Local Discovery by Partitioning

Polynomial-Time Causal Discovery Around Exposure-Outcome Pairs

arXiv

jmaasch

Jacqueline Maasch[1,2,*], **Weishen Pan**[2,3], **Shantanu Gupta**[4], **Volodymyr Kuleshov**[1], **Kyra Gan**[5], **Fei Wang**[2,3]

[1]*Department of Computer Science, Cornell Tech;* [2]*Institute of AI for Digital Health, Weill Cornell Medicine;* [3]*Department of Population Health Sciences, Weill Cornell Medicine;* [4]*Machine Learning Department, Carnegie Mellon University;* [5]*Department of Operations Research and Information Engineering, Cornell Tech;* *`maasch@cs.cornell.edu`

Overview

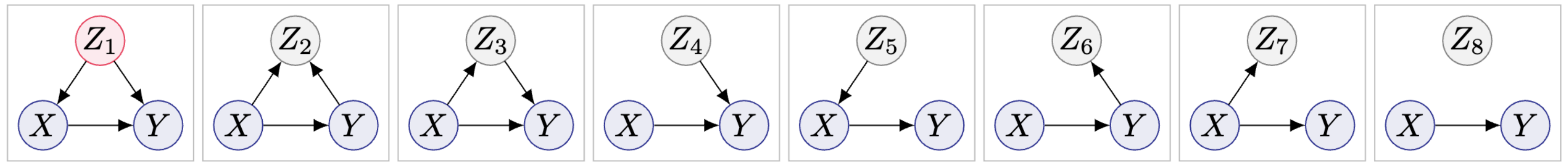
- *Constraint-based confounder discovery:* *Local discovery by partitioning* (LDP) returns a valid adjustment set (VAS) for an exposure X and outcome Y under Pearl's backdoor criterion (Theorem 2).
- *Robust to latent confounding:* If a specific d-separation criterion is passed (Definition 1), the adjustment set is valid under causal insufficiency.
- *Returns partition labels, not graphs:* Instead of learning the causal graph, LDP learns causal partition labels describing the relationship between a given variable, X, and Y (Theorem 1). Partitions are universal properties of arbitrary DAGs.
- *Polynomial time:* Total number of independence tests performed scales quadratically with respect to variable set size (versus worst-case exponential for baselines).
- *Less biased effect estimation:* Adjustment sets from LDP yield less biased and more precise average treatment effect (ATE) estimates than baselines (bottom right).

Universal Property: Causal Partitions

Theorem 1. *Take any arbitrary DAG and a specific exposure X and outcome Y. The eight partitions defined below are exhaustive and disjoint, such that any variable Z falls uniquely under one partition category with respect to $\{X, Y\}$.*

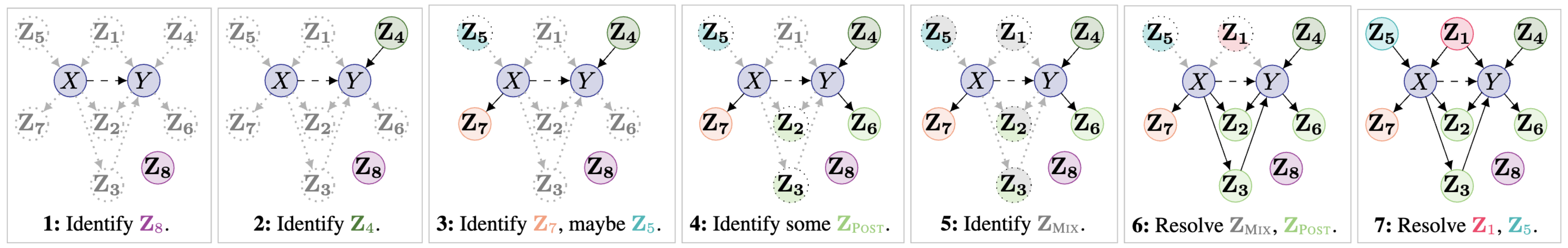
Exhaustive and Mutually Exclusive Causal Partitions

$\mathbf{Z}_1$	*Confounders and their proxies*: Non-descendants of X that lie on an active backdoor path between X and Y (Definition 2.5), and their proxies (Definition B.8).
$\mathbf{Z}_2$	*Colliders and their proxies*: Non-ancestors of $\{X, Y\}$ with at least one active path to X not mediated by Y and at least one active path to Y not mediated by X.
$\mathbf{Z}_3$	*Mediators and their proxies*: Descendants of X that are ancestors of Y, and their proxies (Definition B.8).
$\mathbf{Z}_4$	Non-descendants of Y that are marginally dependent on Y but marginally independent of X (Definition B.3).
$\mathbf{Z}_5$	*Instruments and their proxies*: Non-descendants of X whose causal effect on Y is fully mediated by X, and that share no confounders with Y (Definitions B.1 and B.8).
$\mathbf{Z}_6$	Descendants of Y where all active paths shared with X are mediated by Y.
$\mathbf{Z}_7$	Descendants of X where all active paths shared with Y are mediated by X.
$\mathbf{Z}_8$	All nodes that share no active paths with X nor Y.



Intuition: Partitions generalize the acyclic triples induced by $\{X, Y, Z\}$ to the case of arbitrarily large graphs.

Partitioning for Valid Adjustment Set Discovery



1: Identify $\mathbf{Z}_8$. **2:** Identify $\mathbf{Z}_4$. **3:** Identify $\mathbf{Z}_7$, maybe $\mathbf{Z}_5$. **4:** Identify some $\mathbf{Z}_{\text{POST}}$. **5:** Identify $\mathbf{Z}_{\text{MIX}}$. **6:** Resolve $\mathbf{Z}_{\text{MIX}}$, $\mathbf{Z}_{\text{POST}}$. **7:** Resolve $\mathbf{Z}_1$, $\mathbf{Z}_5$.

Sufficient (not necessary) conditions for correctness: There exists at least one observed member of $\mathbf{Z}_4$ and $\mathbf{Z}_5$.

Robustness to Latent Confounding

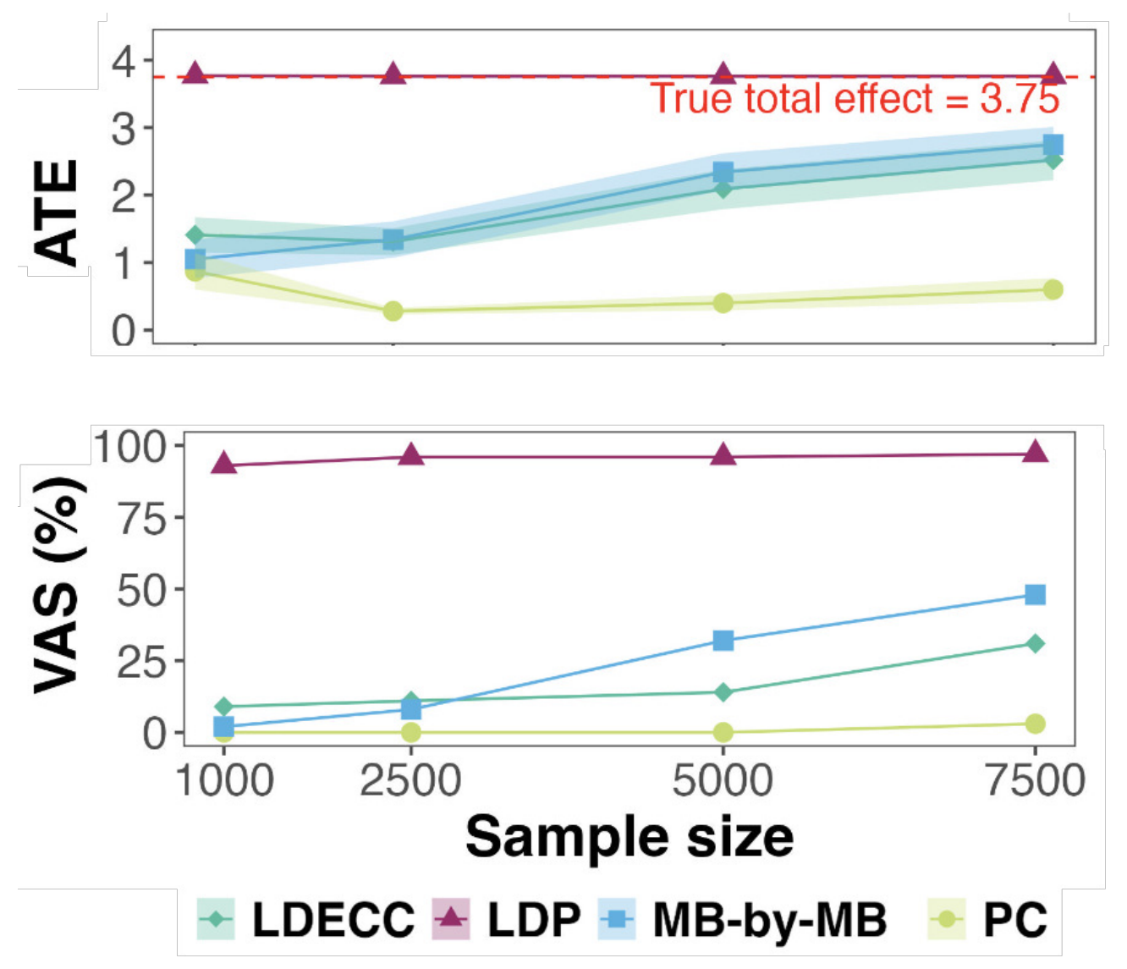
Lemma 1. *LDP does not place descendants of X in $\mathbf{Z}_1$ under sufficient conditions.*

Definition 1 ($\mathbf{Z}_5$ criterion)**.** True if $\exists Z_5 \in \mathbf{Z}_5$ that is d-separable from Y given X and $\mathbf{Z}_1$ ($Z_5 \perp\!\!\!\perp Y | X \cup \mathbf{Z}_1$).

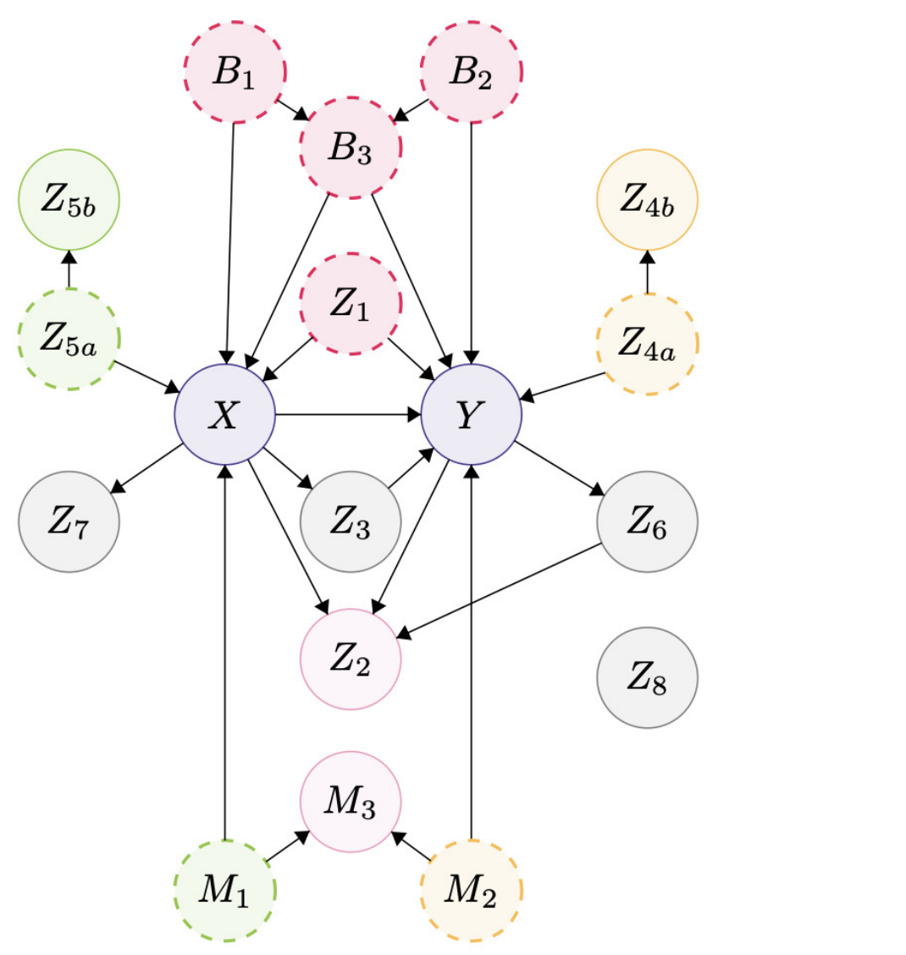
Lemma 2. *Passing the $\mathbf{Z}_5$ criterion is a valid indicator that $\mathbf{Z}_1$ blocks all backdoor paths.*

Theorem 2 (LDP returns a VAS for $\{X, Y\}$ under the backdoor criterion)**.** *Following from Lemmas 1 and 2, if the $\mathbf{Z}_5$ criterion is passed, then the $\mathbf{Z}_1$ returned by LDP is asymptotically guaranteed to be a VAS for $\{X, Y\}$.*

Valid Adjustment Sets for ATE Estimation



Latent	VAS Exists	$\mathbf{Z}_5$ Crit	% Valid
$B_1 \in \mathbf{Z}_1$	✓	✓	100
$B_2 \in \mathbf{Z}_1$	✓	✓	99
$Z_{4a} \in \mathbf{Z}_4$	✓	✓	99
$M_2 \in \mathbf{Z}_4$	✓	✓	100
$Z_{5a} \in \mathbf{Z}_5$	✓	✓	99
$M_1 \in \mathbf{Z}_5$	✓	✓	100
$Z_1 \in \mathbf{Z}_1$	✗	✗	0
$B_3 \in \mathbf{Z}_1$	✗	✗	0



LDP returns a higher proportion of VAS for a ten-node DAG, compared to baselines (left). Results of the $\mathbf{Z}_5$ criterion are consistent with whether a VAS exists in latently confounded variable sets (center, right).

Acknowledgments The authors would like to acknowledge support from the NSF Graduate Research Fellowship (author JM); NSF 1750326, 2212175; and NIH R01AG080991, R01AG076234 for this research.

Presented at the 40th Conference on Uncertainty in Artificial Intelligence (UAI 2024), Barcelona, Spain. Copyright 2024 by the authors.