Chapter 6

Experimental Design

In this lecture, we will discuss how one might *design* interventions to either learn more about the underlying system (e.g., the causal structure over variables) or optimize for some desired outcomes. In general, this moves from the *passive* learning setting discussed in Chapter 5 to an *active* setting, where one performs interventions sequentially using the samples gathered in previous experiments.

We will mainly talk about the *noiseless* setting, where enough samples are assumed for each intervention to decide its effect. For this setting, we first characterize the information that can be learned via each intervention, which involves the concepts of interventional Markov equivalence classes and Meek rules. We then consider experimental design for *full-graph identification* and discuss about verifying intervention sets and competitive ratios. We showcase a popular adaptive approach to cast such concepts into usage.

Aside from full-graph identification, one may be interested in more *targeted* information. We conclude this section with a brief discussion on this. We will also provide a discussion on the *noisy* setting, where *uncertainty* quantification are essential to calibrate for finite samples. These discussions lead to more active areas of open research.

6.1 Identifiability from Interventions

In the previous chapters, we have shown that the causal graph can be identified up to its Markov equivalence class (MEC) given observational data. Such an MEC is characterize by the essential graph, where the unknown orientations are denoted undirected edges. With interventions, one may learn additional orientations. We will now develop an understanding of which undirected edges in the essential graph can be oriented by an intervention.

We begin by defining interventions. Consider random variables $X = (X_1, \dots, X_p)$ that factorizes with respect to DAG $\mathcal{G}$, i.e., the joint distribution $P(X) = \prod_{i \in [p]} P(X_i \mid X_{\mathrm{pa}_{\mathcal{G}}(i)})$.[1]

Definition 6.1 (Intervention)**.** *An* **intervention** $I \subseteq [p]$ *is an experiment where the conditional distributions* $P(X_i \mid X_{\mathrm{pa}_{\mathcal{G}}(i)})$ *for* $i \in I$ *are changed. The* **interventional distribution** P^I *is defined as*

$$P^I(X) = \prod_{i \in I} P^I(X_i \mid X_{\mathrm{pa}_{\mathcal{G}}(i)}) \prod_{i \notin I} P^I(X_i \mid X_{\mathrm{pa}_{\mathcal{G}}(i)}).$$

Observational data is a special case where $I = \varnothing$. *Hard* interventions refer to changes that eliminate the dependency between i and $\mathrm{pa}_{\mathcal{G}}(i)$. For example, *do* interventions are a special case of hard interventions, where $P^I(X_i \mid X_{\mathrm{pa}_{\mathcal{G}}(i)}) = 1\{X_i = a\}$ for some $a \in \mathbb{R}$. *Soft* interventions modify this dependency without removing it. For example, *shift* interventions are a special case of soft interventions, where $P^I(X_i = x + a \mid X_{\mathrm{pa}_{\mathcal{G}}(i)}) = P(X_i = x \mid X_{\mathrm{pa}_{\mathcal{G}}(i)})$ for some $a \in \mathbb{R}$ and all x in the support of X_i.

[1] Here $[p] = \{1, \dots, p\}$ and $\mathrm{pa}_{\mathcal{G}}(i) = \{j \in [p] \mid j \to i \text{ in } \mathcal{G}\}$.

Next we define interventional Markov equivalence classes, which involves a set of probability pairs that are $\mathcal{I}$-Markov with respect to $\mathcal{G}$.

Definition 6.2 ($\mathcal{I}$-Markov Equivalence Class)**.** *For a set of interventions $\mathcal{I} = \{I_1, ..., I_k\}$, the pair $(f, \{f^I\}_{I\in\mathcal{I}})$ is* **$\mathcal{I}$-Markov** *w.r.t. $\mathcal{G}$ if f factorizes w.r.t. $\mathcal{G}$ and f^I factorizes according to*

$$f^I(X) = \prod_{i\in I} f^I(X_i \mid X_{\mathrm{pa}_\mathcal{G}(i)}) \prod_{i\notin I} f^I(X_i \mid X_{\mathrm{pa}_\mathcal{G}(i)}).$$

Two DAGs $\mathcal{G}_1, \mathcal{G}_2$ are in the same **$\mathcal{I}$-Markov equivalence class ($\mathcal{I}$-MEC)** *or* **$\mathcal{I}$-Markov equivalent** *if any positive distribution that is $\mathcal{I}$-Markov w.r.t. $\mathcal{G}_1$ is also $\mathcal{I}$-Markov w.r.t. $\mathcal{G}_2$.*

Without additional (e.g., parametric) assumptions on the interventional distributions, it is clear from this definition that the underlying DAG $\mathcal{G}$ can only be identifiable up to its $\mathcal{I}$-MEC with interventions $\mathcal{I}$. However, to guarantee that the $\mathcal{I}$-MEC of the underlying DAG $\mathcal{G}$ can be identified, *interventional faithfulness* assumptions on $(P, \{P\}_{I\in\mathcal{I}})$ are needed (c.f., Yang et al. (2018)). We will skip this detail and assume that the $(P, \{P\}_{I\in\mathcal{I}})$ satisfies the interventional faithfulness assumptions. In this case:

With infinite observational data and interventional data from $\mathcal{I}$, DAG $\mathcal{G}$ is identifiable to its $\mathcal{I}$-MEC.

We now characterize $\mathcal{I}$-MECs using graphical properties.

Proposition 6.1. *The following graphical characterizations of MEC and $\mathcal{I}$-MEC are known:*

- *Two DAGs are in the same MEC if and only if they share the same skeleton (adjacencies) and v-structures (induced subgraphs $i \to j \leftarrow k$). See Verma and Pearl (1991) for proofs.*
- *Two DAGs are in the same $\mathcal{I}$-MEC, if they are in the same MEC and they have the same directed edges $\{i \to j \mid i \text{ adjacent to } j, i \in I, j \notin I, I \in \mathcal{I}\}$. See Hauser and Bühlmann (2014) for proofs.*

Example 6.1. *Let $\mathcal{G} = \{1 \to 2\}$ and $\mathcal{G}' = \{1 \leftarrow 2\}$. Let $\mathcal{I} = \{I_1\}$. Then:*

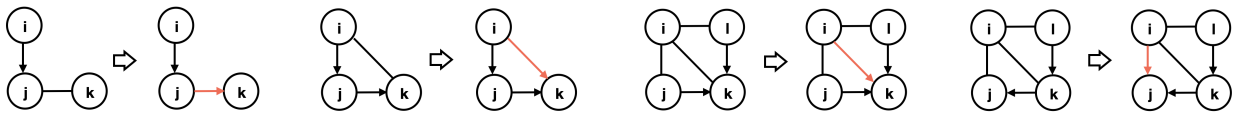
- *$\mathcal{G}$ and $\mathcal{G}'$ are Markov equivalent.*
- *If $I_1 = \{1\}$, then $\mathcal{G}$ and $\mathcal{G}'$ are not $\mathcal{I}$-Markov equivalent, since $1 \leftarrow 2$ in $\mathcal{G}'$ but not in $\mathcal{G}$.*
- *Similarly, if $I_1 = \{2\}$, then $\mathcal{G}$ and $\mathcal{G}'$ are not $\mathcal{I}$-Markov equivalent.*
- *However, if $I_1 = \{1, 2\}$, then $\mathcal{G}$ and $\mathcal{G}'$ are $\mathcal{I}$-Markov equivalent. For example, if we perform a do-intervention on both variables, then we can't tell which node is downstream of the other.*

Proposition 6.1 showed that an intervention I orients all the edges that are cut by I. Additional edges can be oriented using a set of logical relations known as Meek rules Meek (1995).

Proposition 6.2 (Meek Rules, Meek (1995))**.** *We can infer all shared directed edges in $\mathcal{I}$-MEC using the following four rules:*

1. *If $i \to j - k$ and i is not adjacent to k, then $j \to k$.*
2. *If $i \to j \to k$ and $i - k$, then $i \to k$.*
3. *If $i - j, i - k, i - l, j \to k, l \to k$ and j is not adjacent to l, then $i \to k$.*
4. *If $i - j, i - k, i - l, j \leftarrow k, l \to k$ and j is not adjacent to l, then $i \to j$.*

Figure 6.1 illustrates the four Meek rules.



(a) Rule 1 (b) Rule 2 (c) Rule 3 (d) Rule 4

Figure 6.1: Meek Rules.

Proof. We will show that these four rules always hold. To see that they are complete (i.e., *all* shared directed edges in the $\mathcal{I}$-MEC can be inferred from these four rules), see Meek (1995).

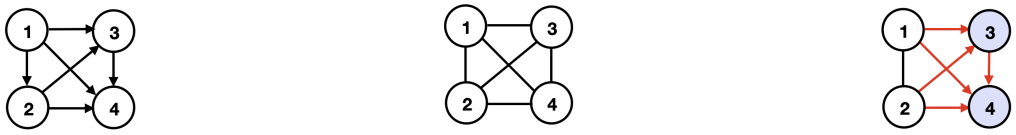
Rule 1: by the fact that all v-structures are oriented in the essential graph and $j - k$ is undirected. It can not be $i \to j \leftarrow k$. Thus we must have $j \to k$.

Rule 2: this is apparent from acyclicity.

Rule 3: if $i \leftarrow k$, then from acyclicity we have $i \leftarrow j$ and $i \leftarrow l$. Then $j \to i \leftarrow l$ creates a v-structure. A contradiction. Therefore there must be $i \to k$.

Rule 4: if $i \leftarrow j$, then from acylicity there is $i \leftarrow k$ and then $i \leftarrow l$. Then $j \to i \leftarrow l$ creates a v-structure. A contradiction. Therefore there must be $i \to j$. □

Using Proposition 6.1 and 6.2, we can fully determine if two DAGs belong the same $\mathcal{I}$-MEC. We can also define *$\mathcal{I}$-essential graphs* similar to essential graphs for MECs. In particular, the **$\mathcal{I}$-essential graphs** for $\mathcal{G}$ is a partially directed graph such that $i \to j$ is oriented if $i \to j$ in *every* DAG in the $\mathcal{I}$-MEC of $\mathcal{G}$, and $i - j$ is undirected if there exists $\mathcal{G}_1, \mathcal{G}_2$ in the $\mathcal{I}$-MEC of $\mathcal{G}$ such that $i \to j$ in $\mathcal{G}_1$ and $i \leftarrow j$ in $\mathcal{G}_2$. Figure 6.2 shows an example. Given the $\mathcal{I}$-essential graph of $\mathcal{G}$, it is easy to check if another DAG $\mathcal{G}'$ is $\mathcal{I}$-Markov equivalent to $\mathcal{G}$.



(a) DAG $\mathcal{G}$. (b) Essential graph. (c) $\mathcal{I}$-essential graph.

Figure 6.2: An example illustrating essential graph and $\mathcal{I}$-essential graph. **(a)** the ground-truth DAG $\mathcal{G}$. **(b)** the essential graph of $\mathcal{G}$. **(c)** the $\mathcal{I}$-essential graph of $\mathcal{G}$ given a set of two interventions $\mathcal{I} = \{\{3\}, \{4\}\}$, where edges oriented by $\mathcal{I}$ are indicated in red.

6.2 Verifying Intervention Sets

We now consider the setting where we want to learn the entire causal graph. In this scenario, we may ask: how can one tell if a set of interventions contain enough information to identify the causal graph? The definition of verifying intervention set aims to articulate this question.

Definition 6.3 (Verifying Intervention Set)**.** *Let $\mathcal{G}$ be a causal graph and $\mathcal{I}$ be a set of interventions. We call $\mathcal{I}$ a* **verifying intervention set** *for $\mathcal{G}$ if the $\mathcal{I}$-MEC only contains $\mathcal{G}$.*

Example 6.2. *If $\mathcal{G}$ is a tree with root node X_i, then an intervention on X_i is a verifying intervention set. For a complete graph with topological order $X_1 < X_2 < \ldots < X_p$, the set of single-node intervention on all odd-numbered nodes is a verifying intervention set. See Figure 6.3.*

We now characterize the necessary and sufficient condition for $\mathcal{I}$ to be a verifying intervention set.

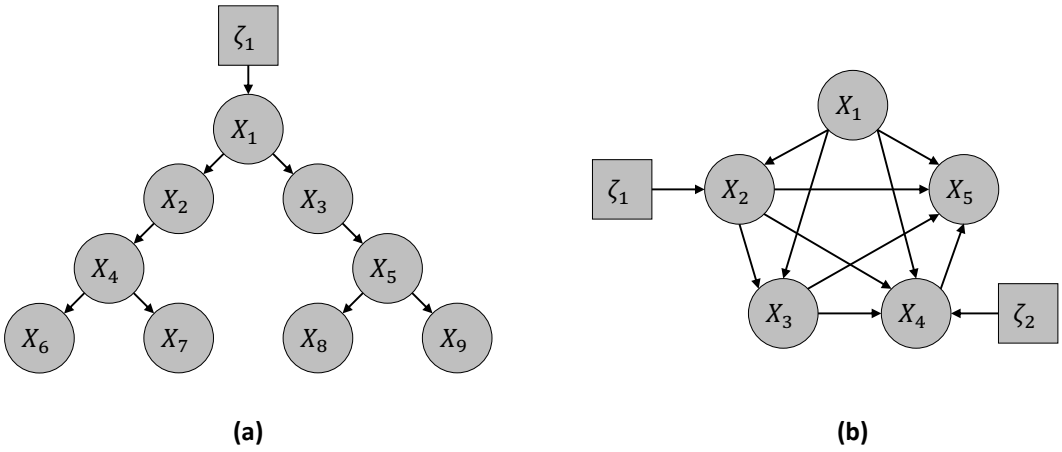


Figure 6.3: Verifying intervention sets for a tree graph and a complete graph.

Theorem 6.1 (Choo et al. (2022))**.** *$\mathcal{I}$ is a verifying intervention set for $\mathcal{G}$ if and only if for every covered*[2] *edge $i \to j$ in $\mathcal{G}$, there exists some $I \in \mathcal{I}$ such that $|I \cap \{i, j\}| = 1$.*

Proof. We will show the necessity. For sufficiency, see Lemma 28 in Choo et al. (2022) for proofs.

Only if direction: suppose $\mathcal{I}$ is a verifying set but there exists a covered edge $i \to j$ such that no $I \in \mathcal{I}$ satisfies $|I \cap \{i, j\}| = 1$. Then it is easy to see that $i \to j$ will not be oriented using Proposition 6.2. In addition $i \to j$ will also not be oriented by any of the four Meek rules in Proposition 6.2. Rule 1: if $i \to j$ is oriented by this, then there is $k \to i$ but k not adjacent to i, which contradicts $i \to j$ being a covered edge. Rule 2: if $i \to j$ is oriented by this, then there is $i \to k$ and $k \to j$, which contradicts $i \to j$ being a covered edge. Rule 3: if $i \to j$ is oriented by this, then there is either a new v-structure $k \to i \leftarrow l$ or $i \to k \to j$ or $i \to l \to j$, which contradicts $i \to j$ being a covered edge. Rule 4: if $i \to j$ is oriented by this, then there is either a cycle $k \to i \to l \to k$ or $i \to k \to j$ or $l \to i \to j$, which contradicts $i \to j$ being a covered edge. □

Example 6.3. *For a tree graph $\mathcal{G}$ with root X_i, the only covered edges are of the form $X_i \to X_j$ for $X_j \in \mathrm{ch}_{\mathcal{G}}(X_i)$. Hence, for a tree graph, an intervention set with only an intervention targeting X_i is a separating set.*

For a complete graph $\mathcal{G}$ with topological order $X_1 < X_2 < \ldots < X_p$, the covered edges are $X_1 \to X_2, X_2 \to X_3, \ldots, X_{p-1} \to X_p$. Thus, if $\mathcal{I}$ is a separating set for $\mathcal{G}$, then for every pair of nodes (X_i, X_{i+1}), one must be intervened. Thus, just even-numbered nodes suffice.

To show the implications on experimental design, consider the minimum verification number. This evolves the concept of bounded interventions: an intervention I of size bounded by k satisfies $|I| \leq k$.

Definition 6.4. *For bounded interventions of size at most k, the* **minimum verification number** *$\nu_k(\mathcal{G})$ denotes the size of the minimum size verifying intervention set.*

Remark 6.1. *Based on Theorem 6.1, we know that $\nu_1(\mathcal{G})$ is equal to the size of the minimum vertex cover of $\mathcal{G}$'s covered edges.*

Remark 6.2 (Verification Number as Lower Bound for Experimental Design)**.** *For bounded interventions of size at most k, any algorithm (adaptive or non-adaptive, randomized or deterministic) would require at least $\nu_k(\mathcal{G})$ interventions to identify $\mathcal{G}$.*

[2] Recall that $i \to j$ is covered edge when $\mathrm{pa}_{\mathcal{G}}(j) = \mathrm{pa}_{\mathcal{G}}(i) \cup \{i\}$

6.3 Adaptive Intervention Design: an Example

In many scenarios, we might consider running experiments *sequentially* instead of all at once. For the t-th intervention, we can make use of the orientations discovered by the $t-1$ interventions, denoted $\mathcal{I}_{t-1}$, already performed to make our choice. Formally, in the noiseless setting where a-priori observational data is assumed, we can think of it using the following definition.

Definition 6.5. *A (noiseless)* **adaptive intervention policy** π *is a set of maps* π_t*, where each* π_t *is a map from previous experiments, which is fully characterizes by a sequence of interventional MECs:* $\mathcal{I}_1$*-MEC, ...,* $\mathcal{I}_{t-1}$*-MEC, to an intervention* I_t*. Here* $\mathcal{I}_s = \{I_1, \ldots, I_s\}$ *for any non-negative integer* s*.*

We now showcase one popular adaptive intervention policy. In general, as the space of possible DAGs is combinatorial and super-exponential in the number of nodes, it is helpful to design interventions that drastically shrink the possible space by a certain ratio. This is very difficult for general DAGs. However, it becomes easier for certain types of DAGs, e.g., tree graphs.

Theorem 6.2 (Greenewald et al. (2019))**.** *Let* $\mathcal{G}$ *be a tree graph with* p *nodes and consider single-node interventions (i.e., bounded interventions with size* 1*). Then there exists an adaptive intervention policy which performs at most* $\lceil \log_2 p \rceil + 1$ *interventions to identify* $\mathcal{G}$*.*

Proof. An intervention on X_i splits a tree into components, with at most one component upstream of X_i (otherwise, we would form an unshielded collider at X_i). By the first Meek rule, all edges downstream of X_i are oriented. Thus, we only need to consider the component that is upstream of X_i. In the worst case, this component is the largest component amongst all components. The size of the largest component is minimized by picking a **central node**, which is guaranteed to split the tree into components with at most $p/2$ nodes. Thus, we reduce the problem size by $1/2$ at each time, ending in at most $\lceil \log_2 p \rceil$ steps. □

In the previous section, we show that the verification number can serve as a lower bound for any policy. One natural question to ask is: *how does my (adaptive) intervention policy compared to this lower bound?* This is essentially a question asking how good is our designed policy. The *competitive ratio* is a quantitative measurement that can be used to answer this question.

Definition 6.6 (Competitive Ratio)**.** *Consider bounded interventions with size at most* k*, the* **competitive ratio** *of any policy* π *w.r.t. benchmark* $\nu_k(\mathcal{G})$ *is defined as* $\frac{T}{\nu_k(\mathcal{G})}$*, where* T *is the smallest step such that* $\mathcal{I}_T$*-MEC only contains* $\mathcal{G}$*.*

Example 6.4. *Consider single-node interventions and tree graph* $\mathcal{G}$*, Theorem 6.2 showed that the central-node policy satisfies* $T \leq \lceil \log_2 p \rceil + 1$*. From the previous section, we know that* $\nu_1(\mathcal{G}) = 1$*. Therefore the competitive ratio of the central-node policy is bounded by* $\lceil \log_2 p \rceil + 1$*.*

In general, it is good to obtain a competitive ratio of $O(\log p)$, as the trivial policy that intervenes on all single nodes would have a competitive ratio of $O(p)$.

Remark 6.3. *In Example 6.4, we showed how to achieve a competitive ratio of* $O(\log p)$ *in tree graphs. It is possible to extend this result to general DAGs Choo et al. (2022).*

Algorithmically, the central-node algorithm acts like a binary search *among the tree-graph space. For general DAGs, such an idea can be extended, where technical details are required to develop such a binary-search-like algorithm. Recent result in Shiragur et al. (2023) shows how to do this, where a competitive ratio of* $O(\log p)$ *is also proven.*

6.4 Additional Remarks

- **Batched or Budgeted Setting:** In this lecture, we mainly considered the case where (1) one intervention is selected at each step, (2) the goal is to minimize the total number of interventions needed

for full-graph identification. Alternatively, one may be interested in the batch-setting, where a set of interventions are selected at each step; see for example Sussex et al. (2021). One might also be interested in the setting where we are given a budget B, and we wish to maximize some measure of learned information about the $\mathcal{I}$-essential graph (e.g., the number of oriented edges) subject to $\text{cost}(\mathcal{I}) \leq B$. See for example Ghassami et al. (2018).

- **Different Objectives:** We considered picking interventions with the end goal of fully orienting the graph. However, this may be overkill for targeted tasks or not sufficient for estimation tasks (which requires estimating the mechanisms in addition to the structure). For example, we may only care about learning a certain sub-structure or some properties (c.f., Agrawal et al. (2019); Shiragur et al. (2023); Zhang et al. (2021)). Instead of learning, we may care about optimization, e.g., maximizing the value of a certain causal variable (e.g., the variable corresponding to the effect), see for example Aglietti et al. (2020). Aside from these active-learning settings where only the last time step matters, we may also care about the accumulated time step in a bandit-like setting (c.f. Simchi-Levi and Wang (2023)).

- **Noisy & Finite-Sample Setting:** So far we have only shown results where infinite samples are assumed after each intervention to decide its effect. Realistically we would need to account for noises coming from finite samples. Many works mentioned under different objectives deal with finite samples. In such settings, one needs to quantify the *uncertainty* of estimating the effect of an intervention. Then the policy to pick the next intervention should account for such uncertainty. Such problems have been extensively studied outside of the causality community, where the interventions are often referred to as arms or actions or data points. When a causal system is considered, most current works deal with the setting where the structure is pre-given but the mechanisms are unknown (c.f., Aglietti et al. (2020); Zhang et al. (2023)).

Bibliography

Aglietti, V., Lu, X., Paleyes, A., and González, J. (2020). Causal bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3155–3164. PMLR.

Agrawal, R., Squires, C., Yang, K., Shanmugam, K., and Uhler, C. (2019). Abcd-strategy: Budgeted experimental design for targeted causal structure discovery. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3400–3409. PMLR.

Choo, D., Shiragur, K., and Bhattacharyya, A. (2022). Verification and search algorithms for causal dags. *arXiv preprint arXiv:2206.15374*.

Ghassami, A., Salehkaleybar, S., Kiyavash, N., and Bareinboim, E. (2018). Budgeted experiment design for causal structure learning. In *International Conference on Machine Learning*, pages 1724–1733. PMLR.

Greenewald, K., Katz, D., Shanmugam, K., Magliacane, S., Kocaoglu, M., Boix Adsera, E., and Bresler, G. (2019). Sample efficient active learning of causal trees. *Advances in Neural Information Processing Systems*, 32.

Hauser, A. and Bühlmann, P. (2014). Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4):926–939.

Meek, C. (1995). Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, page 403–410, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Shiragur, K., Zhang, J., and Uhler, C. (2023). Meek separators and their applications in targeted causal discovery. *arXiv preprint arXiv:2310.20075*.

Simchi-Levi, D. and Wang, C. (2023). Multi-armed bandit experimental design: Online decision-making and adaptive inference. In *International Conference on Artificial Intelligence and Statistics*, pages 3086–3097. PMLR.

Sussex, S., Uhler, C., and Krause, A. (2021). Near-optimal multi-perturbation experimental design for causal structure learning. *Advances in Neural Information Processing Systems*, 34:777–788.

Verma, T. and Pearl, J. (1991). *Equivalence and synthesis of causal models.* UCLA, Computer Science Department.

Yang, K., Katcoff, A., and Uhler, C. (2018). Characterizing and learning equivalence classes of causal dags under interventions. In *International Conference on Machine Learning*, pages 5541–5550. PMLR.

Zhang, J., Cammarata, L., Squires, C., Sapsis, T. P., and Uhler, C. (2023). Active learning for optimal intervention design in causal models. *Nature Machine Intelligence*, 5(10):1066–1075.

Zhang, J., Squires, C., and Uhler, C. (2021). Matching a desired causal state via shift interventions. *Advances in Neural Information Processing Systems*, 34:19923–19934.