) \coloneqq w_{G,D}\mathbb{1}(G \in \hat{\mathcal{G}}_T)$. The DAG learning algorithm used for this purpose must be able to handle a mix of observational and interventional data. Two recent methods that have been developed for this purpose are given in Hauser and Bühlmann (2012) and Wang et al. (2017). We summarize constructing an approximate posterior via the DAG bootstrap in Algorithm 1.

Algorithm 1 `DAGBootSample`

Input: N datapoints D_N, number of samples T
Output: T bootstrap DAG samples $\hat{\mathcal{G}}_T$

$\hat{\mathcal{G}}_T \leftarrow \emptyset$
for $s = 1 : T$ **do**
  $\tilde{D}_N \leftarrow N$ datapoints sampled (with replacement) from D_N
  $G_s \leftarrow$ `DAGLearner`(D_N) e.g. (Wang et al., 2017; Hauser and Bühlmann, 2012)
  $\hat{\mathcal{G}}_T \leftarrow \hat{\mathcal{G}}_T \cup G_s$
return $\hat{\mathcal{G}}_T$

Given $\hat{\mathcal{G}}_T$, which can be constructed from Algorithm 1 or sampled from a Markov chain, we next discuss how to compute the expectation over θ. Recall that $U^f_{\text{M.I.}}(\xi; D)$ is given by

$$\begin{aligned}&\mathbb{E}_{G|D}\left[\mathbb{E}_{y|G,\xi}U^f_{\text{M.I.}}(y,\xi;D)\right]\\ &= \mathbb{E}_{G|D}\left[\mathbb{E}_{\theta|G,D}\mathbb{E}_{y|G,\theta,\xi}U^f_{\text{M.I.}}(y,\xi;D)\right].\end{aligned} \tag{7}$$

Instead of carrying out the expensive expectation over $\theta \mid G, D$ in Eq. (7), we use the MLE of θ for each sampled G. This approximation is justified by the *Bernstein-von Mises Theorem*, which implies that

$$\begin{aligned}\mathbb{P}(\theta \mid G, D) &\to N(\hat{\theta}^G_{\text{MLE}}, \frac{1}{n}I(\theta_G)^{-1}),\\ \hat{\theta}^G_{\text{MLE}} &\coloneqq \arg\max_\theta \mathbb{P}(D \mid G, \theta).\end{aligned} \tag{8}$$

Here, n is the number of datapoints in D, and $I(\theta_G)$ is the Fisher information matrix of the parameter θ_G, which is the asymptotic limit of the maximum likelihood estimator $\hat{\theta}^G_{\text{MLE}}$ (Van der Vaart, 2000, Chapter 10). Therefore, the posterior distribution $\theta \mid D, G$ concentrates around $\hat{\theta}^G_{\text{MLE}}$ at the standard $O(1/\sqrt{n})$ statistical rate. Hence, for moderate n (e.g., when a moderate amount of observational data is provided at the start of the experimental design), Eq. (8) implies

$$\begin{aligned}&\mathbb{E}_{G|D}\mathbb{E}_{\theta|G,D}\left[\mathbb{E}_{y|G,\theta,\xi}U^f_{\text{M.I.}}(y,\xi;D)\right]\\ &\approx \mathbb{E}_{G|D}\left[\mathbb{E}_{y|G,\hat{\theta}^G_{\text{MLE}},\xi}U^f_{\text{M.I.}}(y,\xi;D)\right].\end{aligned} \tag{9}$$

In Hauser and Bühlmann (2015, Section 6.1), the authors provide a closed-form expression for $\hat{\theta}^G_{\text{MLE}}$ when $y \mid G, \theta$ is multivariate Gaussian. In this case, $\hat{\theta}^G_{\text{MLE}}$ is a simple function of the sample covariance matrix.

4.2 Approximating Mutual Information

While in the previous subsection we showed how to approximate the expectations in Eq. (9), computing $U^f_{\text{M.I.}}(y, \xi)$ even for a fixed y is intractable since we must sum over all possible DAGs. Recall from Eq. (3) that the mutual information utility function is

$$U^f_{\text{M.I.}}(y, \xi; D) = H(f \mid D) - H(f \mid D, y = y, \xi). \tag{10}$$

Note that $H(f \mid D)$ is a constant and does not matter in the optimization over ξ. More care is required for computing the second term in Eq. (10), since the posterior of G changes as a result of observing y, the realizations of the interventions specified by ξ. We therefore cannot immediately use the samples in $\hat{\mathcal{G}}_T$ to approximate this term. To overcome this problem, we propose to use *weighted importance sampling* and approximate $H(f \mid D, y, \xi)$ by a weighted average of DAGs in $\hat{\mathcal{G}}_T$. We define the importance sample weights for DAG G_i, $1 \le i \le T$, by

$$w_i \coloneqq \frac{\mathbb{P}(D, y \mid G_i, \xi)}{\mathbb{P}(D \mid G_i)}. \tag{11}$$

In general, w_i is not equal to $\mathbb{P}(y \mid G, \xi)$ since D and y are dependent without conditioning on θ. While $\mathbb{P}(D, y \mid G, \xi)$ can be computed in closed-form if the

prior on $\theta \mid G$ belongs to one of the families described in Geiger and Heckerman (1999), the dependence on previous samples in the importance weights makes greedily building up the intervention set ξ expensive. In particular, since Eq. (11) does not factorize, the importance weights must be recomputed with every new additional intervention, which again requires an integration over all parameters. Motivated by the approximation in Section 4, where the parameters of each sampled $G \in \hat{\mathcal{G}}_T$ are not marginalized out, we instead propose using the importance sample weights

$$\begin{aligned} \hat{w}_i &:= \frac{\mathbb{P}(D, y \mid G_i, \hat{\theta}^{G_i}_{\text{MLE}}, \xi)}{\mathbb{P}(D \mid G_i, \hat{\theta}^{G_i}_{\text{MLE}})} \\ &= \mathbb{P}(y \mid G_i, \xi, \hat{\theta}^{G_i}_{\text{MLE}}); \end{aligned} \tag{12}$$

$\hat{w}_i$ has the natural interpretation of re-weighting each DAG by the likelihood of the newly observed data y.

Recall from Eq. (3), that $U^f_{\text{M.I.}}(y, \xi; D)$ is based on weighting each DAG according to its posterior probability $\mathbb{P}(G \mid D) \propto \int_\theta \mathbb{P}(D \mid G, \theta)\mathbb{P}(\theta \mid G)\mathbb{P}(G)$. Using the importance sample weights $\hat{w}_i$ translates into approximating the mutual information against a different posterior distribution in Eq. (3), namely

$$\tilde{\mathbb{P}}(G \mid D) \propto \mathbb{P}(D \mid G, \hat{\theta}^G_{\text{MLE}})\mathbb{P}(G), \tag{13}$$

which is a specific instance of an empirical Bayes approximation. In what follows, we denote the mutual information score based on the posterior in Eq. (13) by $\tilde{U}^f_{\text{M.I.}}(y, \xi; D)$.

4.3 Greedy Optimization

The cardinality constraint $|\xi| = N_b$ makes our optimization problem a difficult integer program. In the following, we show how to overcome this final computational hurdle using a generalized notion of *submodularity* for multisets (Soma and Yoshida, 2016). In particular, we prove that greedily selecting interventions provides a $(1 - \frac{1}{e})$ guarantee on optimization quality.

Algorithm 2 `GreedyDesign`

Input: Utility function U, number of samples N_b, intervention family $\mathcal{I}^*$
Output: Multiset of interventions ξ

$\xi \leftarrow \emptyset$
for $s = 1 : N_b$ **do**
  $I^* \in \arg\max_{I \in \mathcal{I}^*} U(\xi \cup I)$
  $\xi \leftarrow \xi \cup I^*$
return ξ

Theorem 4.1. *Suppose $f(G) = G$ i.e. the goal is to recover the full graph as in Tong and Koller (2001); Cho et al. (2016); Murphy (2001); Ness et al. (2018). Then the difference between the global optimum*

$$v^*_b = \max_{\xi \in \mathbb{Z}^{\mathcal{I}^*} \cap C_{N,b}} \mathbb{E}_{G|D_{b-1}} \mathbb{E}_{y|G, \hat{\theta}^G_{MLE}, \xi} \tilde{U}^f_{M.I.}(y, \xi; D)$$

and $\tilde{v}_b = \texttt{GreedyDesign}(\tilde{U}^f_{M.I.}, N_b, \mathcal{I}^*)$, the output of Algorithm 2 in batch b, satisfies $\tilde{v}_b \geq (1 - \frac{1}{e})v^*_b$, where $C_{N,b}$ is defined as in Eq. (4).

Remark. We conjecture that Theorem 4.1 holds for arbitrary functions f, but we currently only have a proof (see Appendix) for the case when $f(G) = G$.

We conclude this section by summarizing the developed *Active Budgeted Causal Design Strategy* (*ABCD-Strategy*) in Algorithm 3 and then summarizing all the proposed approximations.

Algorithm 3 `ABCD`-Strategy

Input: Target functional f, interventional data collected D_{b-1}, observational data D_{obs}, number of batch samples N_b, intervention family $\mathcal{I}^*$, number of DAGs T, number of datasets M
Output: Multiset of interventions ξ

$\xi \leftarrow \emptyset$
$G_T \leftarrow$ `DAGBootSample`$([D_{obs}, D_{b-1}], T)$
Compute $\hat{U}^f_{\text{M.I.}}$ via Eq. (14)
return `GreedyDesign`$(\hat{U}^f_{\text{M.I.}}, N_b, \mathcal{I}^*)$

In terms of approximations, Eq. (9) implies

$$\begin{aligned} &\mathbb{E}_{G,\theta|D}\left[\mathbb{E}_{y|G,\theta,\xi}\tilde{U}^f_{\text{M.I.}}(y, \xi; D)\right] \\ &\approx \mathbb{E}_{G,|D}\left[\mathbb{E}_{y|G,\hat{\theta}^G_{\text{MLE}},\xi}\tilde{U}_{\text{M.I.}}(y, \xi; D)\right] \\ &\approx \sum_{t=1}^{T}\sum_{m=1}^{M}\tilde{U}^f_{\text{M.I.}}(y_{tm}, \xi; D), \\ &\quad \text{s.t. } y_{tm} \overset{\text{i.i.d}}{\sim} y \mid G_t, \hat{\theta}^G_{\text{MLE}}, \xi \\ &\approx \sum_{t=1}^{T}\sum_{m=1}^{M}\hat{U}^f_{\text{M.I.}}(y_{tm}, \xi; D), \text{ where} \end{aligned} \tag{14}$$

$$\hat{U}^f_{\text{M.I.}}(y_{tm}, \xi; D) := H_1(f \mid D) - H_2(f \mid D),$$
$$\hat{\mathbb{P}}_1(G \mid D) := w_{G,D}\mathbb{1}(G \in \hat{\mathcal{G}}_T),$$
$$\hat{\mathbb{P}}_2(G \mid D, y, \xi) := \frac{w_{G,D}\mathbb{P}(y \mid G, \xi, \hat{\theta}^G_{\text{MLE}})}{\sum_{t=1}^{T} w_{G_t,D}\mathbb{P}(y \mid G_t, \xi, \hat{\theta}^{G_t}_{\text{MLE}})},$$

where M is the number of synthetic datasets generated, H_1 and H_2 are the entropies induced by $\hat{\mathbb{P}}_1$ and $\hat{\mathbb{P}}_2$ respectively, and $w_{G,D}$ is defined in Eq. (6). Note that $\hat{U}^f_{\text{M.I.}}$ is based on the importance sample weights given in Eq. (12).

Proposition 4.2. *The total runtime of Algorithm 2 with input utility function $\hat{U}^f_{M.I.}$ is $O(pT\kappa^3 +$*

$|\mathcal{I}^*|MT^2N_b\kappa p)$, *where* κ *is the maximum indegree of a graph in* $\hat{\mathcal{G}}_T$.

See Appendix A.5 for the proof of Proposition 4.2.

5 Experiments

We begin by considering a simple case to demonstrate the behavior of our ABCD-strategy under easily interpretable conditions. Consider the chain graph on $2m-1$ nodes,

$$1 \to 2 \to \ldots \to m \to \ldots \to p = 2m-1. \tag{15}$$

The corresponding essential graph is completely undirected, and the MEC has $2m-1$ members, one with each node as the source. Assume that sufficient observational data is available to identify the MEC, and we are interested in fully identifying the DAG. Then, our ABCD-strategy selects interventions in order to minimize the expected entropy of the posterior over this MEC. Given a limit of one intervention per batch but infinite samples per batch, Proposition 3.2 implies the expected entropy after intervening at node i or $2m-i$, $1 \le i \le m$, is

$$\frac{1}{2m-1}\Big(\sum_{j<i}\log(i-2) + \sum_{j>i}\log(m-(i+2))\Big),$$

which is minimized by choosing the midpoint $i = m$. Analogously, we see that the updated $\{\emptyset, \{m\}\}$-MEC is of the same form, so in the second batch, the optimal intervention will be halfway through the remaining nodes. This process of bisection is illustrated in Figure 1 and matches the behavior of our algorithm even in the finite-sample regime as described next.

Figure 2 illustrates the performance of our ABCD-strategy on fifty 11-node chain graphs with random edge weights sampled from $[-1,-.25] \cup [.25,1]$. For comparison, we consider a random intervention strategy that uniformly distributes the samples in each batch to k interventions picked uniformly at random,

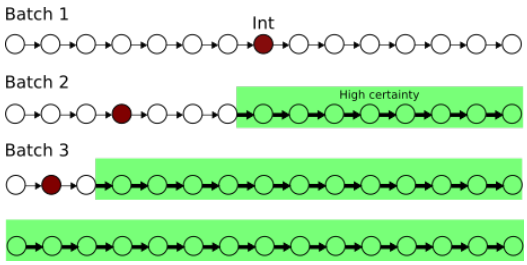


Figure 1: Illustration of active learning on a chain graph, beginning with a known MEC on a simulated dataset with $p = 15$ nodes. The brown circles indicate the interventions selected in each batch.

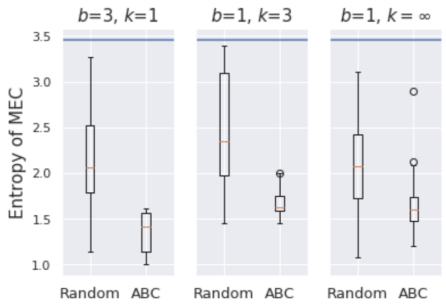


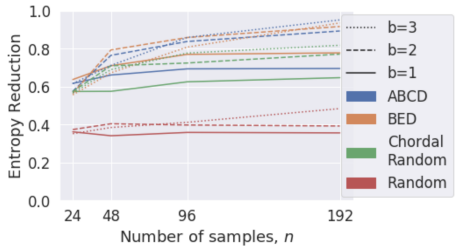
Figure 2: Box plots for 50 runs of the random strategy versus our ABCD-strategy on the graph in Figure 5 with $p = 11$ and $n = 30$ samples. The horizontal line indicates the entropy of the prior distribution, i.e. uniform over the MEC. Note that $k = \infty$ corresponds to the case with no constraints on the number of unique interventions.

where k is the maximum number of unique interventions allowed per batch. Whereas the median-performing random strategy barely reduces the entropy, the ABCD-strategy reduces the entropy significantly in all runs. When all k interventions are picked for the same batch, so that ABCD receives no feedback, the median-performing run of active learning still reduces the entropy as much as the best-performing runs of the random strategy.

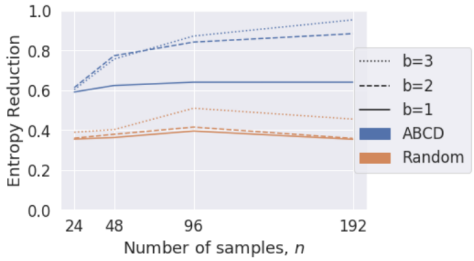
Having demonstrated the behavior of ABCD for a simple case, we now analyze the performance of our method on more general DAGs. The skeleton of each graph is sampled from an Erdös-Rényi model with density $\rho = 0.25$. The edges of these graphs are directed by sampling a permutation of the nodes uniformly at random and orienting the edges accordingly. To avoid long runtimes when enumerating the MEC, we disposed of graphs with more than 100 members in their MEC.[1] When the MEC is known, we may define a variant of the random strategy, *Chordal-Random*, which only intervenes on nodes that are in chordal components of the essential graph, i.e., nodes adjacent to at least one undirected edge. Since the Meek rules can only propagate by intervening within chordal components, Chordal-Random is a more fair baseline strategy for comparison than simple random sampling.

Figure 3a demonstrates the improvement in selecting interventions using the ABCD-strategy as compared to Chordal-Random when the number of unique interventions per batch is bounded by one. The entropy reduction for an interventional data set D_ξ is defined as

[1] From a sample of 10,000 graphs, only 54 had MEC size greater than 100. Based on the results by Gillispie and Perlman (2001), we expect the MECs to be typically small.



(a) Single MEC



(b) Multiple MECs.

Figure 3: Performance of intervention strategies for batch sizes b as a function of the total number of samples, computed from 50 Erdös-Rényi DAGs with density $\rho = 0.25$.

$\frac{H(G)-H(G|D_\xi)}{H(G)}$, and it is used as a metric so that MECs of different sizes are comparable. Since the number of total possible unique interventions is kB, an increase in the number of batches also increases the variability of the interventions, reflected in the increase of entropy reduction with batch size. Already with only 192 samples and 3 total batches, our ABCD-strategy is able to learn most graphs with complete certainty. The comparable performance of the Budgeted Experiment Design (BED) strategy (Ghassami et al., 2018) suggests that for the given experimental setup, the interventions that orient the most edges correspond well to those that most reduce entropy as we discussed in Proposition 3.2. Figure 3b shows that the performance of the ABCD-strategy remains strong even when the MEC of the graph is not known. Specifically, up to 3 additional MECs were generated by randomly flipping non-covered edges that did not create cycles, and again only graphs for which the union of these MECs had cardinality less than 100 were kept. Note that we are not able to compare with BED since BED requires that the MEC is known.

DREAM4 Synthetic Dataset. Finally, we applied our experimental design strategy to gene expression data from the DREAM4 10-node in-silico network reconstruction challenge (Schaffter et al., 2011). These data are generated from stochastic differential equations and simulate microarray data of gene regulatory networks. We constructed an observational dataset

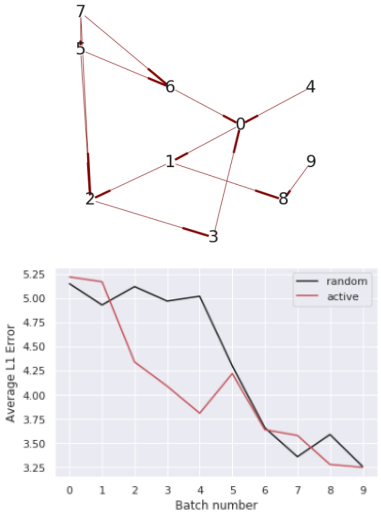


Figure 4: Top: DREAM4 ground truth 10-node network. Bottom: Performance of intervention strategies on predicting the descendants of gene 0.

from the wild-type, multifactorial perturbation, and time 0 time-series samples (16 samples in total), and similarly, interventional datasets from the knockdown and knockout samples (2 samples each).

Previous work on experimental design applied to biological datasets (Cho et al., 2016) has focused on learning the entire network. In practice, practitioners may be specifically interested in performing experiments to elucidate a functional of the network, such as the pathway or local network surrounding a gene of interest. To emulate this setting, we applied our ABCD-strategy towards learning the *downstream genes* of select genes from the true network (Figure 4, top). Despite high variations in learning due to the small size of the dataset, we observed an improvement over the random strategy for several central genes (Fig. 4, bottom; Fig. 6). These results illustrate the promise of applying targeted experimental design for applications to genomics.

6 Concluding Remarks

We proposed *Active Budgeted Causal Design Strategy* (*ABCD-Strategy*), an experimental method based on optimal Bayesian experimental design with provable guarantees on approximation quality. Empirically, we demonstrated that ABCD yields considerable boosts over random sampling for both targeted and full causal structure discovery. Such experimental design strategies are particularly relevant for applications to genomics, where the number of possible experiments is huge due to the possibility of intervening on combinations of genes.

Acknowledgements

R. Agrawal was partially supported by IBM. K.D. Yang was supported by an NSF graduate fellowship and ONR (N00014-18-1-2765). C. Uhler was partially supported by NSF (DMS-1651995), ONR (N00014-17-1-2147 and N00014-18-1-2765), IBM, and a Sloan Fellowship.

References

R. Agrawal, T. Broderick, and C. Uhler. Minimal I-MAP MCMC for scalable structure discovery in causal DAG models. In *International Conference on Machine Learning*, 2018.

S. A. Andersson, D. Madigan, and M. D. Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *Annals of Statistics*, 25(2):505–541, 1997.

K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10:273–304, 1995.

H. Cho, B. Berger, and J. Peng. Reconstructing causal biological networks through active learning. *PLoS ONE*, 2016.

A. Dixit, O. Parnas, B. Li, J. Chen, C. Fulco, L. Jerby-Arnon, N. Marjanovic, D. Dionne, T. Burks, R. Raychowdhury, B. Adamson, T. Norman, E. Lander, J. Weissman, N. Friedman, and A. Regev. Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, pages 1853–1866, 2016.

B. Ellis and W. H. Wong. Learning causal Bayesian network structures from experimental data. *Journal of the American Statistical Association*, 103:778–789, 2008.

N. Friedman and D. Koller. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50:95–125, 2003.

N. Friedman, M. Goldszmidt, and A. J. Wyner. Data analysis with Bayesian networks: A bootstrap approach. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 1999.

N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3-4):601–620, 2000.

D. Geiger and D. Heckerman. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 1999.

A. Ghassami, S. Salehkaleybar, N. Kiyavash, and E. Bareinboim. Budgeted experiment design for causal structure learning. In *International Conference on Machine Learning*, 2018.

S. B. Gillispie and M. D. Perlman. Enumerating Markov equivalence classes of acyclic digraph models. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, 2001.

M. Grzegorczyk and D. Husmeier. Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Machine Learning*, 71:265–305, 2008.

A. Hauser and P. Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13(1):2409–2464, 2012.

A. Hauser and P. Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55:926–939, 2014.

A. Hauser and P. Bühlmann. Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society Series B*, 77(1):291–318, 2015.

D. Heckerman, C. Meek, and G. Cooper. A Bayesian approach to causal discovery. Technical report, Microsoft Research, 1997.

J. Kuipers and G. Moffa. Partition MCMC for inference on acyclic digraphs. *Journal of the American Statistical Association*, 112:282–299, 2017.

J. Kuipers, G. Moffa, and D. Heckerman. Addendum on the scoring of Gaussian directed acyclic graphical models. *The Annals of Statistics*, 42:1689–1691, 2014.

S. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

D. Madigan and J. York. Bayesian graphical models for discrete data. *International Statistical Review*, 63:215–232, 1995.

K. Murphy. Active learning of causal Bayes net structure. Technical report, 2001.

R. O. Ness, K. Sachs, P. Mallick, and O. Vitek. A Bayesian active learning experimental design for inferring signaling networks. *Journal of Computational Biology*, 25(7):709–725, 2018.

T. Niinimaki, P. Parviainen, and M. Koivisto. Structure discovery in Bayesian networks by sampling partial orders. *Journal of Machine Learning Research*, 17:2002–2048, 2016.

J. Pearl. Causality: Models, reasoning, and inference. *Econometric Theory*, 19(675-685):46, 2003.

J. M. Robins, M. A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology, 2000.

T. Schaffter, D. Marbach, and D. Floreano. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27:2263–2270, 2011.

T. Soma and Y. Yoshida. Maximizing monotone submodular functions over the integer lattice. In *International Conference on Integer Programming and Combinatorial Optimization*, pages 325–336. Springer, 2016.

T. Soma, N. Kakimura, K. Inaba, and K. Kawarabayashi. Optimal budget allocation: Theoretical guarantee and efficient algorithm. In *International Conference on International Conference on Machine Learning*, 2014.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

S. Tong and D. Koller. Active learning for structure in Bayesian networks. In *International Joint Conference on Artificial Intelligence*, 2001.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

T. S. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Uncertainty in Artificial Intelligence*, volume 6, page 255, 1991.

T. S. Verma and J. Pearl. An algorithm for deciding if a set of observed independencies has a causal explanation. In *Uncertainty in Artificial Intelligence*, 1992.

Y. Wang, L. Solus, K. Yang, and C. Uhler. Permutation-based causal inference algorithms with interventions. In *Advances in Neural Information Processing Systems*, pages 5824–5833, 2017.

K. D. Yang, A. Katcoff, and C. Uhler. Characterizing and learning equivalence classes of causal DAGs under interventions. In *International Conference on Machine Learning*, 2018.