Introduction to Experimental Design for Causal Discovery

Chandler Squires

Agenda

Settings for experimental design in causal structure learning and preliminaries
9:30-10:30

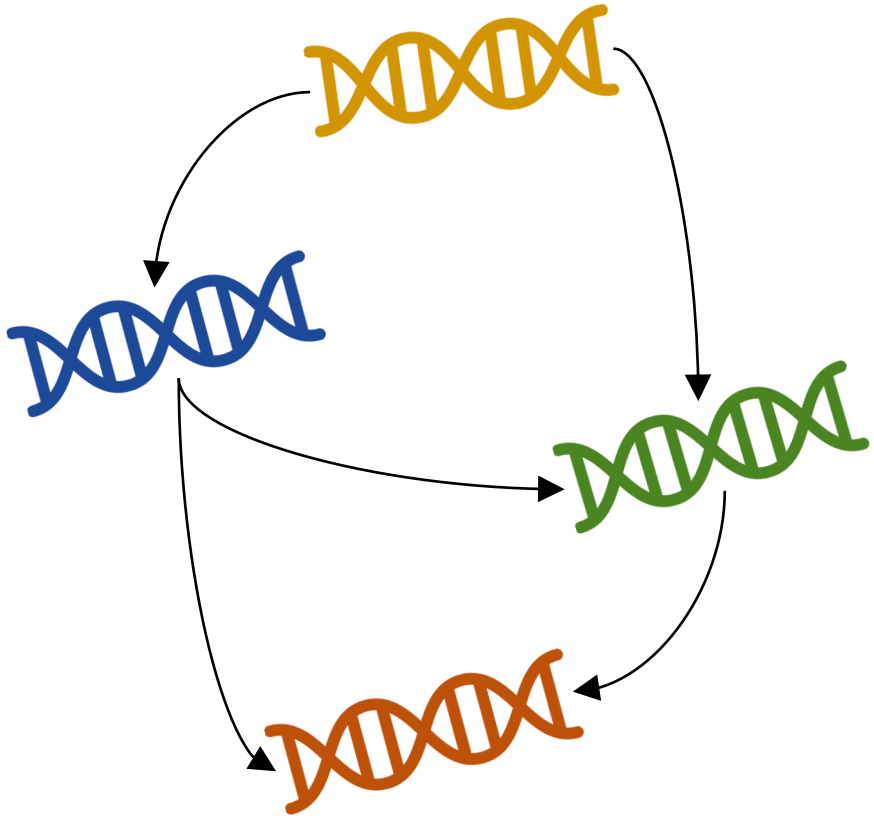
Non-adaptive experimental design strategies
11-12

Adaptive experimental design strategies
2-3

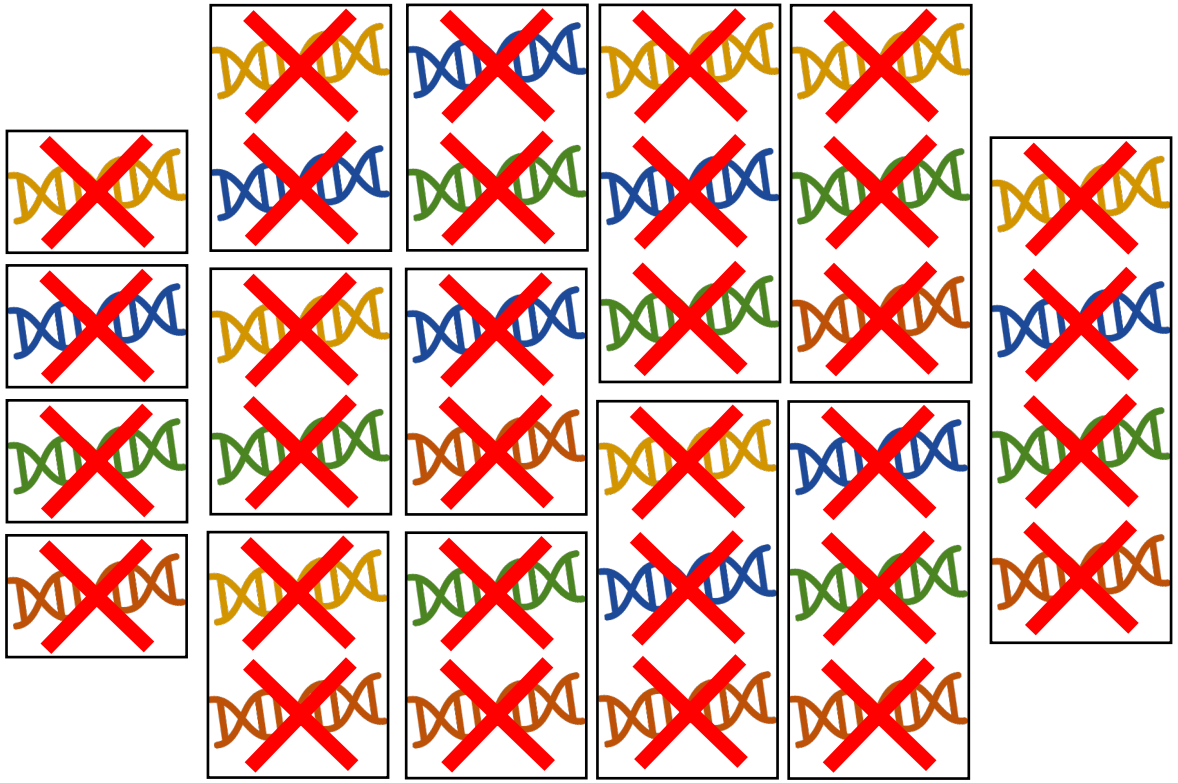
Targeted experimental design and other open challenges
3:30-4:30

Part I: Settings for experimental design in causal discovery and preliminaries

Motivating Example: Genomics

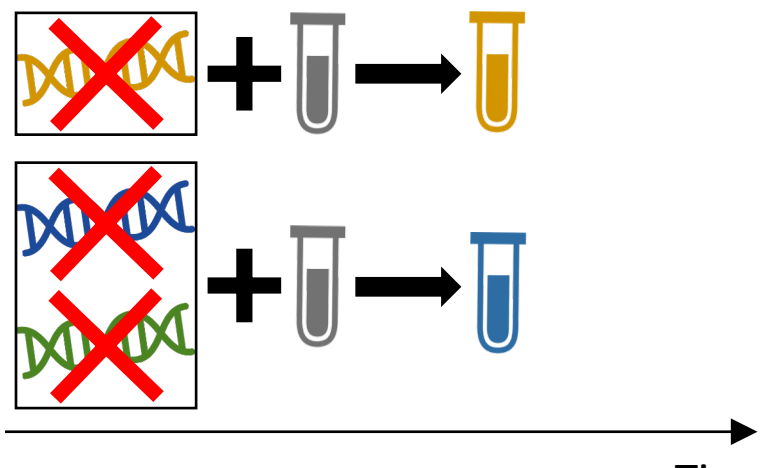


Knockout experiments:



How do we pick experiments to learn the underlying causal graph?

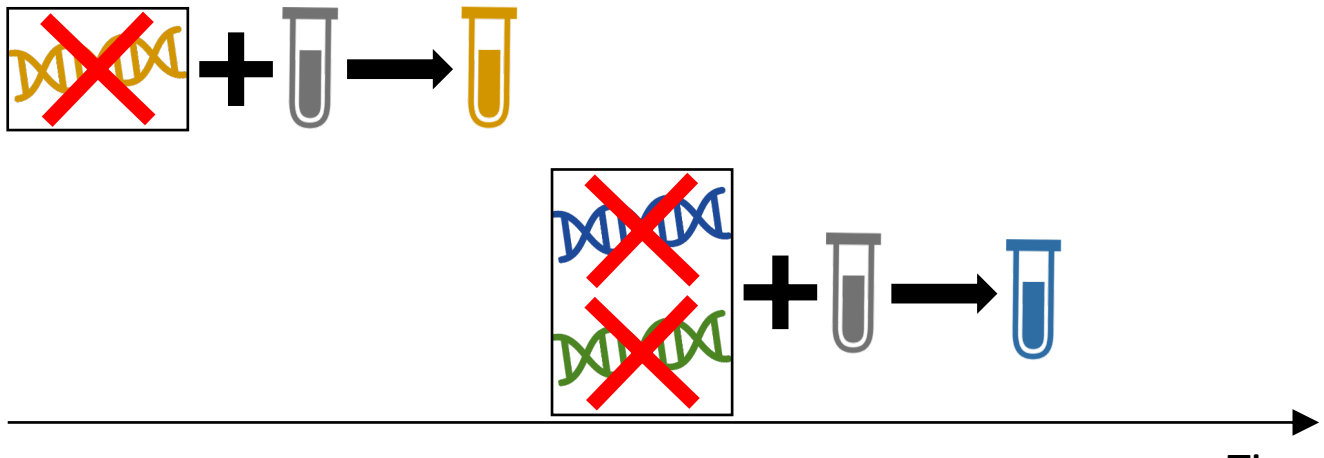
What are our experimental limitations?



Time

Non-adaptive

Single-round
Passive
Parallel
Batch
Fixed

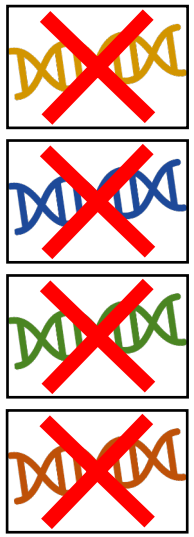


Time

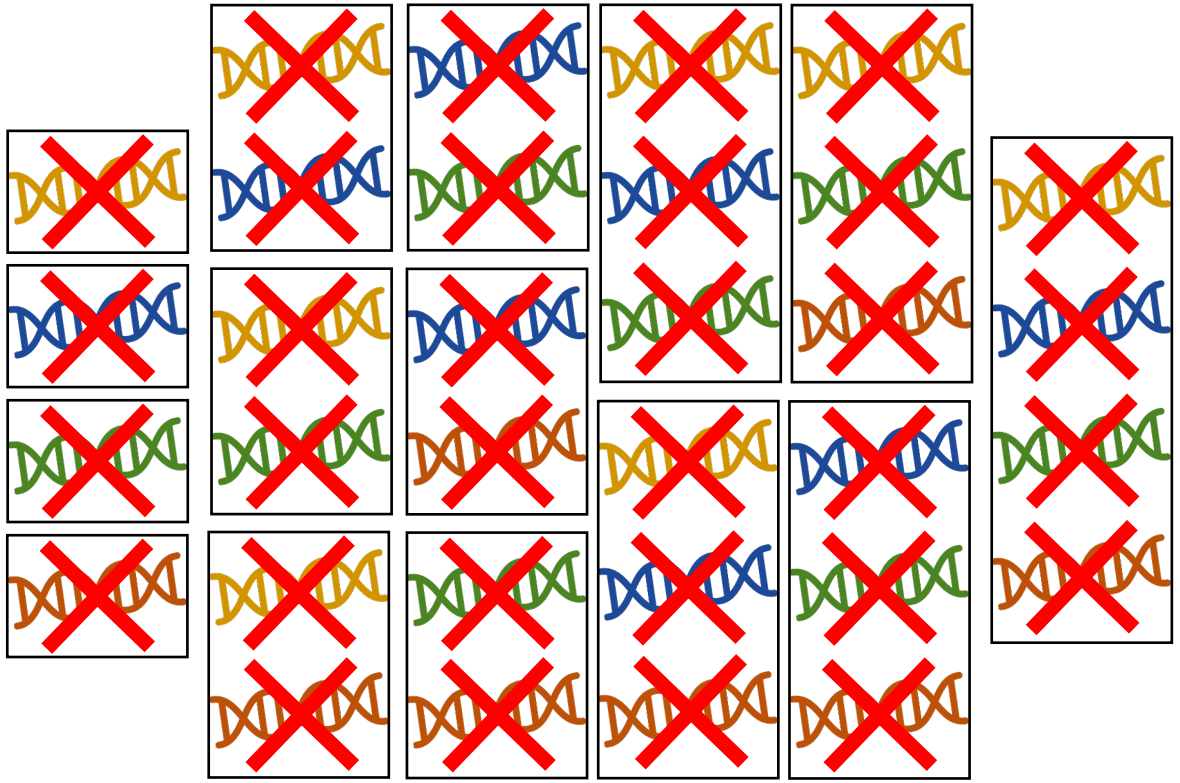
Adaptive

Multi-round
Active
Sequential

What are our experimental limitations?

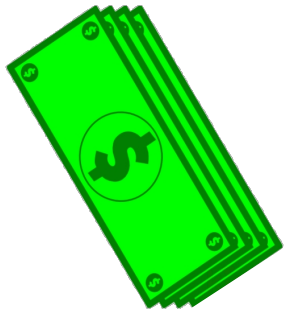


Bounded



Unbounded

What are our experimental limitations?



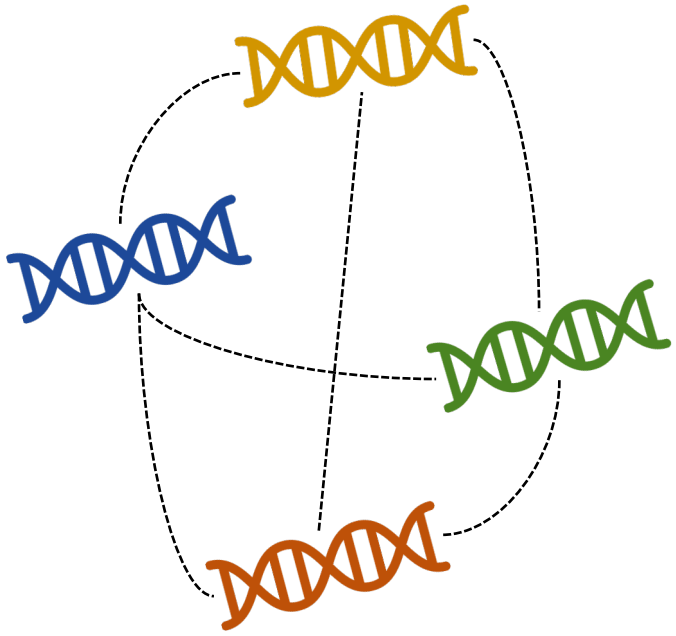
max $info(interventions)$
s.t. $cost(interventions) \leq budget$

Fixed budget

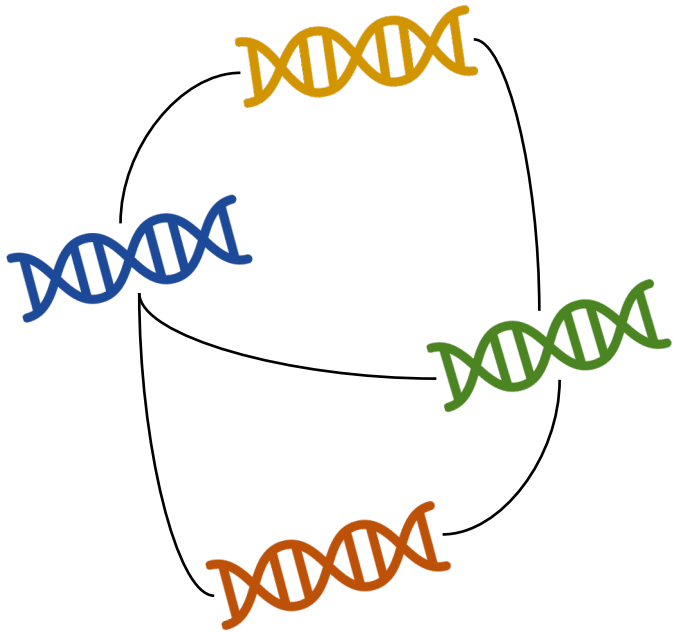
min $cost(interventions)$
s.t. $interventions\ identify\ G^*$

Minimum cost identification

What are our experimental limitations?

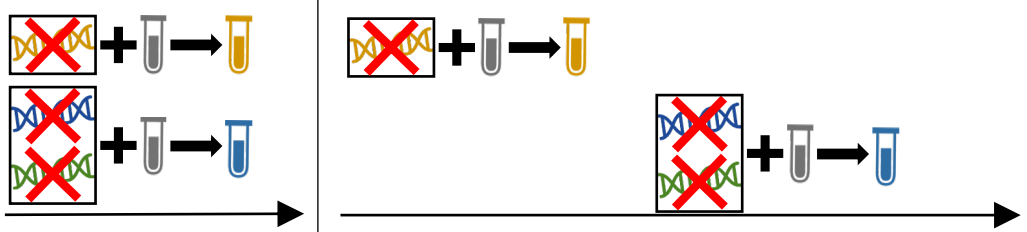


No observational data



Observational data

What are our experimental limitations?



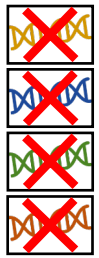
Non-adaptive | Adaptive

$$\max \quad info(interventions)$$
$$\text{s.t.} \quad cost(interventions) \leq budget$$

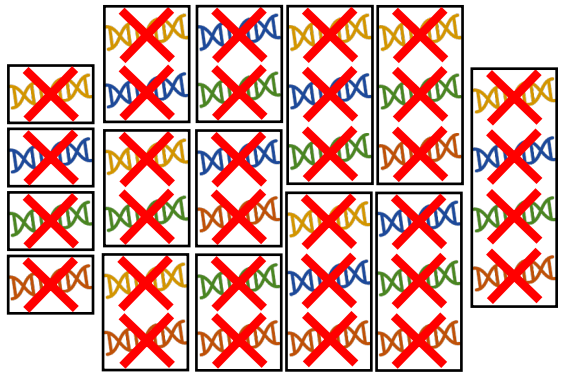
Fixed budget

$$\min \quad cost(interventions)$$
$$\text{s.t.} \quad interventions\ identify\ G^*$$

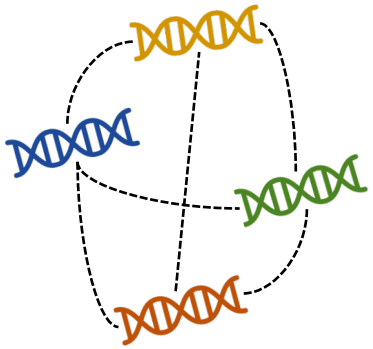
Min-cost identification



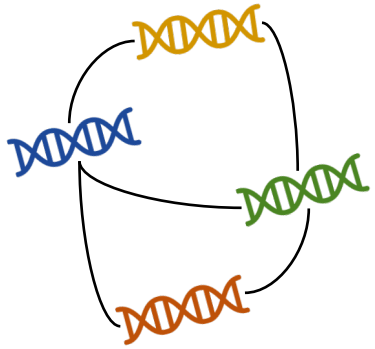
Bounded



Unbounded



No observational data

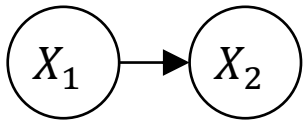


Observational data

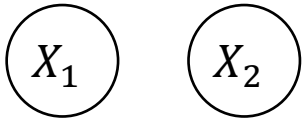
Additional assumptions for parts I, II, and III

- **Noiseless setting:** we obtain an infinite amount of data from each intervention, so that the only uncertainty is due to unidentifiability and not statistical noise.
  - i.e., if the distribution associated with an intervention I_k is f^k, then we assume access to *conditional invariance* oracle that tells us whether $f^k(X_i \mid X_C) = f^{k'}(X_i \mid X_C)$ for any node i and set C
  - In practice, we can imagine setting a threshold ϵ and obtaining enough interventional samples to distinguish any two distributions that are further than ϵ under some measure (e.g., total variation distance)
- **Causally sufficient setting:** there are no unobserved confounders.

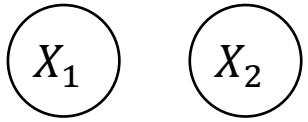
Types of Interventions



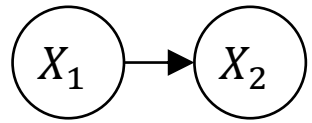
$X_1 \leftarrow f_1(\varepsilon_1)$
$X_2 \leftarrow f_2(X_1, \varepsilon_2)$



$X_1 \leftarrow f_1(\varepsilon_1)$
$X_2 \leftarrow x_2$



$X_1 \leftarrow f_1(\varepsilon_1)$
$X_2 \leftarrow f_2'(\varepsilon_2')$

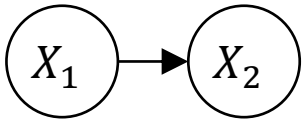


$X_1 \leftarrow f_1(\varepsilon_1)$
$X_2 \leftarrow f_2'(X_1, \varepsilon_2')$

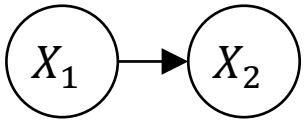
Do-interventions $\subset$ Perfect interventions $\subset$ Soft interventions

aka: mechanism changes

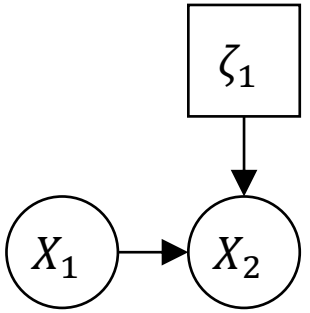
Interventional Augmented Graphs



$$X_1 \leftarrow f_1(\varepsilon_1)$$
$$X_2 \leftarrow f_2(X_1, \varepsilon_2)$$

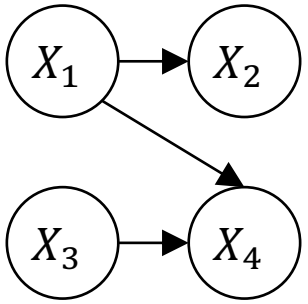


$$X_1 \leftarrow f_1(\varepsilon_1)$$
$$X_2 \leftarrow f_2'(X_1, \varepsilon_2')$$



$$X_1 \leftarrow f_1(\varepsilon_1)$$
$$X_2 \leftarrow (1 - \zeta_1) \cdot f_2(X_2, \varepsilon_2) + \zeta_1 \cdot f_2'(X_1, \varepsilon_2')$$

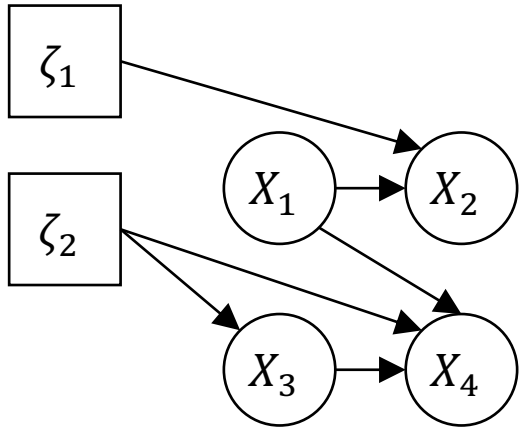
Interventional Augmented Graphs



$$I_1 = \{X_2\}$$

$$I_2 = \{X_3, X_4\}$$

$$\mathcal{I} = \{\emptyset, I_1, I_2\}$$



$\mathcal{I}$-DAG

Given a DAG G and a set of interventions $\mathcal{I} = \{\mathrm{I}_1, \ldots, I_K\}$, the **interventional DAG** ($\mathcal{I}$-DAG), denoted $G^{\mathcal{I}}$, is obtained by introducing **intervention variables** $\zeta_1, \ldots, \zeta_K$ and edges $\zeta_k \rightarrow X_i$ for $X_i \in I_k$

Joint Causal Inference from Multiple Contexts (Mooij et al., 2020)

Characterizing and Learning Equivalence Classes of Causal DAGs under Interventions (Yang et al., 2018)

Interventional Markov Equivalence

An indexed set of distribution $\left(f^k\right)_{k=1}^K$ is $\boldsymbol{\mathcal{I}}$**-Markov** w.r.t. G if:

- each f^k is Markov w.r.t. G
- $f^k(X_i \mid X_{pa(i)}) = f^{k'}(X_i \mid X_{pa(i)})$ for all k, k' such that $X_i \notin I_k \cup I_{k'}$.

We call two DAGs G_1 and G_2 $\boldsymbol{\mathcal{I}}$**-Markov equivalent** if every $(f_k)_{k=1}^K$ which is $\mathcal{I}$-Markov w.r.t. G_1 is $\mathcal{I}$-Markov w.r.t. G_2 and vice versa.

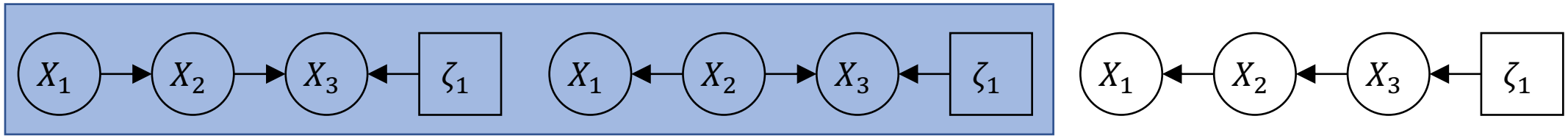
The $\mathcal{I}$-Markov equivalence class of G is denoted $[G]_{\mathcal{I}}$. The observational Markov equivalence class (i.e., $\mathcal{I} = \{\emptyset\}$) is denoted $[G]$.

Characterization of Interventional Markov Equivalence

Theorem (Hauser and Bühlmann, 2012, Yang et al., 2018)

Two DAGs G_1 and G_2 are $\mathcal{I}$-Markov equivalent iff. their $\mathcal{I}$-DAGs $G_1^{\mathcal{I}}$ and $G_2^{\mathcal{I}}$ have the same skeletons and v-structures.

Equivalently: Two DAGs are G_1 and G_2 $\mathcal{I}$-Markov equivalent iff. they are Markov equivalent and when $X_i \to X_j$ in G_1 and $|\{X_i, X_j\} \cap I_k| = 1$ for some $I_k \in \mathcal{I}$, then $X_i \to X_j$ in G_2.



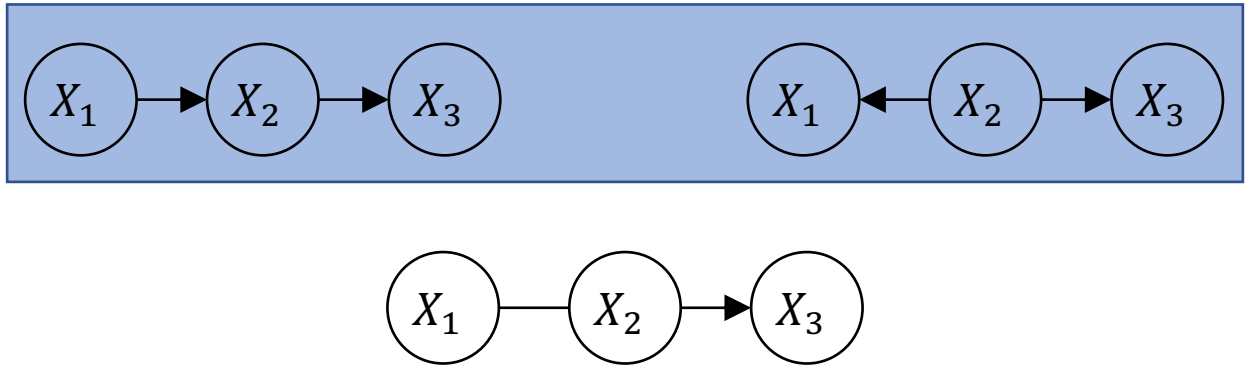
Characterization and Greedy Learning of Interventional Markov Equivalence Classes of Directed Acyclic Graphs (Hauser and Bühlmann, 2012)

Characterizing and Learning Equivalence Classes of Causal DAGs under Interventions (Yang et al., 2018)

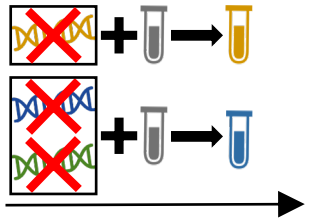
$\mathcal{I}$-essential graphs

Define the $\mathcal{I}$-essential graphs graph of G^*, denoted $\mathcal{E}_{\mathcal{I}}(G^*)$, as the mixed graph (directed and undirected edges) with:

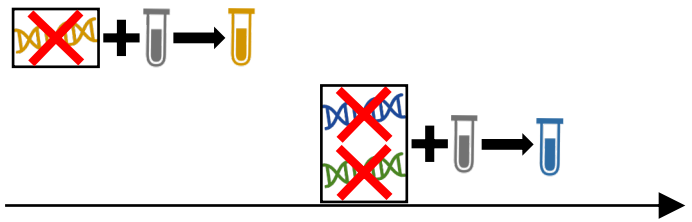
- The same adjacencies as G^*
- $X_i \to X_j$ in $\mathcal{E}_{\mathcal{I}}(G^*)$ if $X_i \to X_j$ for all $G \in [G^*]_{\mathcal{I}}$
- $X_i - X_j$ otherwise



Part II: Non-adaptive experimental design strategies



Non-adaptive



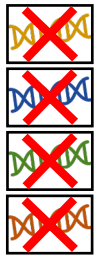
Adaptive

max $info(interventions)$
s.t. $cost(interventions) \leq budget$

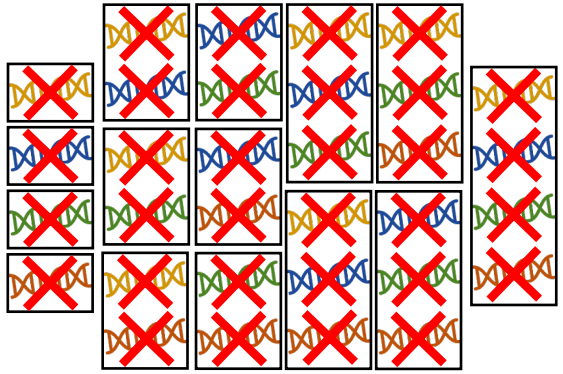
Fixed budget

min $cost(interventions)$
s.t. $interventions\ identify\ G^*$

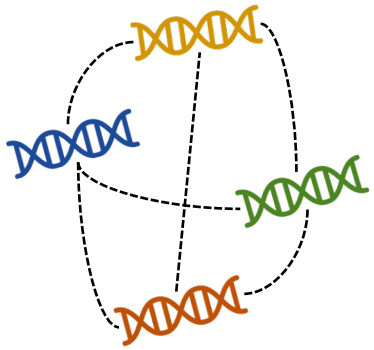
Min-cost identification



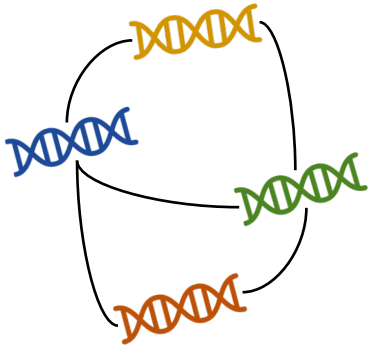
Bounded



Unbounded



No observational data



Observational data

Structural information from interventions

- If $\{X_i, X_j\} \cap I_k = \{X_i\}$, we call I_k an X_i**-orientation test** for (X_i, X_j)
- If $\{X_i, X_j\} \cap I_k = \emptyset$, we call I_k an **adjacency test** for (X_i, X_j)

Lemma (Eberhardt, 2008)

To determine the presence/absence and orientation of an edge between X_i and X_j, either of the following are sufficient and in the worst case necessary:

1. A structural orientation test and a structural adjacency test.
2. An X_i-orientation test and a X_j-orientation test.

Single-node interventions

***Q:** How many single-node interventions are sufficient and worst-case necessary for finding any G^* on $p \geq 3$ nodes?*

$$\begin{aligned} \min \quad & |\mathcal{I}| \\ \text{s.t.} \quad & \forall\, G^*,\ \mathcal{I}\ identifies\ G^* \\ & \forall\, I \in \mathcal{I}, |\mathrm{I}| \leq 1 \end{aligned}$$

A*[1]*: p-1

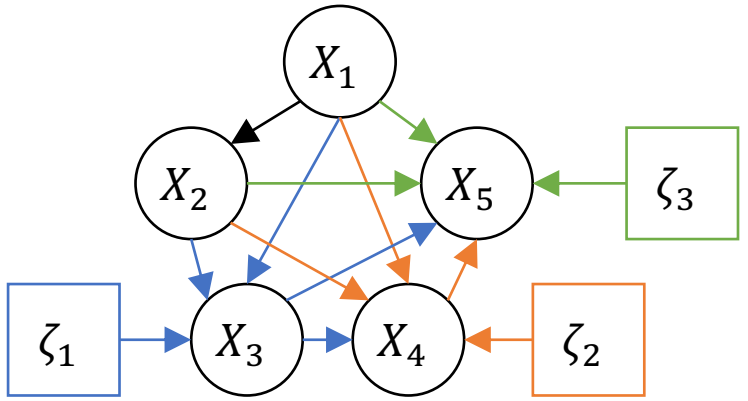
Sufficiency

All pairs of variables are subject to adjacency tests.

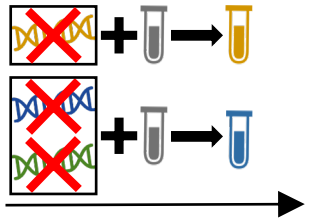
All pairs have at least one member subject to an orientation test.

Worst-case necessity

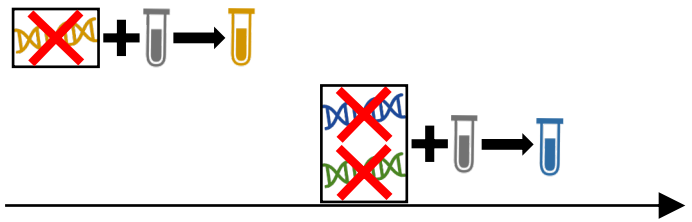
If we only intervene on $p - 2$ variables, and the two we don't intervene on happen to be the most upstream.



[1] Causation and Intervention (Eberhardt, 2008)



Non-adaptive



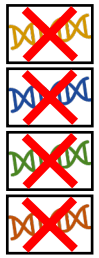
Adaptive

max $info(interventions)$
s.t. $cost(interventions) \leq budget$

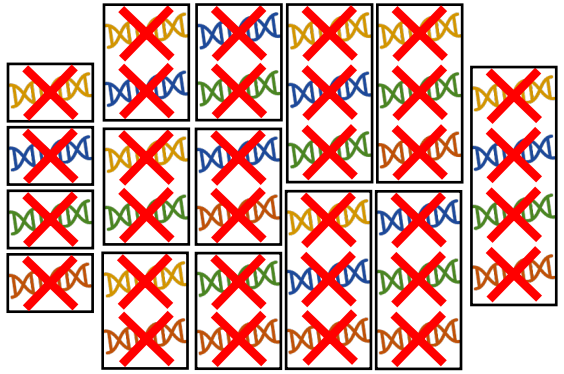
Fixed budget

min $cost(interventions)$
s.t. $interventions\ identify\ G^*$

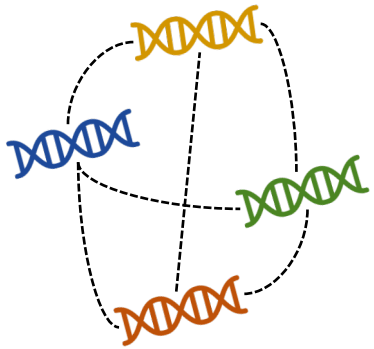
Min-cost identification



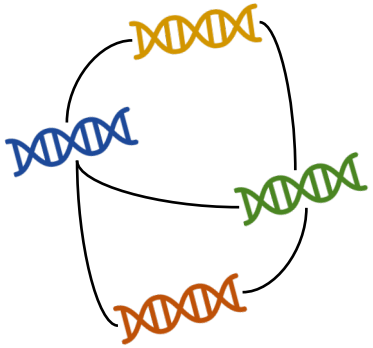
Bounded



Unbounded



No observational data



Observational data

Unbounded interventions

Q: *How many (unrestricted) interventions are sufficient and worst-case necessary for finding any G^* on p nodes?*

$$\begin{aligned} \min \quad & |\mathcal{I}| \\ \text{s.t.} \quad & \forall\, G^*,\ \mathcal{I}\ identifies\ G^* \end{aligned}$$

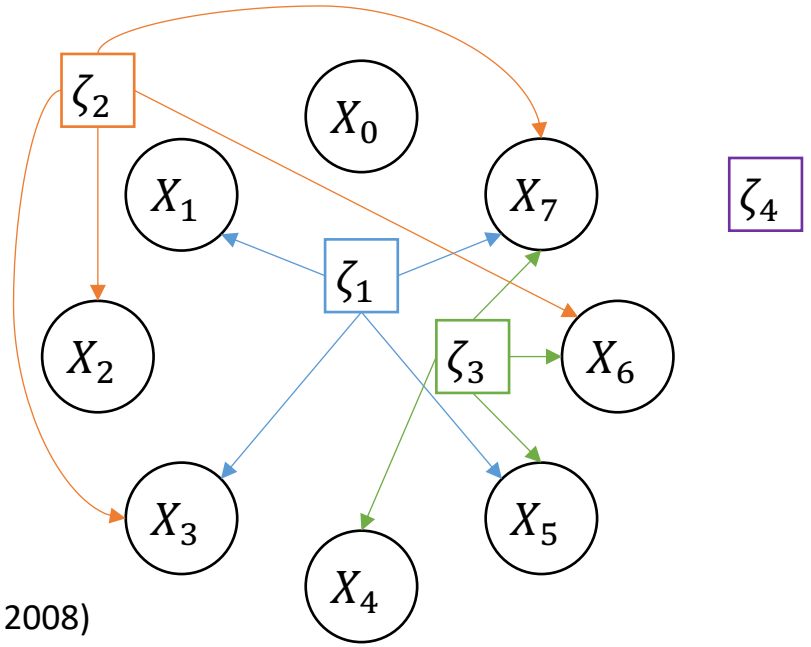
A*[1]*: $\lfloor \log_2 p \rfloor + 1$

Sufficiency

Let $b(k)$ denote the binary representation of an integer k, and $b(k)_j$ the j-th rightmost digit in this representation.

$b(6) = 110, b(6)_1 = 0, b(6)_2 = 1$, etc.

Label the vertices $0, \dots, p-1$ and let $I_j = \{k \mid b(k)_j = 1\}$, for $j = 1, \dots, \lfloor \log_2 p \rfloor + 1$.



[1] Causation and Intervention (Eberhardt, 2008)

Unbounded interventions

Q: *How many (unrestricted) interventions are sufficient and worst-case necessary for finding any G^* on p nodes?*

$$\begin{aligned} \min \quad & |\mathcal{I}| \\ \text{s.t.} \quad & \forall\, G^*,\ \mathcal{I}\ identifies\ G^* \end{aligned}$$

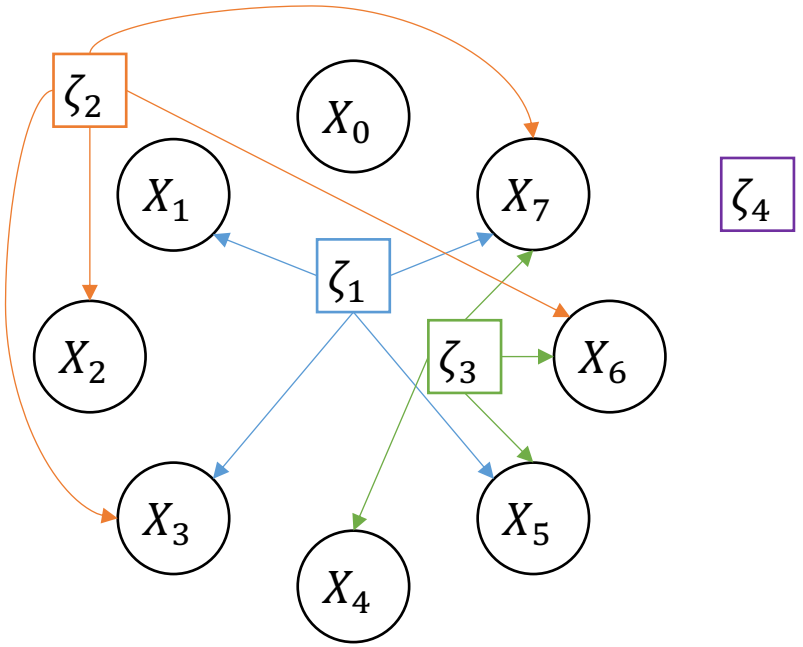
A: $\lfloor \log_2 p \rfloor + 1$

Sufficiency

Each pair is subjected to an orientation test, since any two numbers differ in at least one position in their binary expansions.

If p is a power of 2, then there's an empty intervention which gives an adjacency test for all pairs.

Otherwise, we can use the fact that every binary expansion has at least one zero to show that every pair is subjected to either the opposing orientation test, or an adjacency test.



Unbounded interventions

Q: *How many (unrestricted) interventions are sufficient and worst-case necessary for finding any G^* on p nodes?*

$$\begin{array}{ll} \min & |\mathcal{I}| \\ \text{s.t.} & \forall\, G^*,\ \mathcal{I}\ identifies\ G^* \end{array}$$

A: $\lfloor \log_2 p \rfloor + 1$

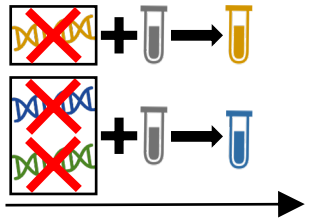
Worst-case necessity

(induction) Assume the result holds for all $q < p$. Base cases on 2, 3, and 4 nodes can be done by hand.

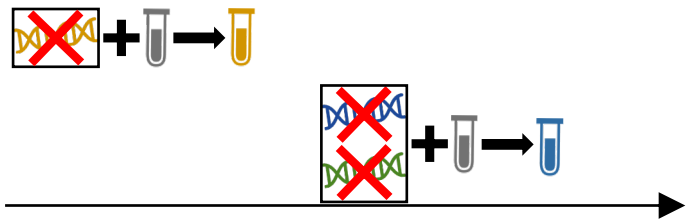
Consider an intervention I_1 on any number $\mathrm{K} \leq p$ nodes.

In the worst case, all $p - K$ non-intervened nodes are upstream of these K nodes and form a complete graph, and there is a complete graph over the K intervened nodes as well.

By the induction hypothesis, one of the remaining graphs requires at least $\lfloor \log_2 \frac{p}{2} \rfloor + 1 = \lfloor \log_2 p \rfloor$ interventions.



Non-adaptive



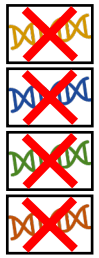
Adaptive

max $info(interventions)$
s.t. $cost(interventions) \leq budget$

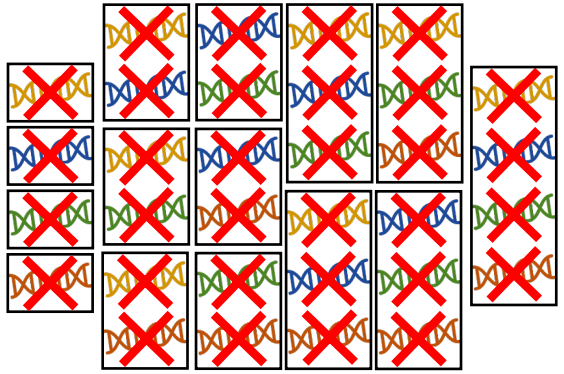
Fixed budget

min $cost(interventions)$
s.t. $interventions\ identify\ G^*$

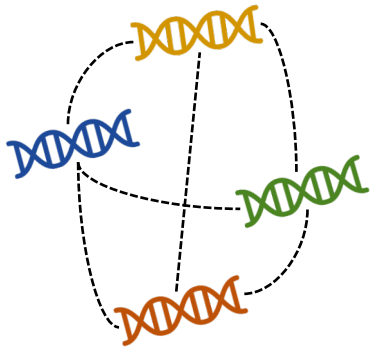
Min-cost identification



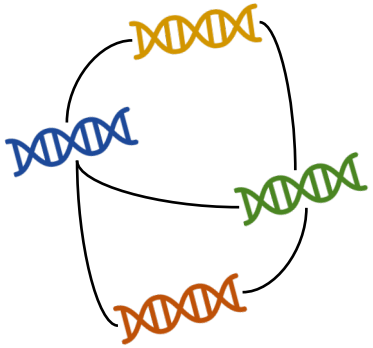
Bounded



Unbounded



No observational data



Observational data

Interventional Essential Graphs

The interventional essential graph is a *chain graph* with *chordal* chain components.

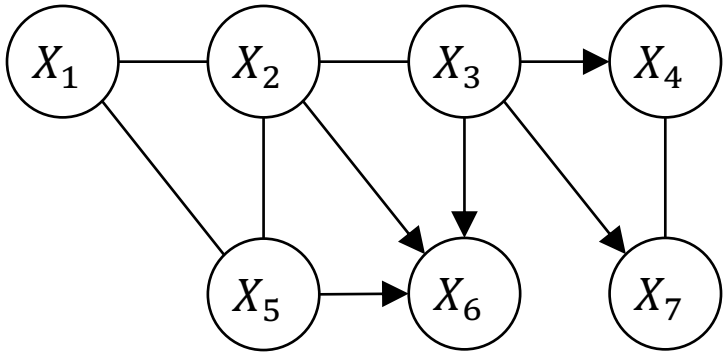
Furthermore, the orientations in each chain component are logically independent of one another.

Characterization and Greedy Learning of Interventional Markov Equivalence Classes of Directed Acyclic Graphs, Propositions 15 and 16 (Hauser and Bühlmann, 2012)

Chain graphs

A **partially directed cycle** in a mixed graph (i.e., a graph with both directed and undirected edges) is a sequence of distinct vertices $v_0, v_1, \ldots, v_n$ such that:

- for each v_i, either $v_{i-1} - v_i$ or $v_{i-1} \to v_i$, where $v_{-1} = v_n$
- at least one edge in the path is directed



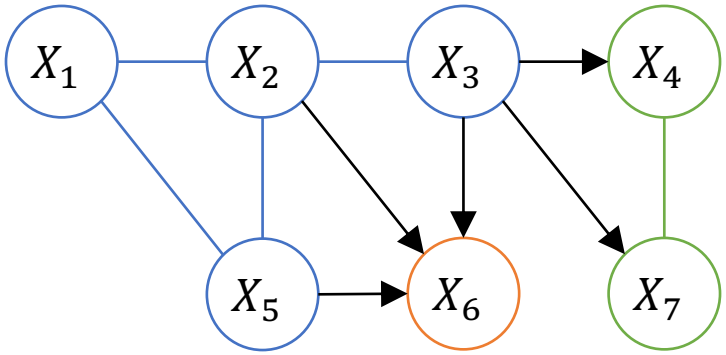
A **chain graph** is a mixed graph with no partially directed cycles.

Chain graphs

A **partially directed cycle** in a mixed graph (i.e., a graph with both directed and undirected edges) is a sequence of distinct vertices $v_0, v_1, \ldots, v_n$ such that:

- for each v_i, either $v_{i-1} - v_i$ or $v_{i-1} \to v_i$, where $v_{-1} = v_n$
- at least one edge in the path is directed

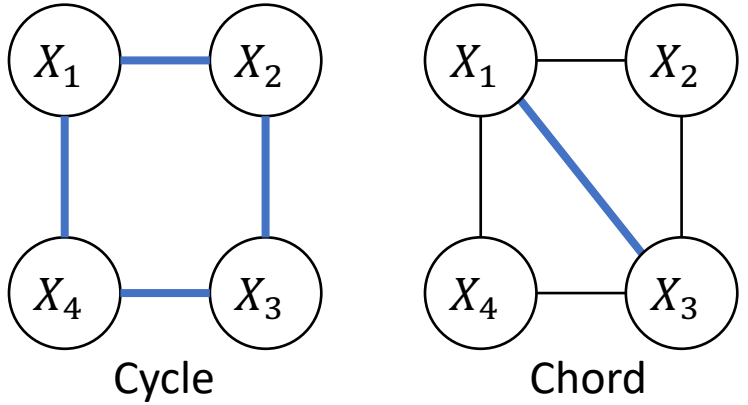
A **chain graph** is a mixed graph with no partially directed cycles.



The **chain components** of a chain graph are the connected components of the subgraph containing only the undirected edges.

Chordal graphs

Given an undirected cycle, a **chord** is an edge that is not in the cycle but connects two nodes in the cycle.



A **chordal graph** is an undirected graph where all cycles of length ≥ 4 have a chord.

The clique number of a graph

A **clique** in a graph is a set of nodes which are all adjacent.

The **clique number** of a graph G is the size of the largest clique, denoted $\omega(G)$.

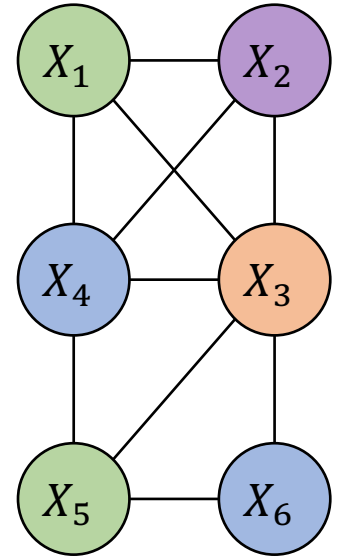
Note, this can be computed in linear time for a chordal graph G.

Colorings of a graph

A **coloring** assigns a color to each node such that adjacent nodes do not have the same color.

The **chromatic number** of a graph is the smallest number of colors sufficient to form a coloring.

In a chordal graph G, the chromatic number is equal to the clique number $\omega(G)$, and an optimal coloring can be found in linear time.



Unbounded interventions for a fixed MEC

Q: *How many (unrestricted) interventions are sufficient and worst-case necessary for finding any* $G^* \in [G^*]$ *on* p *nodes?*

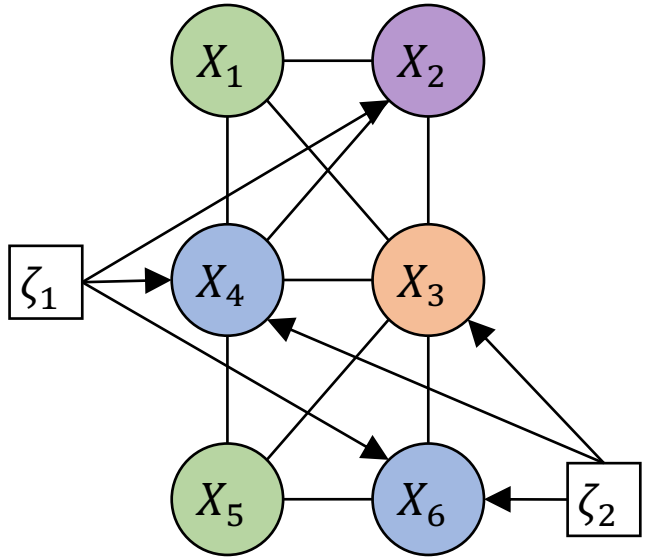
$$\begin{aligned} &\min && |\mathcal{I}| \\ &\text{s.t.} && \forall\, G^* \in [G^*],\ \mathcal{I}\ identifies\ G^* \end{aligned}$$

A: $\lceil \log_2 \omega(\mathcal{E}(G^*)) \rceil$

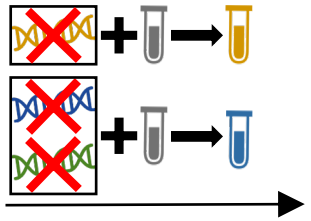
Sufficiency

Assign an optimal coloring to each chordal component, with each color corresponding to a number $0, \ldots, \omega(\mathcal{E}(G^*))$. Let $c(j)$ denote the color (number) assigned to node j.

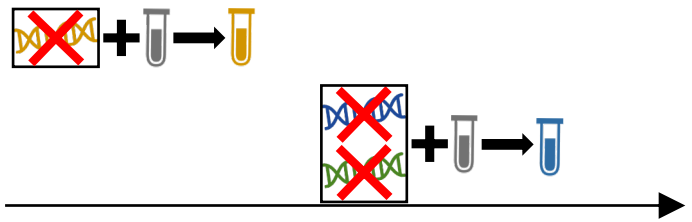
Then, let $\mathrm{I_k} = \{j \mid b\big(c(j)\big)_k = 1\}$. The argument follows the same as in the case with no observational data.



Two Optimal Strategies for Active Learning of Causal Models From Interventional Data (Hauser and Bühlmann, 2012)



Non-adaptive



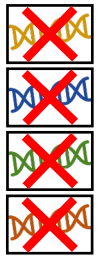
Adaptive

max $info(interventions)$
s.t. $cost(interventions) \leq budget$

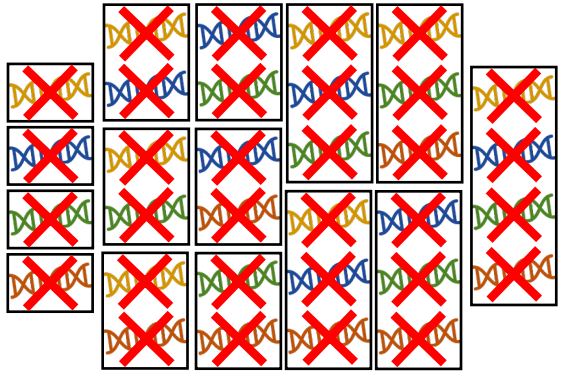
Fixed budget

min $cost(interventions)$
s.t. $interventions\ identify\ G^*$

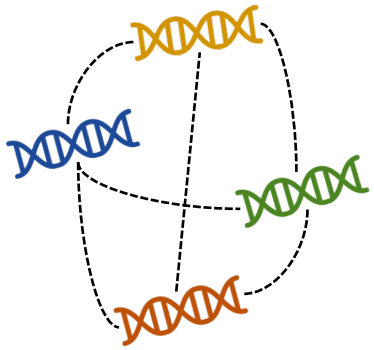
Min-cost identification



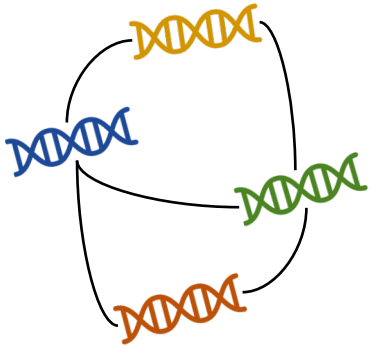
Bounded



Unbounded

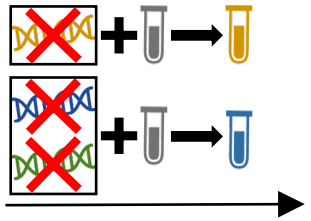


No observational data

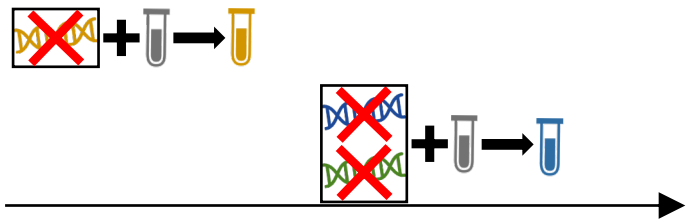


Observational data

- Up to B nodes per intervention: Learning Causal Graphs with Small Interventions (Shanmugam et al., 2015)
- With different costs per intervention: Cost-Optimal Learning of Causal Graphs (Kocaoglu et al., 2017)



Non-adaptive



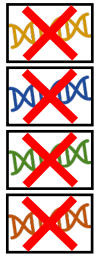
Adaptive

max $info(interventions)$
s.t. $cost(interventions) \leq budget$

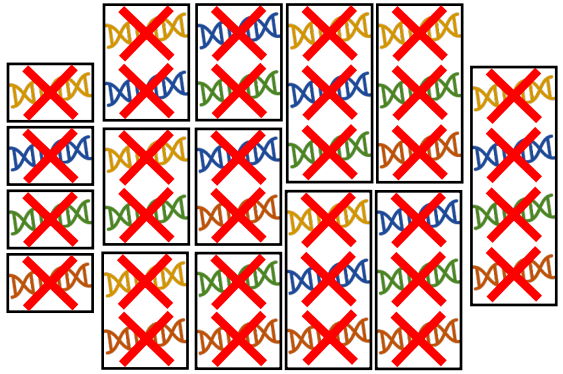
Fixed budget

min $cost(interventions)$
s.t. $interventions\ identify\ G^*$

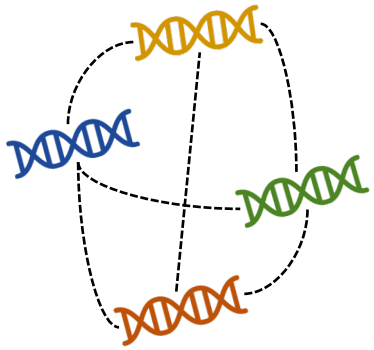
Min-cost identification



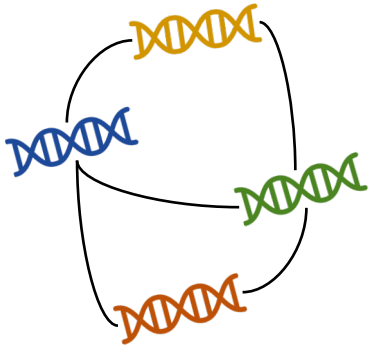
Bounded



Unbounded



No observational data



Observational data

Single-node interventions for a fixed MEC, with a fixed budget

Task: *Pick K single-node interventions to maximize the expected number of oriented edges.*

$$\begin{aligned} \max \quad & \mathbb{E}_{G \sim Unif([G^*])}[dir(\mathcal{E}_{\mathcal{J}}(G))] \\ \text{s.t.} \quad & |\mathcal{J}| \leq K \\ & \forall\, I \in \mathcal{J}, |\mathrm{I}| \leq 1 \end{aligned}$$

Submodularity

- A set function $f: 2^{\Omega} \rightarrow \mathbb{R}$ is **submodular** if for all $A \subset B$ and $b \in \Omega \setminus B$, we have $f(A \cup \{b\}) - f(A) \geq f(B \cup \{b\}) - f(B)$
- A set function is **monotonically increasing** if for all $A \subset B$, we have $f(A) \leq f(B)$
- Greedy optimization (up to size K):
  - Start from $A_0 = \emptyset$, then repeat K times:
  - Let $a_{i+1} = \text{argmax}_{a \in \Omega \setminus A_i} f(A_i \cup \{a\})$, and $A_{i+1} = A_i \cup \{a_{i+1}\}$
- For a monotonically increasing submodular function, greedy optimization finds a set A_K such that $f(A_k) \geq \left(1 - \frac{1}{e}\right) \max_{|A| \leq K} f(A)$

Submodularity of directed edges

Theorem (Ghassami et al., 2018)

Let Ω be the set of single-node interventions, let G be any DAG, and let $f_G(\mathcal{I}) = dir(\mathcal{E}_{\mathcal{I}}(G))$. Then $f_G(\mathcal{I})$ is monotonically increasing and submodular.

Budgeted Experiment Design for Causal Structure Learning (Ghassami et al., 2018)

Submodularity of directed edges

Theorem (Ghassami et al., 2018)

Let $E_{\mathcal{I}_1}$ be the set of edges oriented by the set $\mathcal{I}_1$ of single-node interventions, and $E_{\mathcal{I}_2}$ the set of edges oriented by the set $\mathcal{I}_2$ of single-node interventions. Then $E = E_{\mathcal{I}_1} \cup E_{\mathcal{I}_2}$ is complete under Meek's orientation rules.

Why does this imply submodularity?

$$f_D(\mathcal{I} \cup \{i\}) - f_D(\mathcal{I}) = |E_{\{i\}} \setminus E_{\mathcal{I}}|$$

So, checking submodularity reduces to checking that $\mathcal{I} \subset \mathcal{I}' \Rightarrow \left|E_{\{i\}} \setminus E_{\mathcal{I}}\right| \geq \left|E_{\{i\}} \setminus E_{\mathcal{I}'}\right|$

Submodularity of directed edges

Theorem (Ghassami et al., 2018)

Let $E_{\mathcal{I}_1}$ be the set of edges oriented by the set $\mathcal{I}_1$ of single-node interventions, and $E_{\mathcal{I}_2}$ the set of edges oriented by the set $\mathcal{I}_2$ of single-node interventions. Then $E = E_{\mathcal{I}_1} \cup E_{\mathcal{I}_2}$ is complete under Meek's orientation rules.

Proof sketch:

- Check the pre-conditions for each Meek rule which involves 2 oriented edges $X_1 \to X_2$ and $X_3 \to X_4$
- If E is not complete, then $X_1 \to X_2$ must be from $E_{\mathcal{I}_1}$ and $X_3 \to X_4$ must be from $E_{\mathcal{I}_2}$
- Show that $X_1 \to X_2 \in E_{\mathcal{I}_1}$ implies that $X_3 \to X_4 \in E_{\mathcal{I}_2}$

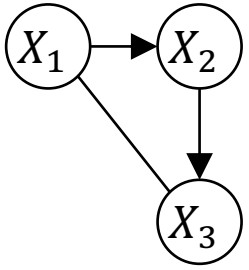
Submodularity of directed edges

Proof sketch:

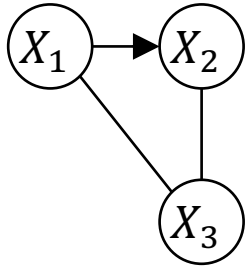
- Check the pre-conditions for each Meek rule which involves 2 oriented edges $X_1 \to X_2$ and $X_3 \to X_4$
- If E is not complete, then $X_1 \to X_2$ must be from $E_{\mathcal{I}_1}$ and $X_3 \to X_4$ must be from $E_{\mathcal{I}_2}$
- Show that $X_1 \to X_2 \in E_{\mathcal{I}_1}$ implies that $X_3 \to X_4 \in E_{\mathcal{I}_2}$

Example:

Say the "no cycles" rule is invoked

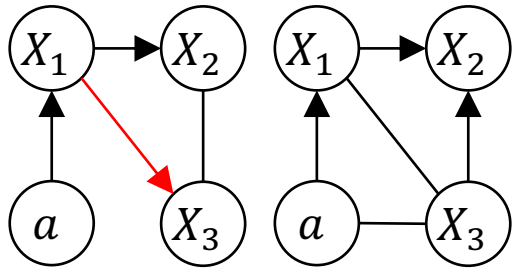


Can we have $X_1 \to X_2 \in E_{\mathcal{I}_1}$ without any of the other edges?



It can't be directly from an intervention on X_1 or X_2, otherwise $X_2 \to X_3$ or $X_1 \to X_3$ would also be oriented

It can't be from the "no new v-structures" rule, or else $X_1 \to X_3$ or $X_3 \to X_2$ would be oriented



Counting and sampling from a Markov equivalence class

Theorem 4, Wienöbst et al., 2020

Let $\mathcal{C}(G)$ denote the set of maximal cliques in a chordal graph G. Then, we can sample uniformly from $[G]$ with p nodes and e edges in $\mathcal{O}(p+e)$ with $\mathcal{O}(|\mathcal{C}(G)|^2 \cdot p \cdot e)$ setup time.

Thus, we may efficiently approximate $\mathbb{E}_{G\sim Unif([G^*])}[dir(\mathcal{E}_{\mathcal{I}}(G))]$ for any $\mathcal{I}$ by sampling and Monte-Carlo averaging.

Submodularity and monotonicity are preserved under positive linear combination, so the full problem is submodular.

Polynomial-Time Algorithms for Counting and Sampling Markov Equivalent DAGs (Wienöbst et al., 2020)

Part III: Adaptive experimental design strategies

Greedy strategies

	Min-max	Bayes
Edge	$\min_I \max_{G \in [G^*]_{\mathcal{J}}} und(\mathcal{E}_{\mathcal{J} \cup \{I\}}(G))$ [1,2]	$\min_I \mathbb{E}_{G \sim Unif([G^*]_{\mathcal{J}})}[und(\mathcal{E}_{\mathcal{J} \cup \{I\}}(G))]$
Entropy	$\min_I \max_{G \in [G^*]_{\mathcal{J}}} \lvert[G]_{\mathcal{J} \cup \{I\}}\rvert$ [3]	$\min_I \mathbb{E}_{G \sim Unif([G^*]_{\mathcal{J}})}[\log\lvert[G]_{\mathcal{J} \cup \{I\}}\rvert]$ [3]
Clique number	$\min_I \max_{G \in [G^*]_{\mathcal{J}}} \omega(\mathcal{E}_{\mathcal{J} \cup \{I\}}(G))$	$\min_I \mathbb{E}_{G \sim Unif([G^*]_{\mathcal{J}})}[\omega(\mathcal{E}_{\mathcal{J} \cup \{I\}}(G))$

[1] Two Optimal Strategies for Active Learning of Causal Models From Interventional Data (Hauser and Bühlmann, 2012)
[2] Learning Causal Graphs with Small Interventions (Shanmugam et al., 2015)
[3] Active Learning of Causal Networks with Intervention Experiments and Optimal Designs (He and Geng, 2008)

Greedy strategies

	Min-max	Bayes
Edge		
Entropy		
Clique number		

Greedy strategies

	Min-max	Bayes
Edge	$\min_{I} \max_{G \in [G^*]_{\mathcal{I}}} und(\mathcal{E}_{\mathcal{I} \cup \{I\}}(G))$ [1,2]	
Entropy		
Clique number		

[1] Two Optimal Strategies for Active Learning of Causal Models From Interventional Data (Hauser and Bühlmann, 2012)
[2] Learning Causal Graphs with Small Interventions (Shanmugam et al., 2015)

Greedy strategies

	Min-max	Bayes
Edge	$\min_I \max_{G \in [G^*]_{\mathcal{I}}} und(\mathcal{E}_{\mathcal{I} \cup \{I\}}(G))$ [1,2]	$\min_I \mathbb{E}_{G \sim Unif([G^*]_{\mathcal{I}})}[und(\mathcal{E}_{\mathcal{I} \cup \{I\}}(G))]$
Entropy		
Clique number		

[1] Two Optimal Strategies for Active Learning of Causal Models From Interventional Data (Hauser and Bühlmann, 2012)
[2] Learning Causal Graphs with Small Interventions (Shanmugam et al., 2015)

Greedy strategies

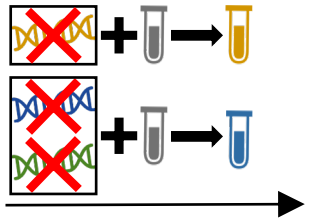
	Min-max	Bayes
Edge	$\min_I \max_{G \in [G^*]_{\mathcal{J}}} und(\mathcal{E}_{\mathcal{J} \cup \{I\}}(G))$ 1,2	$\min_I \mathbb{E}_{G \sim Unif([G^*]_{\mathcal{J}})}[und(\mathcal{E}_{\mathcal{J} \cup \{I\}}(G))]$
Entropy	$\min_I \max_{G \in [G^*]_{\mathcal{J}}} \log\lvert[G]_{\mathcal{J} \cup \{I\}}\rvert$ 3	$\min_I \mathbb{E}_{G \sim Unif([G^*]_{\mathcal{J}})}[\log\lvert[G]_{\mathcal{J} \cup \{I\}}\rvert]$ 3
Clique number		

[1] Two Optimal Strategies for Active Learning of Causal Models From Interventional Data (Hauser and Bühlmann, 2012)
[2] Learning Causal Graphs with Small Interventions (Shanmugam et al., 2015)
[3] Active Learning of Causal Networks with Intervention Experiments and Optimal Designs (He and Geng, 2008)

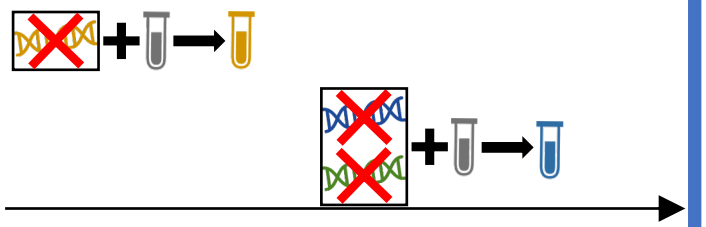
Greedy strategies

	Min-max	Bayes
Edge	$\min_I \max_{G \in [G^*]_{\mathcal{J}}} und(\mathcal{E}_{\mathcal{J} \cup \{I\}}(G))$ [1,2]	$\min_I \mathbb{E}_{G \sim Unif([G^*]_{\mathcal{J}})}[und(\mathcal{E}_{\mathcal{J} \cup \{I\}}(G))]$
Entropy	$\min_I \max_{G \in [G^*]_{\mathcal{J}}} \log\lvert[G]_{\mathcal{J} \cup \{I\}}\rvert$ [3]	$\min_I \mathbb{E}_{G \sim Unif([G^*]_{\mathcal{J}})}[\log\lvert[G]_{\mathcal{J} \cup \{I\}}\rvert]$ [3]
Clique number	$\min_I \max_{G \in [G^*]_{\mathcal{J}}} \omega(\mathcal{E}_{\mathcal{J} \cup \{I\}}(G))$	$\min_I \mathbb{E}_{G \sim Unif([G^*]_{\mathcal{J}})}[\omega(\mathcal{E}_{\mathcal{J} \cup \{I\}}(G))$

[1] Two Optimal Strategies for Active Learning of Causal Models From Interventional Data (Hauser and Bühlmann, 2012)
[2] Learning Causal Graphs with Small Interventions (Shanmugam et al., 2015)
[3] Active Learning of Causal Networks with Intervention Experiments and Optimal Designs (He and Geng, 2008)



Non-adaptive



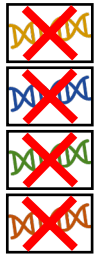
Adaptive

max $info(interventions)$
s.t. $cost(interventions) \leq budget$

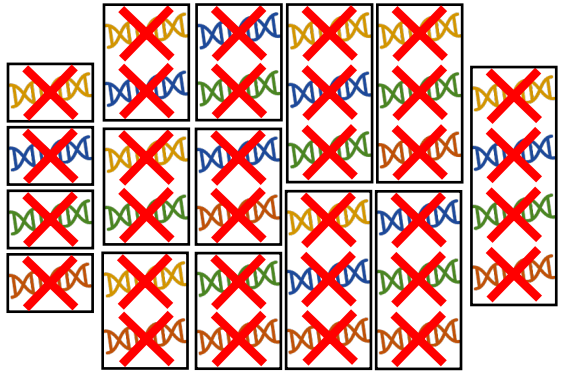
Fixed budget

min $cost(interventions)$
s.t. $interventions\ identify\ G^*$

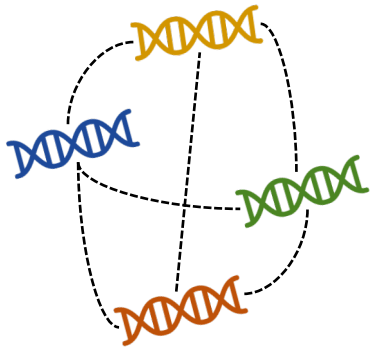
Min-cost identification



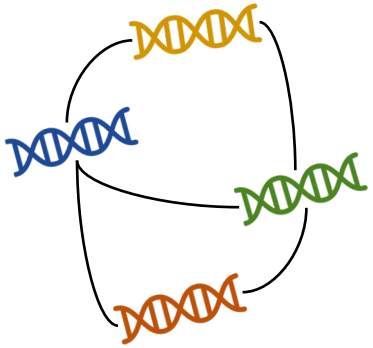
Bounded



Unbounded

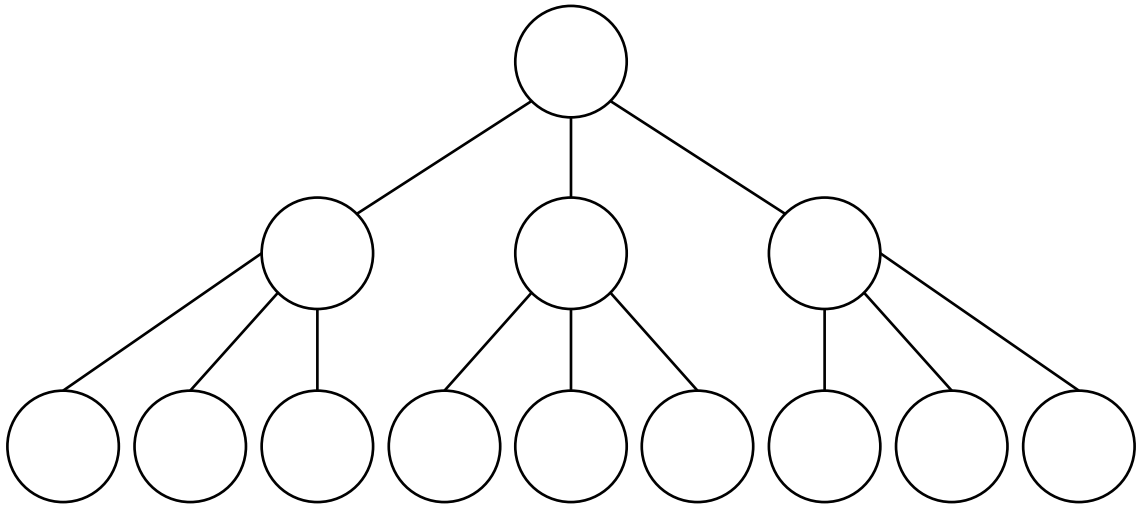


No observational data



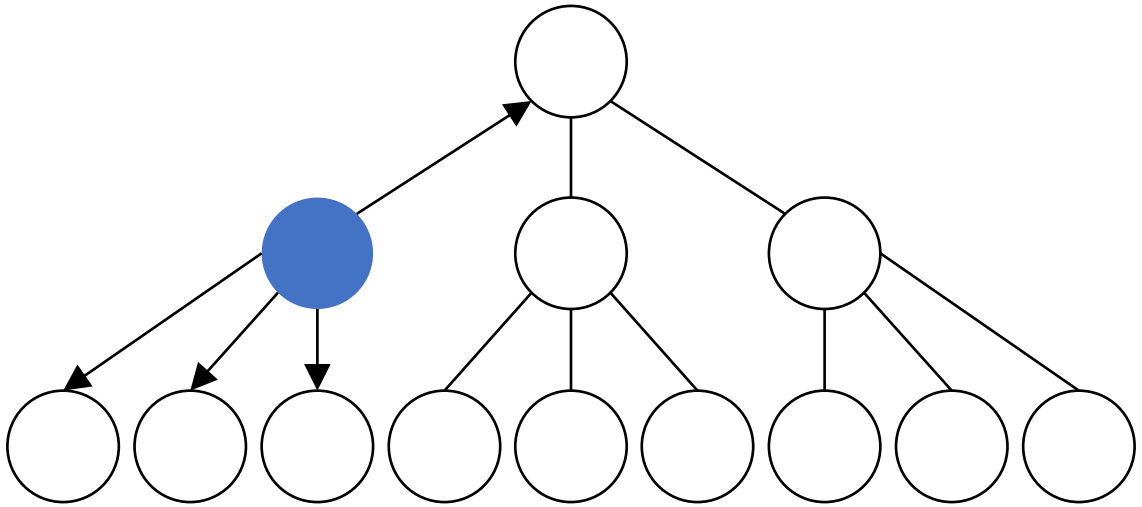
Observational data

Single-node experimental design on trees

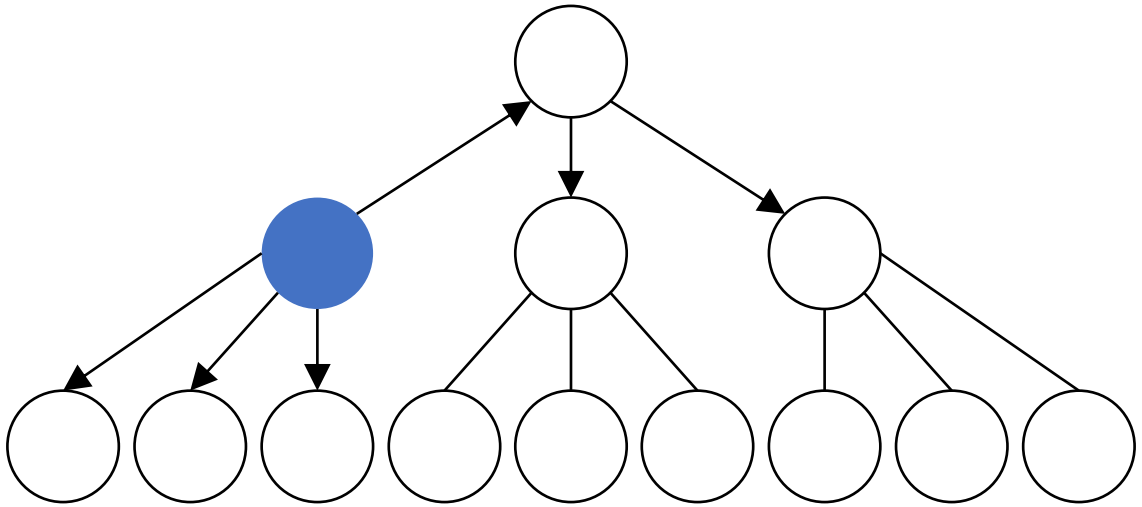


Essential graph

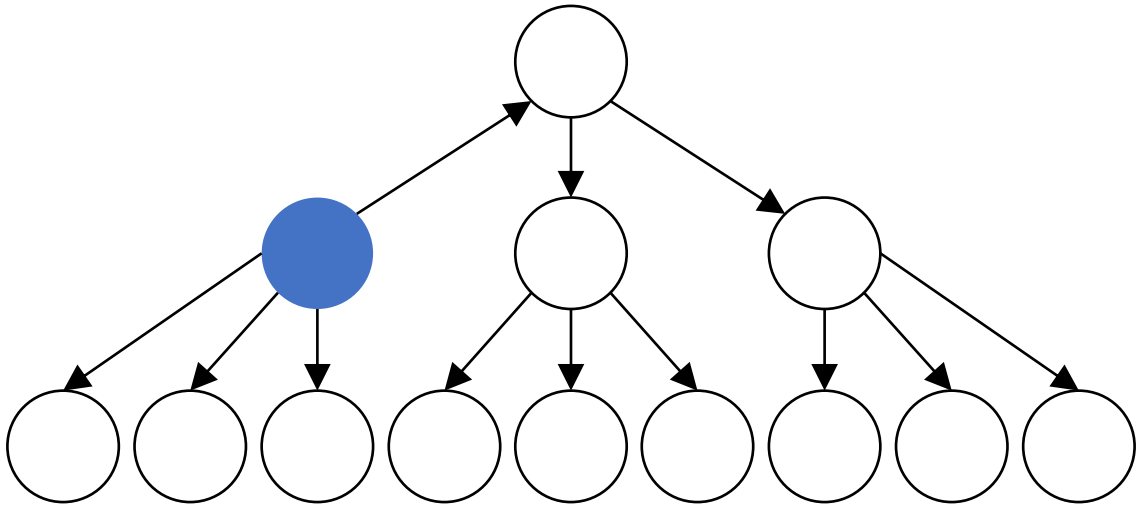
Single-node experimental design on trees



Single-node experimental design on trees

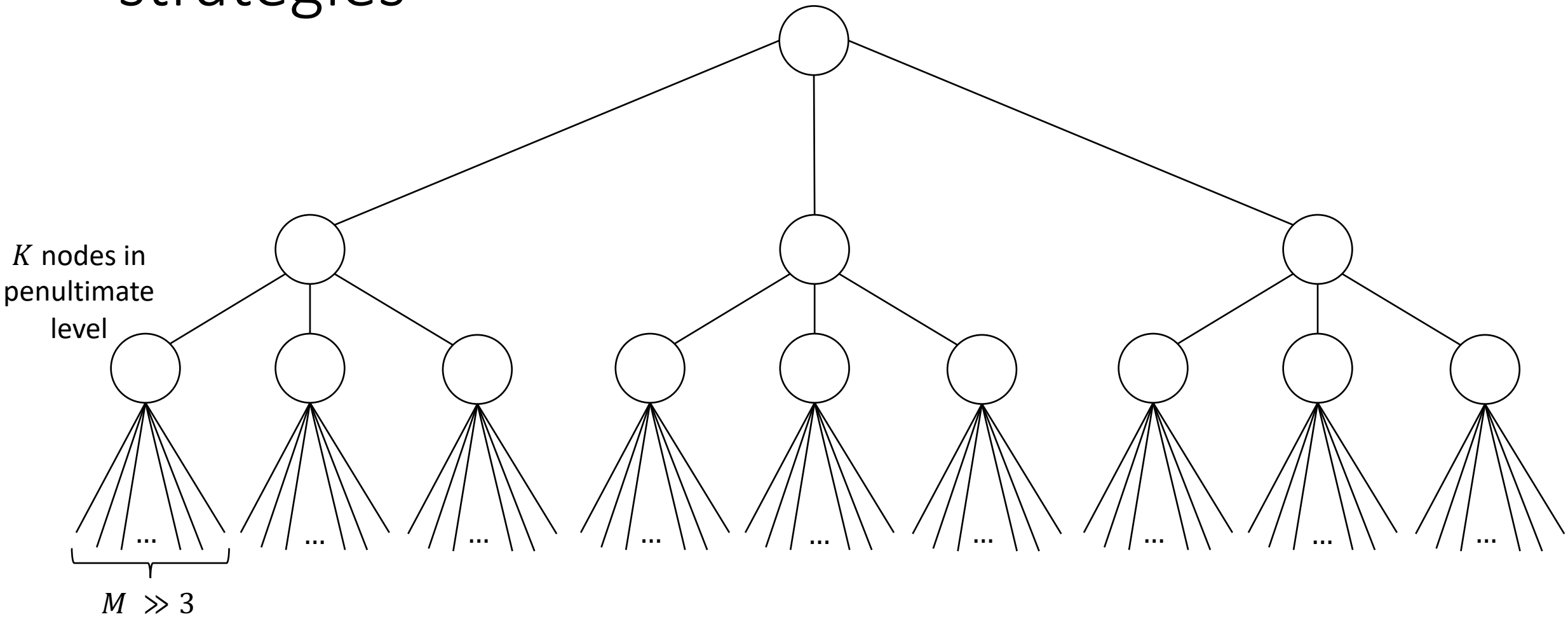


Single-node experimental design on trees



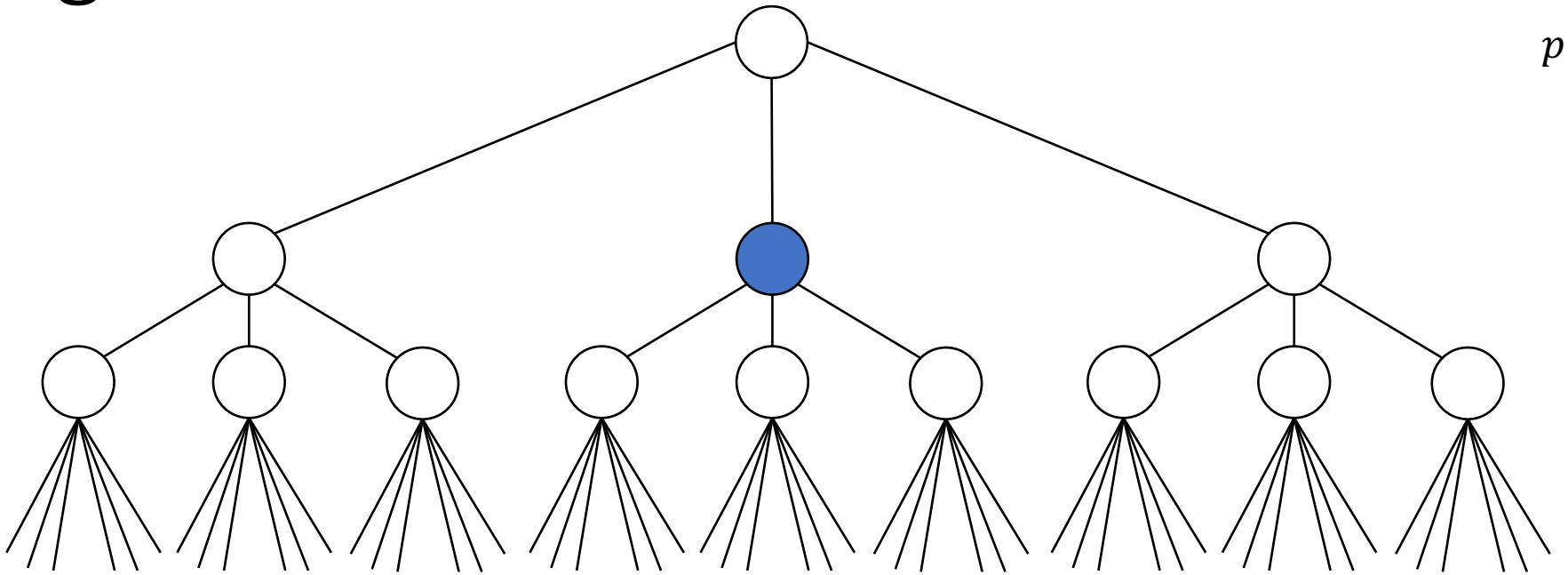
The root node uniquely determines the DAG

Suboptimality of information-greedy strategies



Sample Efficient Active Learning of Causal Trees (Greenewald et al., 2019)

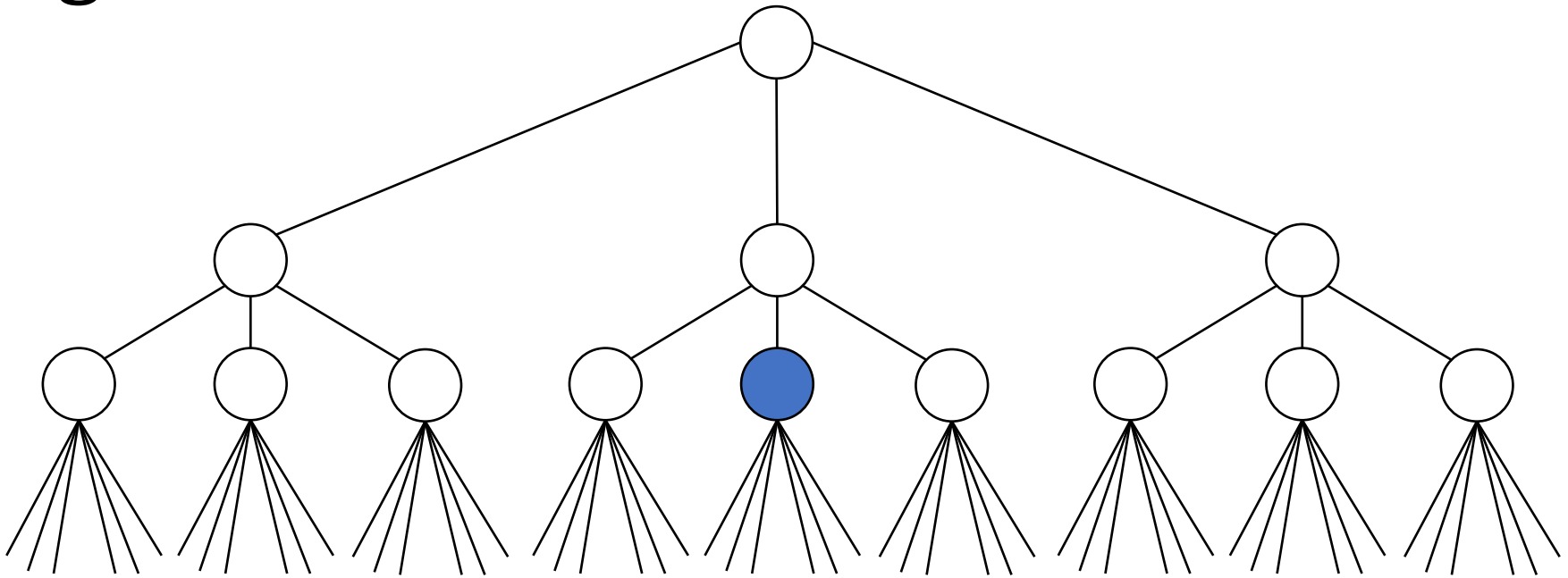
Suboptimality of information-greedy strategies



p total nodes

Expected information gain $\leq \frac{1}{p}\log_2 p + \frac{p-1}{p}\log_2 4 \approx 2$

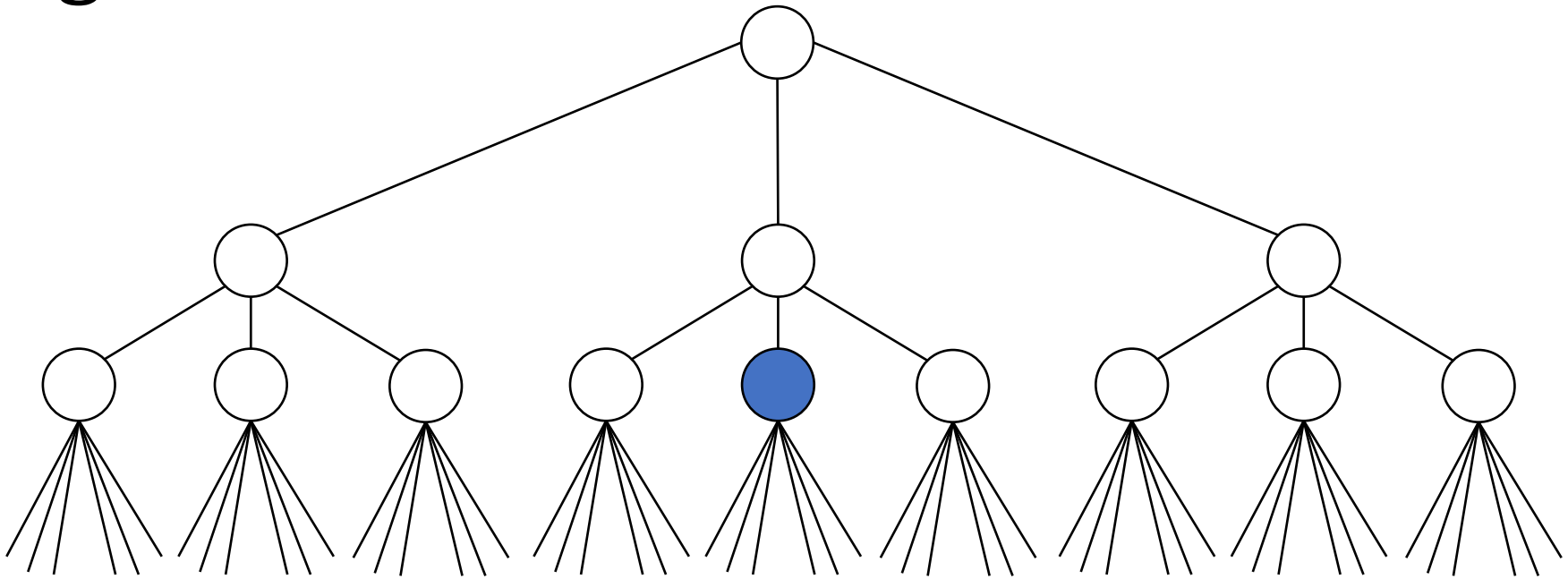
Suboptimality of information-greedy strategies



Expected information gain $\geq \frac{1}{K}\log_2 p \geq \frac{1}{K}\log_2 M$

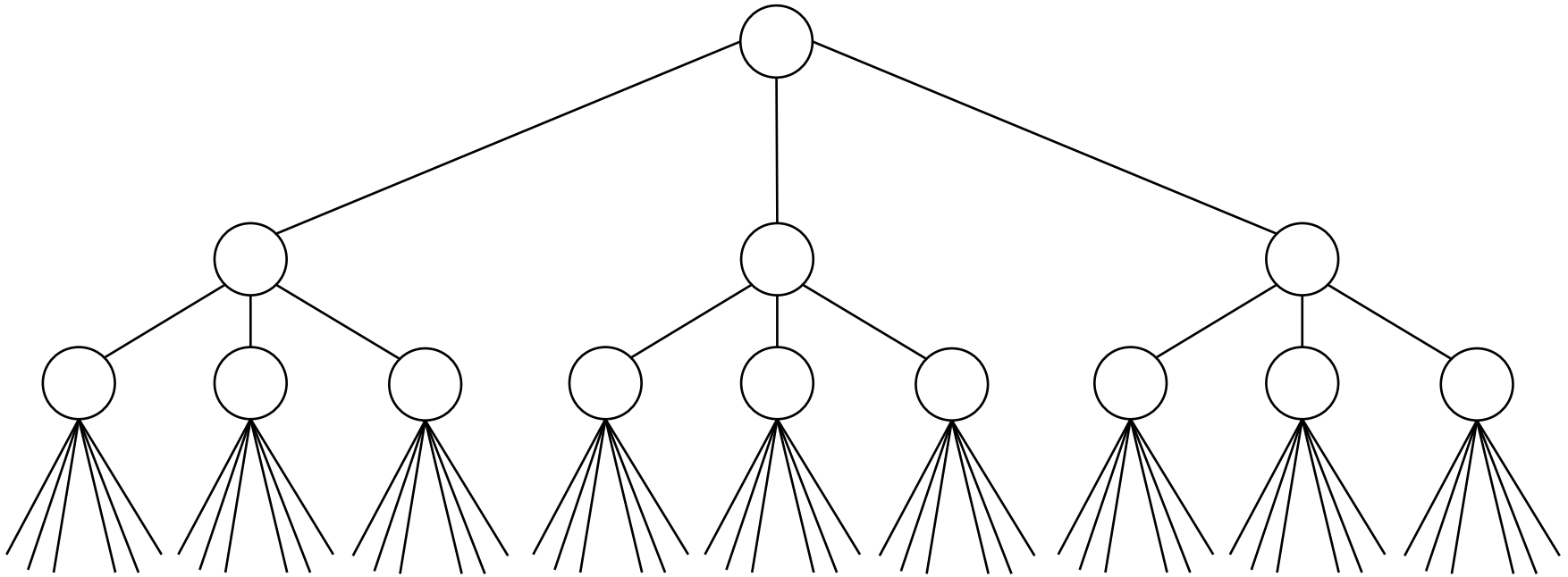
M can be picked so that the nodes in the penultimate layer are always preferable (e.g., $M = O\left(2^{K^2}\right)$)

Suboptimality of information-greedy strategies



The expected number of interventions required by this strategy is at least $K/2$

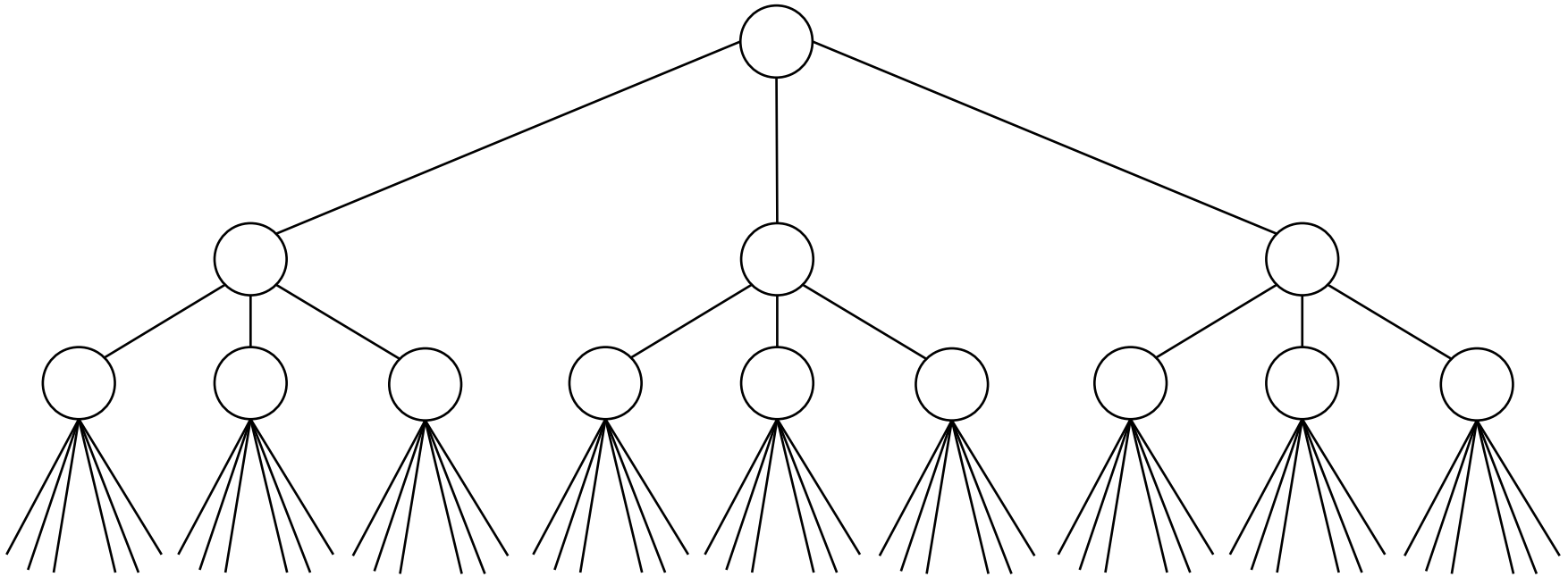
Central node algorithm



Central node algorithm[1] can find the root in at most $(\log_3 K) + 1$ interventions

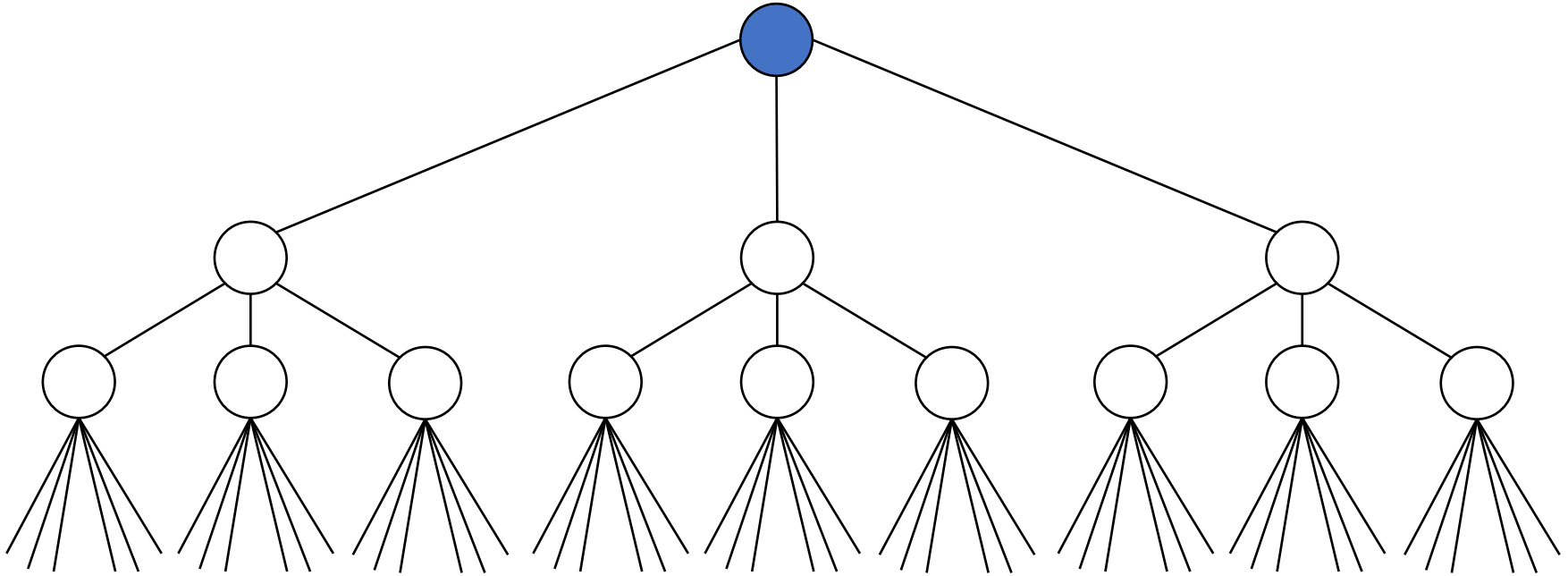
[1] Sample Efficient Active Learning of Causal Trees (Greenewald et al., 2019)

Central node algorithm

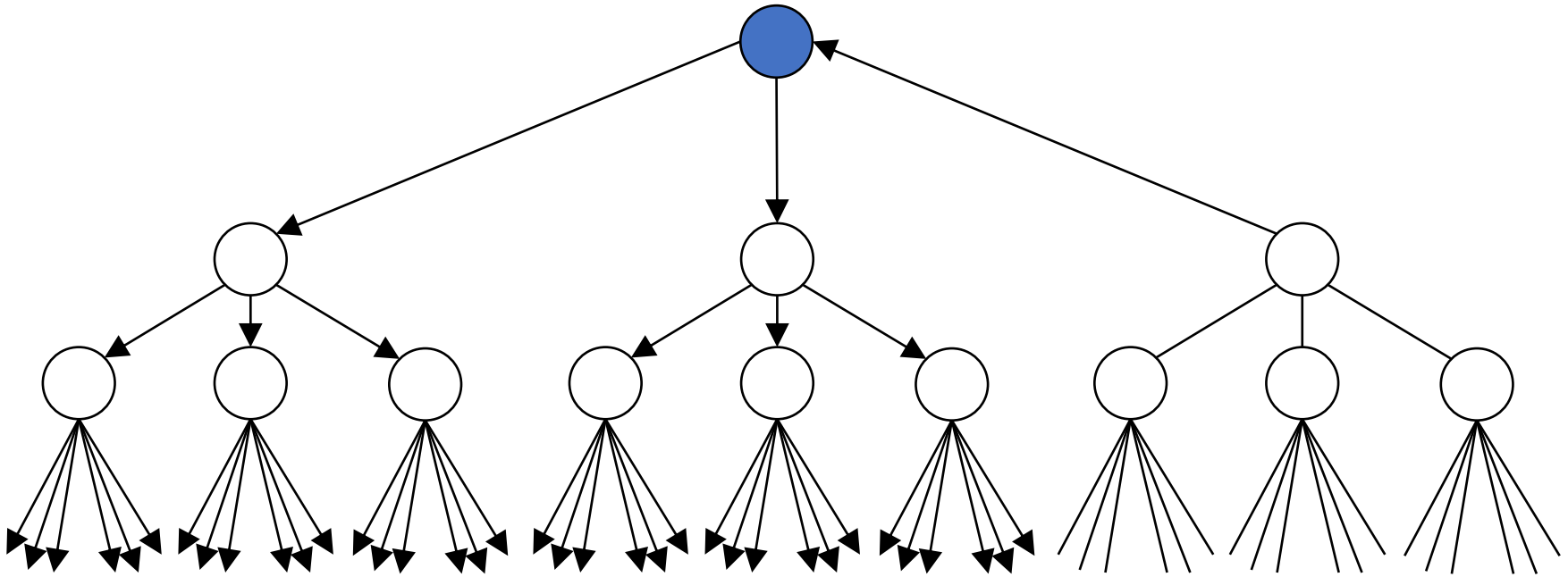


Central node algorithm: Intervene on a node such that, after removing it, the largest remaining connected component is as small as possible. Equivalent to min-max entropy choice.

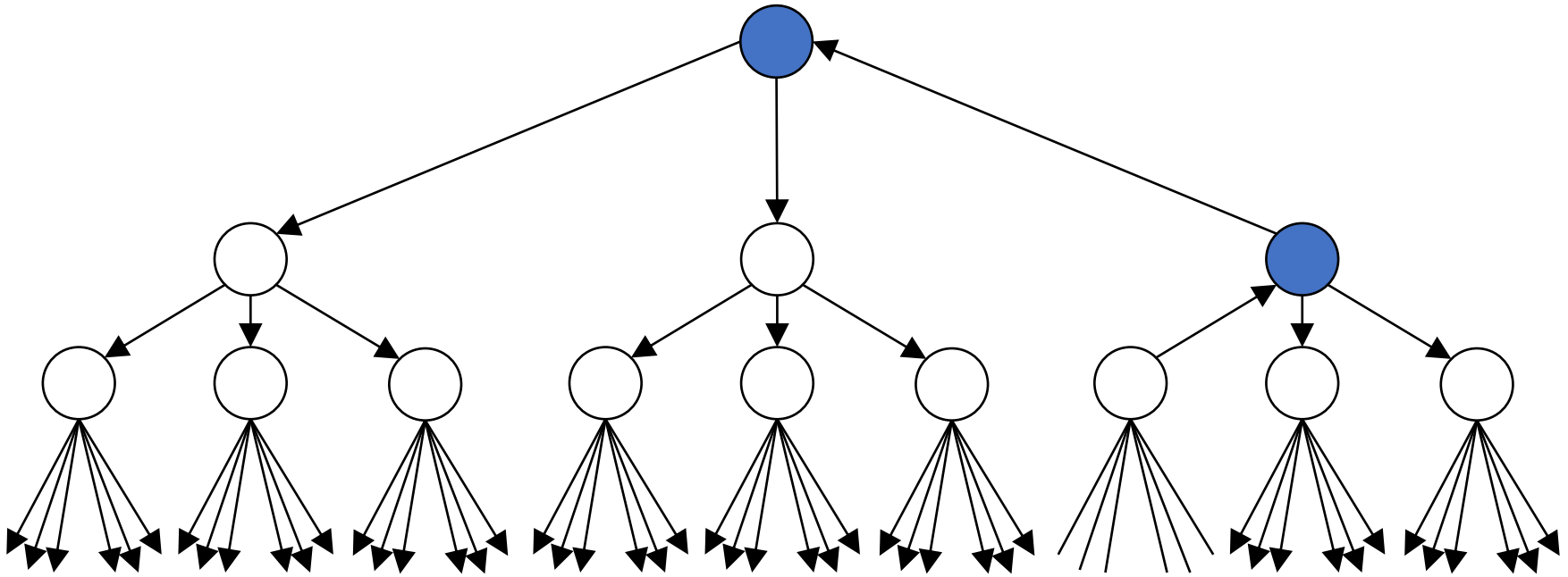
Central node algorithm



Central node algorithm



Central node algorithm



Central node algorithm

- In a tree on p nodes, there is at least one node such that its removal splits the tree into components each having at most $\frac{p}{2}$ nodes.[1]
- So, we can always get at least one bit of information per intervention $\Rightarrow$ a tree on p nodes takes at most $\lceil \log_2 p \rceil$ interventions to orient

Theorem (Greenewald et al., 2019)

The expected number of interventions required by the central node algorithm is at most twice the expected number of interventions required by the optimal algorithm.

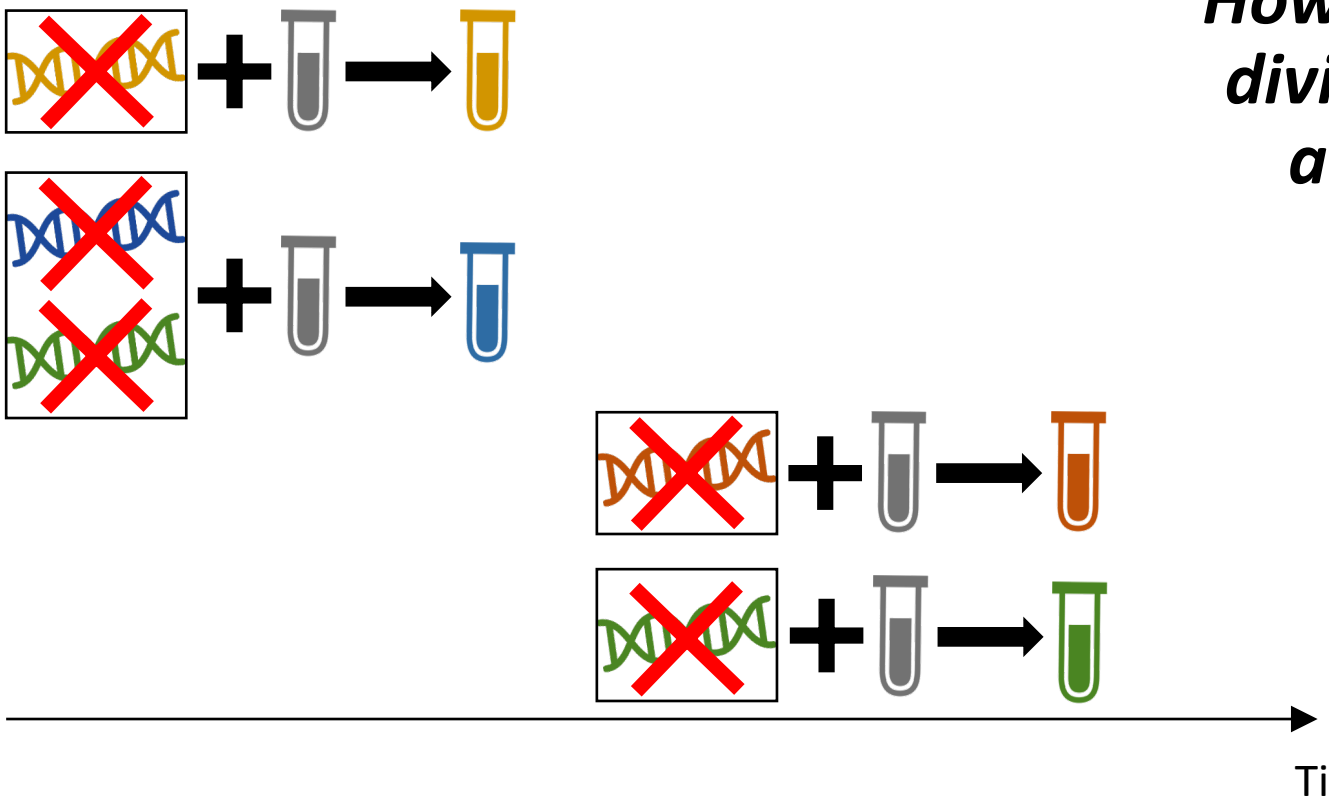
[1]Sur les assemblages de lignes (Jordan, 1869)

Part IV: Targeted experimental design and other open challenges

Targeted experimental design

- So far, we've treated all structural causal information about our graph as equally important
- Rarely true!
  - We may only care about a subset of structural information (e.g., existence of certain paths)
    - ABCD-Strategy: Budgeted Experimental Design for Targeted Causal Structure Discovery (Agrawal et al., 2019)
  - We may only care about finding an "optimal" information for some goal, e.g. so that the interventional distribution is close to some target distribution
    - Matching a Desired Causal State via Shift Interventions
- ***For different learning goals, how does interventional and sample complexity change?***

Batched experimental design



How do we optimally divide interventions across batches?

ABCD-Strategy: Budgeted Experimental Design for Targeted Causal Structure Discovery (Agrawal et al., 2019)

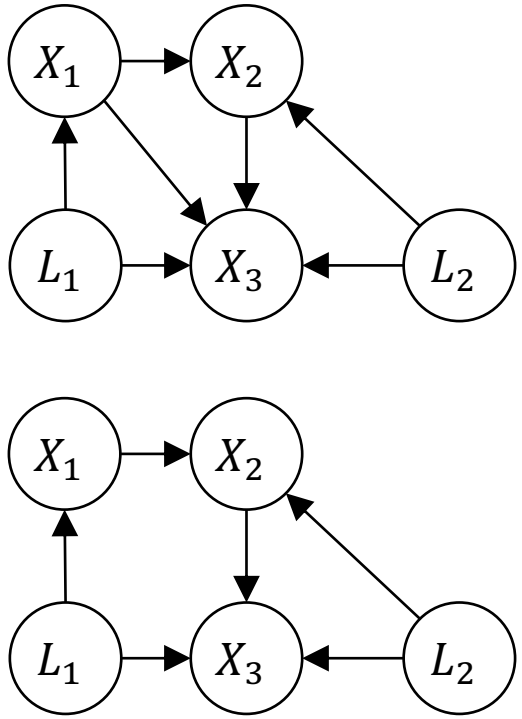
Batched bandit problems (Perchet et al., 2016)

Noisy interventional data

- ***How much better can we do than the "repeat until desired confidence" strategy for dealing with noise?***
  - Sample Efficient Active Learning of Causal Trees (Greenewald et al., 2019) use a multiplicative weights algorithm in the binary setting
  - Learning and Testing Causal Models with Interventions (Acharya et al., 2018) study "goodness-of-fit" testing and other problems in the discrete setting
    - Given a true distribution f_* and a hypothesized distribution $\tilde{f}$ both Markov to G, are all associated interventional distributions (f_*^k and $\tilde{f}^k$ for I_k any do-intervention) with ϵ total variation distance?

Causally insufficient systems

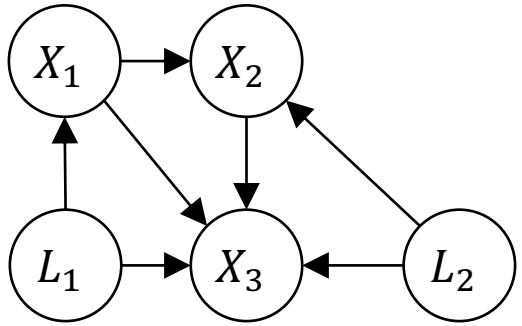
We can’t always identify causal graphs in insufficient systems from just single node interventions

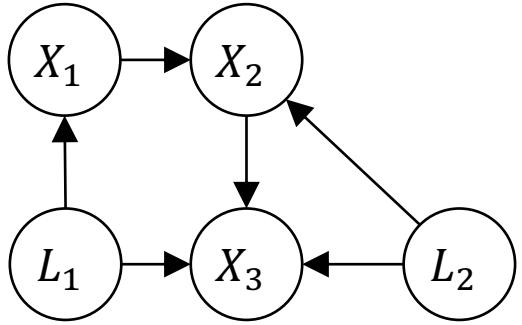


[1]Causation and Intervention (Eberhardt, 2008)

Causally insufficient systems

But we can identify them when the number of intervened nodes is unbounded

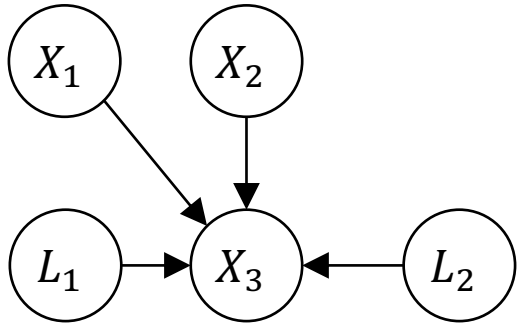


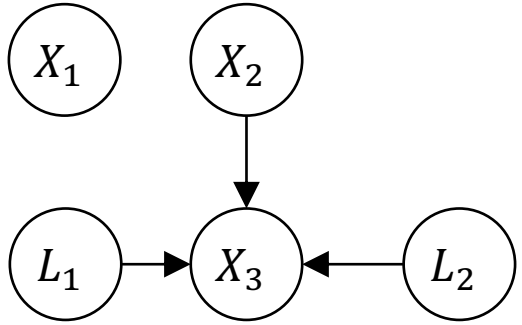


[1] Causation and Intervention (Eberhardt, 2008)

Causally insufficient systems

But we can identify them when the number of intervened nodes is unbounded





$I = \{X_1, X_2\}$

[1] Causation and Intervention (Eberhardt, 2008)

Causally insufficient systems

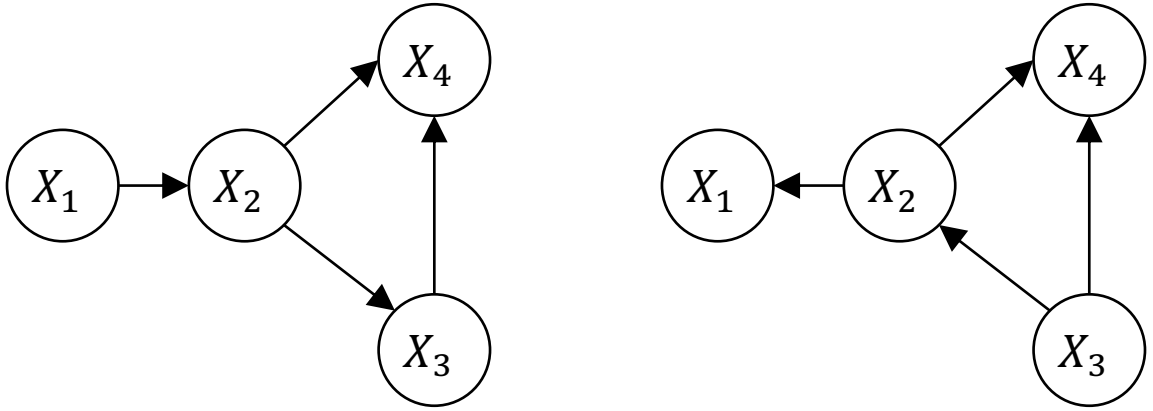
What graphs can we identify when the number of intervention targets is bounded by K?

Lower bounds on interventional complexity

- Suppose a friend tells you the true graph G^*, but you're skeptical and want to check it, while using the minimal number of single-node interventions
- If G^* is a tree: only 1 intervention is required (at the root)
- If G^* is a clique on p nodes: $\lfloor \frac{p}{2} \rfloor$ interventions are required
- Active Structure Learning of Causal DAGs via Directed Clique Trees (Squires et al., 2020) give a polynomial-time algorithm for finding the "minimal verifying intervention set" (MVIS) for any DAG G^*, the size of which is denoted $m(G^*)$

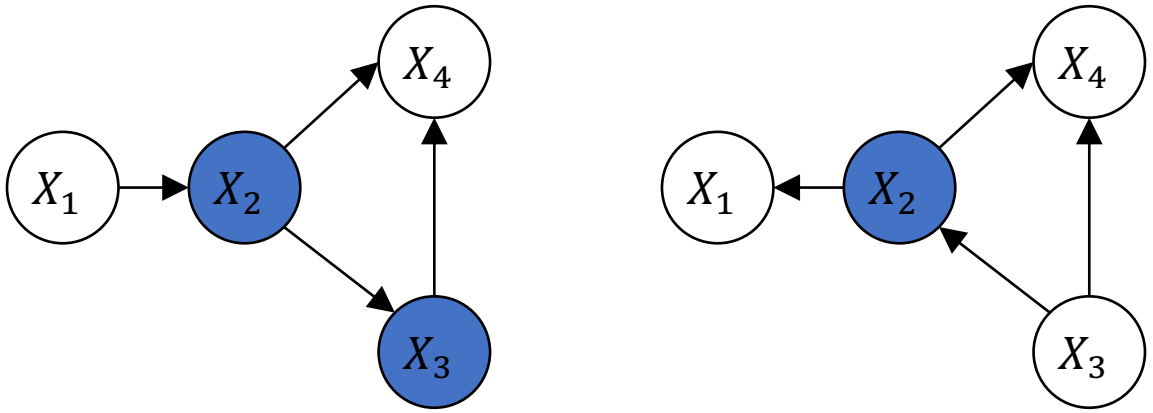
Lower bounds on interventional complexity

- $m(G^*)$ is not necessarily the same across a Markov equivalence class



Lower bounds on interventional complexity

- $m(G^*)$ is not necessarily the same across a Markov equivalence class



- Squires et al. (2020) give a "universal" lower bound on $m(G^*)$ that holds across the equivalence class, improved in Almost Optimal Universal Lower Bound for Learning Causal DAGs with Atomic Interventions (Porwal et al., 2021)