LOCAL DISCOVERY BY PARTITIONING:
Polynomial-Time Causal Discovery Around Exposure-Outcome Pairs

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

arXiv

jmaasch

CORNELL TECH · Weill Cornell Medicine · Carnegie Mellon University

ABSTRACT

- *Constraint-based causal discovery for covariate selection:* Given an exposure-outcome pair $\{X, Y\}$ and a variable set $\mathbf{Z}$ of unknown causal structure, the *Local Discovery by Partitioning* (LDP) algorithm partitions $\mathbf{Z}$ into subsets defined by their relation to $\{X, Y\}$.
- *Differentiating confounders from other variables:* We enumerate eight exhaustive and mutually exclusive partitions of arbitrary $\mathbf{Z}$ and leverage this taxonomy for discovery.
- *No pretreatment assumption:* LDP does not assume that inputs causally precede the exposure, unlike most methods for automated covariate selection.
- *Asymptotic theoretical guarantees:* LDP returns a valid adjustment set for any $\mathbf{Z}$ under sufficient graphical conditions. Partition labels are asymptotically correct under stronger conditions.
- *Polynomial runtimes:* Total independence tests is worst-case quadratic in $|\mathbf{Z}|$, significantly outperforming constraint-based baselines in experiments.
- *Less biased effect estimation:* Adjustment sets from LDP yield less biased and more precise average treatment effect (ATE) estimates than baselines.

BACKGROUND

- Covariate selection is a central task in the design of observational studies.
- The primary goal of covariate selection is to obtain a *valid adjustment set* for an exposure-outcome pair that eliminates confounding bias by adjusting for confounders [1].
- Data-driven approaches can automate the principled selection of covariates, but these often impose strong graphical assumptions that require prior knowledge (e.g., the pretreatment assumption).
- *In the absence of prior knowledge, does there exist a polynomial-time algorithm that can select covariates in a principled, automated, and causality-based manner with guarantees on correctness?*

PARTITIONS OF Z

This work presents **three main theoretical results**: 1) the existence of eight exhaustive and mutually exclusive partitions that define any arbitrary $\mathbf{Z}$ (Theorem 1); 2) LDP yields asymptotically correct partitions of $\mathbf{Z}$ under sufficient conditions (Theorem 2); and 3) LDP returns valid adjustment sets under weakened sufficient conditions (Theorem 3).

Theorem 1. *Any $\mathbf{Z}$ can be partitioned into eight mutually exclusive subsets (of cardinality greater than or equal to zero) defined solely by their relation to exposure X and outcome Y. Thus, each $Z \in \mathbf{Z}$ uniquely belongs to a single partition defined below.*

EXHAUSTIVE AND MUTUALLY EXCLUSIVE PARTITIONS OF ARBITRARY $\mathbf{Z}$

$\mathbf{Z}_1$	*Confounders*: Non-descendants of X that lie on an active backdoor path between X and Y.
$\mathbf{Z}_2$	*Colliders*: 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*: Descendants of X that are ancestors of Y.
$\mathbf{Z}_4$	Non-descendants of Y that are marginally dependent on Y but marginally independent of X.
$\mathbf{Z}_5$	*Instruments*: Non-descendants of X whose causal effect on Y is fully mediated by X, and that share no confounders with Y.
$\mathbf{Z}_6$	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.

LOCAL DISCOVERY BY PARTITIONING (LDP)

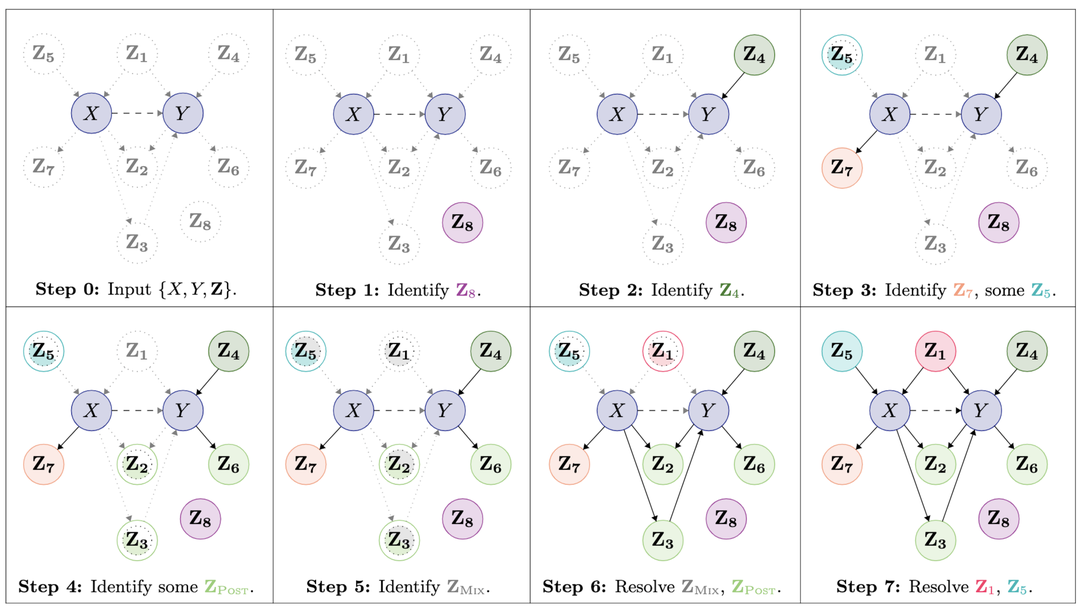

Figure 1: Each step of LDP reveals information about the partitions of $\mathbf{Z}$. The exposure-outcome pair $\{X, Y\}$ serves as a nucleus around which LDP assembles a partial causal graph. Here, each node represents a set corresponding to a single partition of $\mathbf{Z}$, indirect active paths are reduced to length-1 edges, and inter-partition paths are abstracted away.

Sufficient Conditions for Partition Accuracy

C1 The absence of inter-partition active paths that are not fully mediated by $\{X, Y\}$.

C2 The existence of at least one Z_4. Given Condition C1, all Z_2 (if any exist) will be marginally dependent on such a Z_4 and will be identifiable by LDP. This in turn guarantees that all backdoor paths will be blocked by the conditioning set in Step 5 of LDP, which is used to discover $\mathbf{Z}_5$.

C3 Every Z_1 forms a v-structure at X with at least one other variable $Z \in \mathbf{Z}$ ($Z \cdots \rightarrow X \leftarrow \cdots Z_1$) such that $Z \perp\!\!\!\perp Z_1 \wedge Z \not\perp\!\!\!\perp Z_1 | X$. By definition, variable Z can be either in $\mathbf{Z}_5$ or $\mathbf{Z}_1$. Given C1, $\mathbf{Z}_5$ shares no active paths with $\mathbf{Z}_1$ and thus all of $\mathbf{Z}_1$ is marginally independent of $\mathbf{Z}_5$. If $|\mathbf{Z}_5| = 0$, the existence of at least two non-overlapping backdoor paths satisfies this condition.

C4 Causal sufficiency in $\mathbf{Z}$.

Theorem 2 (Correctness of LDP)**.** *Given $\{X, Y, \mathbf{Z}\}$, an independence oracle, and **Conditions C1-C4**, LDP is guaranteed to output a correct partition of $\mathbf{Z}$ that represents the local subgraph surrounding $\{X, Y\}$, where each $Z \in \mathbf{Z}$ is defined solely by its relation to $\{X, Y\}$.*

Theorem 3 (LDP returns valid adjustment sets)**.** *Given $\{X, Y, \mathbf{Z}\}$, an independence oracle, and **Conditions C2-C4**, LDP is guaranteed to return a valid adjustment set.*

Definition 4 (Valid adjustment under the backdoor criterion, [2])**.** Let $\mathbf{A}_{XY}$ be an adjustment set for $\{X, Y\}$ that does not contain $\{X, Y\}$. $\mathbf{A}_{XY}$ is valid if 1) $\mathbf{A}_{XY}$ contains no descendants of X and 2) $\mathbf{A}_{XY}$ blocks all backdoor paths from X to Y.

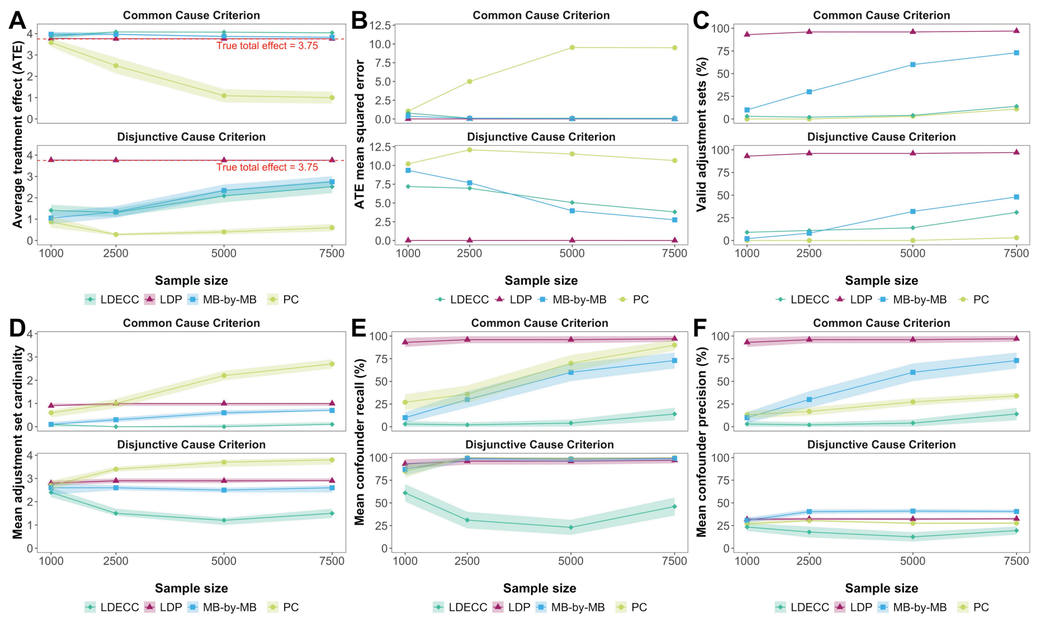

Figure 2: ATE estimation using adjustment sets produced by each baseline for a linear-Gaussian 10-node DAG (Fig. 1). Independence was determined by Fisher-z tests ($\alpha = 0.01$). Results are for 100 replicates per sample size with 95% confidence intervals in shaded regions.

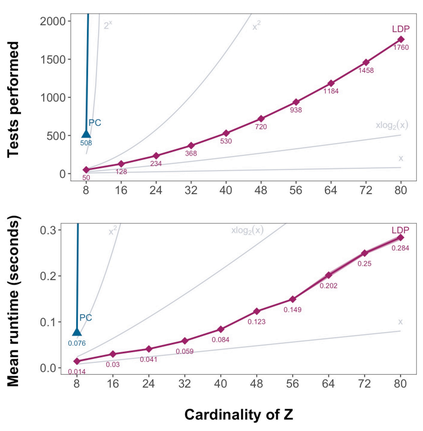

Figure 3: Total tests performed per $\mathbf{Z}$ under an independence oracle (top) and mean runtime over 100 replicates (bottom) as the cardinality of $\mathbf{Z}$ increases, with 95% confidence intervals in shaded regions. Each DAG resembles Figure 1 with equal cardinality per partition ($[1, 10]$).

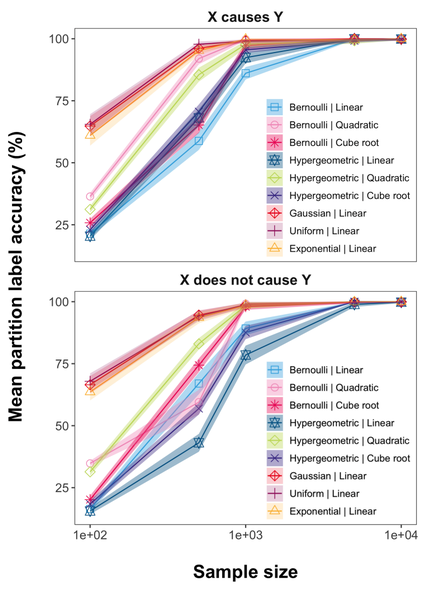

Figure 4: Partition accuracy over 100 replicates of the 10-node DAG (Figure 1), with 95% confidence intervals in shaded regions. Independence was determined by chi-square tests for discrete data and Fisher-z for continuous data ($\alpha = 0.001$).

EMPIRICAL RESULTS

Baseline Methods

1. **PC Algorithm (PC)**, a classic global structure inference algorithm with asymptotic theoretical guarantees [3].
2. **MB-by-MB**, a local Markov blanket learner that infers the local structure around a target node to distinguish parents from children [4].
3. **Local Discovery using Eager Collider Checks (LDECC)**, a local discovery algorithm that leverages unshielded colliders to differentiate the parents of a target from its children [5].

LDP Accurately Partitions Z We measure partition accuracy as the percent of partition labels that are consistent with ground truth. LDP correctly partitions $\mathbf{Z}$ for the 10-node DAG (Fig. 1) under continuous, discrete, linear, and nonlinear data generating processes (Fig. 4).

LDP Enables Less Biased ATE Estimation The ATE was estimated using linear regression for linear-Gaussian 10-node DAGs with a true total effect of 3.75 (Fig. 2). LDP returned the highest quality adjustment sets in terms of ATE mean squared error (MSE), confounder recall, and percent valid. LDP generally produced the least biased ATE estimates and lowest ATE variance, and was the only method to achieve unbiased estimates under the disjunctive cause criterion. Rising MSE for PC may be explained by the cardinality of $\mathbf{A}_{XY}$ increasing with sample size.

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

References

[1] Janine Witte and Vanessa Didelez. Covariate selection strategies for causal inference: Classification and comparison. *Biometrical Journal*, 61(5):1270–1289, September 2019.

[2] Jonas Peters, Dominik Janzing, and Bernhard Scholkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, Cambridge, Massachusetts, 2017.

[3] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*, volume 81 of *Lecture Notes in Statistics*. Springer New York, New York, NY, 2000.

[4] Changzhang Wang, You Zhou, Qiang Zhao, and Zhi Geng. Discovering and orienting the edges connected to a target variable in a DAG via a sequential local learning approach. *Computational Statistics & Data Analysis*, 77:252–266, September 2014.

[5] Shantanu Gupta, David Childers, and Zachary C. Lipton. Local Causal Discovery for Estimating Causal Effects. In *Proceedings of the 2nd Conference on Causal Learning and Reasoning (CLeaR)*. arXiv, February 2023.

PRESENTED AT THE WORKSHOP ON CAUSAL REPRESENTATION LEARNING, 37TH CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NEURIPS 2023), NEW ORLEANS. COPYRIGHT 2023 BY THE AUTHORS.