COMPUTATIONAL AND STATISTICAL GUARANTEES FOR ALGORITHMS ON HIGH-DIMENSIONAL DATA

by

YUHAO WANG

A THESIS SUBMITTED FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

in the

GRADUATE DIVISION

of the

NATIONAL UNIVERSITY OF SINGAPORE

2025

Supervisor:
Associate Professors Arnab Bhattacharyya and Divesh Aggarwal

Examiners:
Associate Professor Jonathan Mark Scarlett
Professor Leong Tze Yun

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Yuhao Wang

31 July 2025

To myself — because this journey was mine to carry.
And to my parents, with the hope that curiosity and lifelong learning find their way into all our lives.

Acknowledgments

The thesis in front of you would not have existed without the help and support of many people. I would like to take this opportunity to thank to them for helping and supporting me to arrive at the end of my PhD journey.

First and foremost, I would like to express my gratitude to my supervisor, Prof. Arnab Bhattacharyya, thanks for leading me to the domain of theoretical computer science. Not only the research. He provided the opportunity and financial support for this doctoral journey. His guidance was valuable in laying the foundation for this work. Especially I want to thank you for giving me all the support and freedom for me to conduct research.

Besides, I would like to sincerely thank Prof Divesh Aggarwal for graciously agreeing to take on my supervision in the final semester, I am very grateful for his support in navigating the final administrative steps of this journey.

I would like to extend my sincere appreciation to my thesis proposal committee members, Prof. Tze Yun and Prof. Jonathan Scarlett. I am grateful for their careful reading of this manuscript and for the insightful comments and thought-provoking questions they provided. Their constructive feedback was invaluable and helped to greatly strengthen the quality and clarity of this dissertation.

Of course, no researcher is an island. Throughout my PhD journey, I have been fortunate to meet and work with many inspiring researchers from around the world. Engaging with the many brilliant minds proved invaluable; the extensive discussions helped me solidify the research directions and formalize the problem statements for my doctoral thesis. I am especially grateful for the opportunity to collaborate and co-author with several of them. I would like to extend my heartfelt thanks to the following individuals, listed alphabetically, whose insights, support, and collaboration have deeply influenced and enriched my research from esteemed faculty members Bryon Aragam, Clément L. Canonne, Constantinos Daskalakis, Sutanu Gayen, Dominik Janzing, Jin Tian, Themis Gouleakis, Vinodchandran Variyam, to talented peers: Davin Choo, Aram Ebtekar, Rishikesh Gajjala, Ming Gao, Wai Ming Tai, and Joy Qiping Yang.

My social interactions throughout this Ph.D. have been an essential source of support, joy, and inspiration. I am deeply grateful to all my lab mates and colleagues

for making this journey so memorable.

Thank you to all the amazing people I met at the School of Computing (SoC) in my first lab after moving to Singapore for showing me the best food spots in the canteens and for all the random chats in the lab: Bingqiao, Hongshi, Naibo, Qian, Qinbin, Shengliang, Shixuan, Xinyu, Yuhang, and Zhaomin. A special thanks to everyone in the AlgoTheory group — Dimitrios, Davin, Eldon, Jiaqi, Joy, Kushagra, Mathews, Naganand, Philips, Sayantan, Sutanu, Themis, Vijeth, Vinh, Vipul, Zeyong, and many others. I'm truly grateful to have been part of such a diverse, vibrant, and active group during my Ph.D. Thank you for the countless research discussions, lunch chats, pizza breaks, coffee moments, snack-sharing sessions, Steam discount alerts, Netflix movie recommendations, soccer banter, cycling rides, and hiking trips — each one made this experience richer.

I also want to thank all my friends around the world who have supported and encouraged me along the way.

Back in Hong Kong, I'm especially grateful to my colleagues and friends: Elmer, Jieshi, Josyl, Ivan, Kanghao, Shawn, Tanuja, Xinyu, and Zhenhui for your research insights, unforgettable food adventures, and countless hikes through the hills.

During my time in Eindhoven, Netherlands at TU/e, I was fortunate to be surrounded by warm, fun, and brilliant colleagues and friends. Special thanks to Anil, Cong, Fang, Ghada, Guangming, Jianpeng, Kaijie, Kelin, Loek, Lu, Marijn, Masoud, Qing, Ricky, Shiwei, Tianjin, Wenjing, Xin, Yulong, and Zhaohuan — for the regular chats, movie nights, hotpot parties, bike trips, bowling events, and cozy home gatherings.

Many thanks to my supervisor, Arnab, for the opportunity to visit the Simons Institute for the Theory of Computing at the University of California, Berkeley, during the first year of my Ph.D. That experience helped me build many of my research collaborations and played a key role in shaping my research direction. I'm also grateful to Professor Kun Zhang for hosting me at Mohamed bin Zayed University of Artificial Intelligence in United Arab Emirates, Abu Dhabi, and for sharing his inspiring vision in research. Thanks as well to Guanglin, Guangwei, Guangyi, Yuewen, Zekang, and many others who supported me when I first arrived — I appreciated the help settling in. Special thanks to Aurélien Bibaut, Alexis Bellot, and Patrick Blöbaum for all the helpful career suggestions during a time when I was

exploring my next steps. Your advice truly made a difference.

In Tübingen, Germany, I had a wonderful time. I'm deeply grateful to my manager Betty Mohler. She's an amazing person who introduced me to the world of industry and helped spark my interest in it. Thanks also to my manager Dominik Janzing, whose insights inspired me both in research and in embracing a healthy lifestyle, not to mention his fascinating travel stories that left a lasting impression. I also want to thank Matthias Poloczek for being a thoughtful stakeholder, and Nefeli Andreou for being such a supportive onboarding buddy.

A big shoutout to Stefan Letz for organizing those lovely Thursday morning breakfast gatherings! It was a joy chatting with brilliant scientists like Elke Kirschbaum, Jonas M. Kübler, Kailash Budhathoki, and Leena Chennuru Vankadara — from lifestyle and career advice to food tips, travel ideas, and skiing recommendations, I learned so much from you all.

I'm also thankful to my talented peers —Anish, Burcu, Filip, Harsha, Haoze, Leello, Philipp, Sergio, Valentyn, Xianghui, Xinyu, and Yiming for all the fun trips, Saturday drinks, and memorable moments we shared. Special thanks to those who encouraged my self-made lunches and salads — you not only motivated me as a researcher, but maybe even as a future chef!

Thank you all for making my Ph.D. journey not just intellectually fulfilling but also filled with friendship, laughter, and unforgettable memories.

I am also grateful to the many other colleagues and researchers and colleagues whose names may not appear here, but whose contributions and interactions have nonetheless shaped my PhD journey. Your insights, questions, and support have enriched my thinking in countless ways, and I remain deeply appreciative.

In the end, I would like to thank my parents. They have always been my greatest supporters and best friends! I'm also deeply grateful to my uncle, Dr. Wang, whose work in space optical technology and scientific outreach has greatly inspired me, and to my uncle, Mr. Zhu, whose dedication to traditional Chinese painting and artistic innovation continues to influence my appreciation for culture and creativity. To all my family members — especially my grandparents — thank you for your unwavering love, encouragement, and support throughout every step of this journey.

Contents

Abstract

Computational and Statistical Guarantees for Algorithms on High-dimensional Data

by

Yuhao Wang

Doctor of Philosophy in Computer Science

National University of Singapore

The importance of computational and statistical guarantees in building reliable decision-making systems is increasingly evident in real-world applications. When working with data, we must ensure that our models are both efficient and reliable, even when the data is biased. This thesis addresses key questions in learning, testing, and inference for high-dimensional data, introducing algorithmic innovations to handle both structured models and biased sampling. The contributions are organized into two themes: (I) Guarantees for unbiased data, ensuring optimal performance in applications such as medical diagnosis and financial risk assessment, and (II) Guarantees for biased data, enabling valid inference in applications such as economics and social sciences, where truncation, censoring, outlier, latent confounding bias challenge traditional methods.

In the first part, we study structured models through three main problems. Firstly, we establish general and rigorous identifiability conditions for linear AnderssonMadiganPerlman (AMP) chain graph models, a broad class that generalizes both linear structural equation models and Gaussian graphical models, even when the chain components are unknown. Secondly, we design statistically optimal algorithms for learning Gaussian trees and polytrees, achieving the best possible rates for both distribution learning (in KL divergence) and exact structure recovery, with finite-sample guarantees and matching lower bounds. Our methods extend the ChowLiu algorithm for trees and adapt the PC algorithm for polytrees. Finally, we develop a mutual information tester with optimal sample complexity for linear models, and apply it to structure learning via the classical ChowLiu algorithm, achieving optimal performance guarantees in this setting.

In the second part, we investigate learning and inference from biased or incomplete data, focusing on how structural constraints, such as truncation, censoring, outliers, and unmeasured confounder bias affect fundamental statistical tasks. Firstly, we study Gaussian mean testing under truncation, where samples from a high-dimensional Gaussian distribution are only observed if they fall within a fixed subset of the domain. We then explore the learning of high-dimensional Gaussians from censored data, where certain coordinates in each sample are missing not at random. Furthermore, we establish quantitative laws governing how outliers propagate through causal mechanisms, offering new insights into model-based anomaly detection. Collectively, these results advance our understanding of how to perform reliable inference when data are incomplete, biased, or shaped by complex causal or observational processes.

List of Figures

List of Tables

List of Notations

Symbol	Description
Vectors and Matrices	
$a \in \mathbb{R}$	Scalar value
$\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$	Deterministic vectors in d-dimensional Euclidean space
$\mathbf{X}, \mathbf{Y} \in \mathbb{R}^d$	Random vectors in d-dimensional Euclidean space
$\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$	Matrix; typically covariance matrix
$\mathbf{I}_d$	Identity matrix of dimension d
A_{ij}	Entry at row i, column j of matrix A
$\langle \mathbf{x}, \mathbf{y} \rangle$	Inner product of vectors $\mathbf{x}$ and $\mathbf{y}$
$\|\cdot\|$	Euclidean (2) norm for vectors
$\|\cdot\|_F$	Frobenius norm for matrices
$\|\cdot\|_{\mathrm{op}}$	Operator (spectral) norm for matrices
$\|\mathbf{x} - \mathbf{y}\|_{\mathbf{\Sigma}}$	Mahalanobis distance with respect to $\mathbf{\Sigma}$
Probability and Distributions	
$\mathbb{E}[X]$	Expectation of random variable X
$\kappa_r(X)$	r-th order cumulant of X
ε	Truncation mass: $\mathbb{P}[(\mathbf{X}, \mathbf{Y}) \notin S]$
α	Accuracy/separation parameter in hypothesis testing
P_X, P_Y, P_{XY}	Marginal and joint distributions
$\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$	Multivariate normal distribution
Information-Theoretic Quantities	
$D_{\mathrm{KL}}(P\|Q)$	Kullback–Leibler divergence between distributions P and Q
$\mathrm{TV}(P, Q)$	Total variation distance between P and Q
$\mathsf{I}(X; Y)$	Mutual information between X and Y

Asymptotic Notation

$\mathcal{O}(\cdot), \Omega(\cdot), \Theta(\cdot)$	Standard asymptotic bounds
$\widetilde{\mathcal{O}}, \widetilde{\Theta}, \widetilde{\Omega}$	Asymptotic bounds up to log factors
$f \lesssim g, f \gtrsim g$	$f \leq Cg$, $f \geq cg$ for constants $C, c > 0$
$A \lesssim B$	$A \leq c \cdot B$ for some absolute constant $c > 0$

Chapter 1

Introduction

1.1 Overview

The analysis of high-dimensional data has become a cornerstone of modern statistics and machine learning, where traditional asymptotic assumptions no longer apply. Foundational work in statistical learning theory, notably through Vapnik's contributions and the development of Probably Approximately Correct (PAC)-style guarantees [Vap99; Val84], established conditions under which generalization is possible despite overparameterization. Subsequent advances in sparse recovery [CT06a; Don06] and high-dimensional estimation [Wai19; DJW18] have introduced structural assumptions, such as sparsity, low-rankness, or manifold constraints, that enable statistical consistency under the asymptotics of dimension. However, these guarantees often presume access to ideal, unbiased data. A more pragmatic challenge arises when this assumption is violated, as classical methods often fail in the presence of truncation, censoring, outliers, and unobserved confounding.

In both the ideal and imperfect data settings, a persistent gap remains between statistical optimality and computational tractability: algorithms that achieve minimax-optimal rates are often computationally infeasible, whereas tractable methods lack rigorous guarantees. This thesis confronts these challenges on two fronts:

1. **Guarantees for Unbiased Data (Part I)**: It addresses the trade-offs between statistical and computational efficiency for inference from unbiased high-dimensional data, focusing on designing methods with optimal, finite-sample guarantees.

2. **Guarantees for Biased Data (Part II)**: It introduces computationally tractable procedures for valid and robust inference in the presence of the common biases, often matching the best possible sample complexity under structural assumptions.

By deriving finite-sample guarantees, characterizing complexity-statistics trade-offs, and leveraging modern optimization theory across these regimes, this work contributes to a unified framework for reliable and scalable inference in realistic high-dimensional settings.

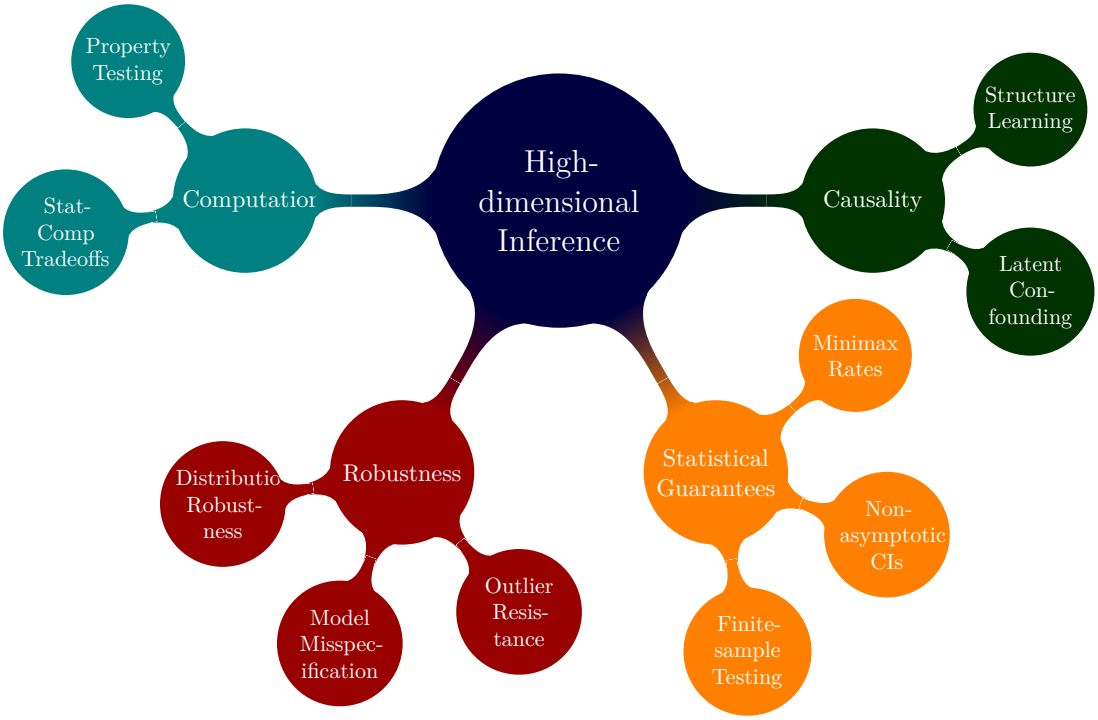


Figure 1.1: Thematic Structure of This Thesis: Domains of High-Dimensional Inference

1.2 Motivation and Problem Statement

The technical challenges inherent in high-dimensional analysis give rise to a fundamental, practical question: how can we be sure that the algorithms we use on complex data will be fast, scalable, and accurate enough for real-world use? The consequences of failing to answer this question are significant. In computational

genomics, for instance, researchers may analyze the expression levels of over 20,000 genes to identify the drivers of a specific cancer. In quantitative finance, algorithms must sift through thousands of market indicators to detect predictive signals for risk management. Similarly, autonomous vehicles rely on processing vast data streams from myriad sensors in real-time to ensure safety. In these high-stakes settings, the risk of mistaking noise for a genuine signal is immense, as uncovering spurious correlations can lead to dire consequences in practice.

To mitigate these risks and ensure the dependability of inference and decision-making systems, this thesis develops algorithmic and statistical frameworks with formal computational and statistical guarantees. The core challenge addressed is the trade-off between statistical rigor (establishing the fundamental limits of learning and inference) and computational feasibility. We develop methods that are either provably optimal in their use of samples or are computationally efficient while providing strong finite-sample guarantees on accuracy and robustness.

1.3 Thesis Outline

This thesis develops algorithmic and statistical frameworks for high-dimensional inference with formal guarantees on performance and efficiency. It is structured around two central themes. The first focuses on inference under unbiased data, where the goal is to design algorithms that achieve optimal statistical accuracy within provable computational limits. The second theme addresses inference under biased or imperfect data, where classical methods often fail due to issues such as truncation, censoring, and outliers.

The progression systematically bridges the gap between foundational statistical theory and the demands of applied high-dimensional analysis. This journey is organized into two complementary parts (Fig. 1.2).

Part I, Learning, Testing, and Inference from Structural Models (Chapter 3 – Chapter 5) establishes a theoretical bedrock. It addresses the fundamental question: Under ideal conditions, what are the absolute limits of statistical inference and how efficiently can we reach them?

- **Chapter 3: Identifiability of Linear AMP Chain Graph Models**

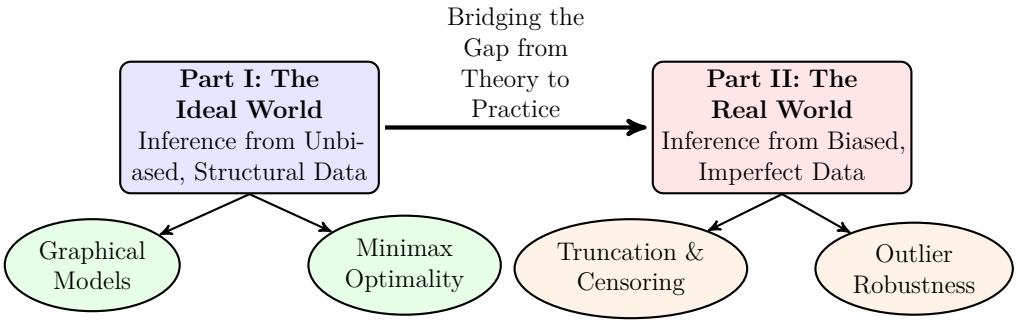


Figure 1.2: Conceptual Flow of the Thesis: From Idealized Models to Robust Inference

[WB21a] investigates rigorous identifiability conditions for linear Andersson-Madigan-Perlman (AMP) chain graph models, a broad class that generalizes linear structural equation models.

- **Chapter 4: Optimal Estimation of Gaussian (Poly) trees** [Wan+24] develops optimal algorithms for learning undirected Gaussian trees and directed Gaussian polytrees, providing finite-sample guarantees and matching lower bounds for both distribution learning and exact structure recovery.
- **Chapter 5: Testing Mutual Information Optimally in Linear Models** introduces a mutual information tester with optimal sample complexity for linear models, which is then applied to structure learning using the classical Chow-Liu algorithm.

Part II, Learning, Testing, and Inference from Biased Data (Chapter 6–Chapter 8) then confronts the complexities of real-world data. It asks: How can we achieve reliable inference when our observations are corrupted by truncation, censoring, or outliers? This part develops robust, tractable methods specifically designed to overcome these common data imperfections.

- **Chapter 6: Gaussian Mean Testing from Truncation** [Can+25] considers the task of testing the mean of a high-dimensional Gaussian distribution when observations are only revealed if they fall within a fixed subset of the domain.
- **Chapter 7: Learning High-dimensional Gaussians from Censored Data** [Bha+25] provides efficient algorithms for distribution learning from

high-dimensional Gaussian data where variables are missing not at random (MNAR).

- **Chapter 8: Toward Universal Laws of Outlier Propagation** [EWJ25] establishes quantitative laws governing how outliers propagate through causal mechanisms, using Algorithmic Information Theory (AIT) to formalize root-cause analysis.

1.4 Summary of Contributions

This thesis develops a unified framework for high-dimensional inference, addressing the critical tension between optimal statistical guarantees and real-world data imperfections. The work systematically bridges foundational theory with practical methodology.

- **Part I (The Theoretical Bedrock)**: We first establish principles for learning from structural models, providing the first general and rigorous identifiability conditions for a broad class of chain graphs. We then develop algorithms for learning Gaussian polytrees with provably minimax-optimal sample complexity.
- **Part II (Robust Inference)**: Building on this foundation, we introduce a suite of robust methods to address common data imperfections. These contributions include the first sharp analysis for statistical testing under truncation, novel algorithms for learning from self-censoring, and a principled theoretical framework that uses Algorithmic Information Theory (AIT) to establish universal laws of anomaly propagation.

1.5 Limitations and Future Directions

No single thesis can solve all challenges in this vast field. The following areas represent the primary limitations of the current work and point to fruitful avenues for future research, which are discussed in detail in Chapter 9: Conclusion and Future Work.

- **Model Assumptions**: Much of the analysis, particularly in Part I, relies on linear or Gaussian assumptions. Extending these optimality guarantees to more complex, non-parametric model families is a significant next step.
- **Computational vs. Practical Scalability**: While the algorithms developed are polynomial-time and often optimal in a theoretical sense, their practical implementation for web-scale datasets (trillions of parameters) may require furthe distributed computation strategies.
- **Known Bias Models**: The methods in Part II assume the nature of the data bias (e.g., the truncation set, the censoring mechanism) is known. Developing methods that can simultaneously learn the model parameters and the nature of the bias is a challenging but critical open problem.

Chapter 2

Preliminaries

This chapter outline the key assumptions used throughout the thesis, grouped by problem setting. These assumptions are critical for guaranteeing statistical identifiability, sample-efficient estimation, and robustness in high-dimensional inference.

2.1 General Notation

This section summarizes the general mathematical notation used throughout this thesis. Scalars are denoted by lowercase letters (e.g., $a \in \mathbb{R}$), vectors by lowercase boldface letters (e.g., $\mathbf{x} \in \mathbb{R}^p$), and matrices by uppercase boldface letters (e.g., $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$). For matrices, A_{ij} denotes the (i, j)-th entry. We adopt standard conventions for matrix norms, including the Euclidean norm $\|\cdot\|$, Frobenius norm $\|\cdot\|_F$, and operator norm $\|\cdot\|_{\text{op}}$. Information-theoretic measures such as Kullback-Leibler divergence $D_{\text{KL}}(P\|Q)$ and total variation distance $D_{\text{TV}}(P, Q)$ are used throughout. Asymptotic behavior is expressed using standard notation $\mathcal{O}(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$, with $\tilde{\mathcal{O}}$ denoting bounds up to logarithmic factors. Inequalities such as $f \lesssim g$ and $f \gtrsim g$ indicate comparisons up to absolute constants. These conventions are followed consistently across all chapters to ensure clarity and uniformity in presentation.

2.2 Probability and Matrix Algebra

We begin by defining the core probability divergence metrics used throughout this thesis to quantify the distance between distributions.

Probability Divergence Metrics

Definition 2.2.1 (Kullback–Leibler Divergence)**.** *Let P and Q be two probability distributions on $\mathcal{X}$ such that $P \ll Q$ (i.e., P is absolutely continuous with respect to Q). The Kullback–Leibler (KL) divergence from P to Q is defined as*

$$D_{\mathrm{KL}}(P \,\|\, Q) := \int_{\mathcal{X}} \log\left(\frac{dP}{dQ}(x)\right) dP(x),$$

where $\frac{dP}{dQ}$ is the Radon–Nikodym derivative of P with respect to Q.

Definition 2.2.2 (Total Variation Distance)**.** *Let P and Q be two probability measures defined on the same measurable space $\mathcal{X}$. The total variation (TV) distance between P and Q is defined as*

$$D_{\mathrm{TV}}(P, Q) := \sup_{A \subseteq \mathcal{X}} |P(A) - Q(A)| = \frac{1}{2} \int_{\mathcal{X}} |p(x) - q(x)|\, dx,$$

where p and q denote the density functions of P and Q, respectively.

Pinskers inequality provides a fundamental connection between KL divergence and TV distance: it guarantees that if two distributions are close in KL divergence, then they are also close in total variation. However, the converse is not generally true, as KL divergence is not symmetric and can be infinite even when TV distance is small.

Proposition 2.2.3 (Pinskers Inequality)**.** *For any two probability distributions P and Q, the total variation distance is bounded in terms of the KL divergence as*

$$D_{\mathrm{TV}}(P, Q) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P \,\|\, Q)}.$$

Next, we introduce several useful notions of distances and norms necessary for formulating our guarantees.

Definition 2.2.4 (Mahalanobis Distance)**.** *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ be vectors and let $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$ be a positive definite matrix. The Mahalanobis distance between $\mathbf{x}$ and $\mathbf{y}$ with respect to $\mathbf{\Sigma}$ is defined as*

$$\|\mathbf{x} - \mathbf{y}\|_{\mathbf{\Sigma}} := \sqrt{(\mathbf{x} - \mathbf{y})^\top \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{y})}.$$

Definition 2.2.5 (ℓ_2 Norm (Euclidean Norm))**.** *Let $\mathbf{x} \in \mathbb{R}^p$ be a vector. The ℓ_2 norm (or Euclidean norm) of $\mathbf{x}$ is defined as*

$$\|\mathbf{x}\| := \sqrt{\sum_{i=1}^{p} x_i^2} = \sqrt{\mathbf{x}^\top \mathbf{x}}.$$

Definition 2.2.6 (Operator Norm (Spectral Norm))**.** *Let $A \in \mathbb{R}^{m\times n}$ be a matrix. The operator norm (also called the spectral norm) of A, induced by the Euclidean norm, is defined as*

$$\|A\|_{\text{op}} := \sup_{\mathbf{x}\in\mathbb{R}^n,\, \|\mathbf{x}\|=1} \|A\mathbf{x}\| = \sigma_{\max}(A),$$

where $\sigma_{\max}(A)$ denotes the largest singular value of A.

Definition 2.2.7 (Frobenius Norm)**.** *Let $A \in \mathbb{R}^{m\times n}$ be a matrix with entries a_{ij}. The Frobenius norm of A is defined as*

$$\|A\|_F := \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2}.$$

Probability The analysis of graphical models in Chapter 3 relies on key probabilistic identities, including the following generalized laws for covariance and entropy.

Fact 2.2.8 (Law of Conditional Covariance)**.** *If X, Y, Z are random variables with strictly positive distributions with each component having finite second moment, then:*

$$\operatorname*{Cov}_X(X \mid Y) = \mathop{\mathbb{E}}_Z[\operatorname*{Cov}_X(X \mid Y, Z) \mid Y] + \operatorname*{Cov}_Z(\mathop{\mathbb{E}}_X[X \mid Y, Z] \mid Y).$$

The following result yields a very useful decomposition for covariance of normal distributions.

Fact 2.2.9. *If $X = (X_A, X_B)$ is distributed jointly as a Gaussian $\mathcal{N}(\mathbf{0}, \Sigma)$, then:*

$$\det(\text{Cov}(X)) = \det(\text{Cov}(X_A)) \cdot \det(\text{Cov}(X_B \mid X_A))$$

where $\text{Cov}(X_B \mid X_A) = \text{Cov}(X_B \mid X_A = x_A)$ is independent of x_A.

Matrix Algebra. Our identifiability condition for chain graph models in Chapter 3 relies on generalized matrix properties. We define the properties required for families of positive semidefinite matrix functions below.

Definition 2.2.10. *Let $\mathbb{C}_n$ denote the cone of $n \times n$ positive semidefinite matrices. We say that a real-valued function $d_n : \mathbb{C}_n \to \mathbb{R}$ is* positive and super-additive *if: (i) $d_n(A) > 0$ for all positive definite matrices A, and (ii) for all positive semidefinite matrices A, B:*

$$d_n(A + B) \geq d_n(A) + d_n(B).$$

A positive and super-additive family *is a collection of functions $f_n : \mathbb{C}_n \to \mathbb{R}$, each of which is positive and super-additive.*

We have several examples of families of positive and super-additive functions:

- Clearly, the projection on any diagonal element and the matrix trace function are positive and super-additive.
- By Minkowski's determinant theorem (see, e.g., [MM92]), it known that for all $A, B \in \mathbb{C}_n$: $(\det(A + B))^{1/n} \geq (\det(A))^{1/n} + (\det(B))^{1/n}$. Hence, $\{\det^{1/n} : \mathbb{C}_n \to \mathbb{R}\}$ is positive and super-additive.
- For χ an irreducible character on a subgroup H of S_n (the permutation group on n elements), define the *generalized matrix function* with respect to H and χ as:

  $$d_\chi^H(A) = \sum_{\sigma \in H} \chi(\sigma) \prod_{i=1}^{n} a_{i,\sigma(i)}$$

  where $A = (a_{i,j})$. [Sch18] showed that $d_\chi^H(A) > 0$ for all positive definite A. It is also known (e.g., [Mer97], p. 228) that they satisfy the super-additivity condition. Hence, the determinant[1], permanent, and the Hadamard matrix function (product of diagonal entries) all form positive and super-additive families.

2.3 Graphical Model Preliminaries

This section introduces the core concepts and notation for graphical models, including Directed Acyclic Graphs (DAGs), Causal Bayesian Networks (CBNs), and

[1]The super-additivity of the determinant is also directly implied by the super-additivity of $\det^{1/n}$.

Chain Graphs. These concepts are foundational, providing the necessary mathematical structure for the entire thesis, from establishing minimax optimality in Part I (Chapter 3–Chapter 5) to modeling outlier propagation in Part II (Chapter 8).

Graphical Definitions For a directed acyclic graph (DAG) $G = (V, E)$, for each node $k \in V$, $\mathrm{pa}(k) = \{j : (j,k) \in E\}$ denotes its parent nodes, descendants $\mathrm{de}(k)$ denotes the nodes that can be reached by k and $\mathrm{nd}(k) = V \setminus \mathrm{de}(k)$ denotes the nondescendants. The skeleton of G, $\mathrm{sk}(G)$, is the undirected graph formed by removing directions of all the edges in G. For any $j, \ell, k \in V$, a triple (j, ℓ, k) is called unshielded if both j, k are adjacent to ℓ but not adjacent to each other, graphically $j - \ell - k$; and is called a v-structure if additionally j, k are parents of ℓ, i.e. $j \to \ell \leftarrow k$. The in-degree of G is $\max_k |\mathrm{pa}(k)|$. A tree is an undirected graph in which any two nodes are connected by exactly one path. A directed tree is a directed graph in which, for some root node u, and any other node v, there is exactly one directed path from u to v. A polytree is a directed graph whose skeleton to be a tree. Denote the set of directed trees (resp. polytrees) over d nodes to be $\mathcal{T}$ (resp. $\widetilde{\mathcal{T}}$). Note that a directed tree is a polytree with in-degree equal to one except the root node who has no parent and $\mathcal{T} \subseteq \widetilde{\mathcal{T}}$.

These structural definitions allow us to introduce the primary graphical models considered in this thesis, starting with the causal Bayesian network.

Definition 2.3.1 (Causal Bayesian Network (CBN))**.** *A causal Bayesian network over variables $X_1, \ldots, X_n$ is defined by a directed acyclic graph (DAG) $G = (V, E)$ and a joint distribution $P(X_1, \ldots, X_n)$ that factorizes as*

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \mathrm{PA}_i), \tag{2.1}$$

where PA_i denotes the parents of X_i in G. Each conditional $P(X_i \mid \mathrm{PA}_i)$ represents an independent causal mechanism.

Gaussian Bayesian Networks In many sections of this thesis, we assume that the joint distribution P is a Gaussian distribution, i.e., $P = \mathcal{N}(0, \Sigma)$. When P is Gaussian, the general Bayesian network factorization (Eq. (2.1)) implies that X

can always be expressed as the following linear structural equation model (SEM):

$$X_k = \beta_k^\top X + \eta_k\,, \quad \eta_k \sim \mathbb{N}(0, \sigma_k^2), \tag{2.2}$$

where $\beta_k \in \mathbb{R}^d$ is supported on $\mathrm{pa}(k)$ and the $\{\eta_k\}_{k=1}^d$ are mutually independent. A Gaussian distribution is said to be T-structured for some directed tree $T \in \mathcal{T}$ (or simply tree-structured when the specific T is not important in the context) if it satisfies Eq. (2.1) with respect to some tree T. For a distribution P and a directed tree T, let

$$P_T := \operatorname*{arg\,min}_{T\text{-structured distribution } Q} D_{\mathrm{KL}}(P \parallel Q),$$

where $D_{\mathrm{KL}}(\cdot \parallel \cdot)$ denotes the KL-divergence. In this paper, we consider both general Gaussians (non-realizable case) as well as tree-structured distributions (realizable and faithful cases), i.e. Eq. (2.2) holds for some directed (poly)tree T.

Faithfulness and Markov Equivalence Class For the purpose of structure learning, a common assumption is faithfulness, under which the DAG is identified up to its Markov equivalence class (MEC). This concept is fundamental to the structure learning algorithms analyzed in Chapter 4 and Chapter 5. The reader is directed to standard references for foundational graphical concepts, such as d-separation; see [KF09a] for more background.

Definition 2.3.2 (Faithfulness)**.** *We say a distribution P is faithful to a DAG G if for any $j, k \in V$ and $S \subseteq V \setminus \{j, k\}$,*

$$X_j \perp\!\!\!\perp X_k \,|\, X_S \Rightarrow j \text{ and } k \text{ are } d\text{-separated by } S\,.$$

Equivalently, for any two nodes j and k not d-separated by set S, faithfulness requires $X_j \not\perp\!\!\!\perp X_k \,|\, X_S$. The MEC of a DAG G is the set of DAGs that encode the same set of conditional independencies as G, which is usually represented by a CPDAG, denoted by $\overline{G}$. A standard approach to learning a CPDAG under faithfulness is the PC algorithm [SG91], which relies on conditional independence testing to recover the skeleton and orient the edges. While faithfulness can be a strong assumption [Uhl+13], it is known that weaker assumptions suffice. For example:

Definition 2.3.3 (Restricted faithfulness)**.** *We say a distribution P is restricted faithful to a DAG G if*

1. *For any* $(j,k) \in E$, $S \subseteq V \setminus \{j,k\}$, $X_j \not\perp\!\!\!\perp X_k \,|\, X_S$;

2. *For any unshielded triple* $j - \ell - k$, *if this is a v-structure, then* $X_j \not\perp\!\!\!\perp X_k \,|\, S$ *for any* $S \subseteq V \setminus \{j,k\}$ *with* $\ell \in S$; *if not, then* $X_j \not\perp\!\!\!\perp X_k \,|\, X_S$ *for any* $S \subseteq V \setminus \{j,k,\ell\}$.

Under general faithfulness, all conditional independence relationships imply d-separations in a DAG. In other words, all instances of d-connections lead to conditional dependence. On the contrary, restricted faithfulness requires only a subset of d-connections to imply conditional dependence. Conventionally, the first part of Definition 2.3.3 is also named *adjacency-faithfulness* and the second part is named *orientation-faithfulness*. With our focus on the setup where the underlying DAG is a polytree, restricted faithfulness can be further relaxed as we will discuss in Section 4.3.

Chain Graphs. This model is central to Chapter 3, where we study the identifiability of linear AMP Chain Graph models. A chain graph is a specialized model that may contain both directed ('$\to$') and undirected ('—') edges. For a general account of chain graph models, we refer the reader to [Lau96] and [Edw12].

A chain graph $\mathcal{C}$ consists of a *vertex set* V and an *edge set* $E \in V \times V$. Two vertices joined by an edge are called *adjacent*. A *path* in $\mathcal{C}$ is a sequence of distinct vertices $\langle v_0, \ldots, v_n \rangle$ such that v_{i-1} and v_i are adjacent for all $1 \leq i \leq n$, and is called a *cycle* if $v_n = v_0$. Moreover, a *semi-directed cycle* exists if $v_1 \to v_2$ is in $\mathcal{C}$ and $v_i \to v_{i+1}$, $v_i \longleftrightarrow v_{i+1}$ or $v_i - v_{i+1}$ is in $\mathcal{C}$ for all $1 < i < n$. For any subset S, the set of parents of v is denoted as $\mathrm{Pa}(v) := \{v \in V \setminus S \mid v \to s \in \mathcal{C} \text{ for some } s \in S\}$, the set of *children* of v is denoted as $\mathrm{Ch}(v) := \{v \in V \setminus S \mid s \to v \in \mathcal{C} \text{ for some } s \in S\}$, the set of *neighbours* is denoted as $\mathrm{Ne}(v) := \{v \in V \setminus S \mid v - s \in \mathcal{C} \text{ for some } s \in S\}$. A chain graph with no directed edges is an undirected graph (UG), while a chain graph with no undirected edges is a DAG. For vertices $(u,v) \in E$ but $(v,u) \notin E$, we write $u \to v$, where vertex u is a *parent* of v. If both $(u,v) \in E$ and $(v,u) \in E$, we denote it by $u - v$, which means u is a *neighbor* of v. The vertex set of a chain graph can be partitioned into *chain components* $\{\tau \mid \tau \in \mathcal{T}\}, (V = \cup_{(\tau \in \mathcal{T})} \tau)$. Edges within chain components are undirected whereas edges between two chain components are directed. A *source* node is any node X_τ such that $\mathrm{Pa}(X_\tau) = \emptyset$. A *sink* node is any

node X_τ such that $\mathrm{Ch}(X_\tau) = \emptyset$. The chain components τ of a chain graph are the connected components of the undirected graph obtained by removing all directed edges from the chain graph. In a DAG, all chain components are singletons. For $S \subseteq V$, $\mathcal{C}_S$ denotes the induced subgraph on S.

By taking into account the directed connections of chain components, an AMP chain graph admits a topological ordering of its chain components. For statistical identifiability of chain graph $\mathcal{C}$, we will consider it sufficient to learn the partition into chain components $\tau_1, \ldots, \tau_t$, and a topological ordering $\prec$ such that $\tau_j \to \tau_k \implies \tau_j \prec \tau_k$. One can learn the directed and undirected edges using standard parameter estimation algorithms.

2.4 Models of Data Imperfection and Bias

In Part II of this thesis (Chapter 6 to Chapter 8), we shift focus from idealized learning settings to realistic ones, where data may be compromised. This section introduces the precise mathematical models that describe truncation, censoring, and outlier generation.

2.4.1 Truncation

The truncation model is the focus of Chapter 6, where we study Gaussian mean testing.

Truncated Gaussian Distribution. Let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ represent the normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, whose probability density function is given by:

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{x}) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{\mu}\|_{\boldsymbol{\Sigma}}^2\right).$$

We denote the truncated normal distribution restricted to a set S as $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)$, with the probability mass of S under this distribution written as $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; S)$. The corresponding probability density function is:

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S; \boldsymbol{x}) = \begin{cases} \frac{1}{\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; S)} \cdot \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{x}) & \boldsymbol{x} \in S \\ 0 & \boldsymbol{x} \notin S \end{cases}.$$

We can then write the population negative log-likelihood $\bar{\ell}(\cdot)$ for data coming from a truncated normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{I}_d$ as:

$$\bar{\ell}(\boldsymbol{v}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{I}_d,S)}\left[\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x} - \boldsymbol{v}^T\boldsymbol{x}\right] + \log\left(\int_S \exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{z} + \boldsymbol{v}^T\boldsymbol{z}\right)d\boldsymbol{z}\right). \tag{2.3}$$

In Chapter 6, we focus on spherical Gaussian (covariance matrix is $\boldsymbol{I}_d$), so our log-likelihood function only has one parameter. We can write the gradient of the negative log-likelihood function $\bar{\ell}$ as with respect to $\boldsymbol{v}$ ($\nabla\bar{l}(\boldsymbol{v})$) as follows:

$$\frac{\partial\bar{l}(\boldsymbol{v})}{\partial\boldsymbol{v}} = -\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{I}_d,S)}[\boldsymbol{x}] + \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{v},\boldsymbol{I}_d,S)}[\boldsymbol{z}] \tag{2.4}$$

Throughout Chapter 6, we will use S to indicate the support of P after truncation and $\boldsymbol{\mu}_S = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{I}_d,S)}[\boldsymbol{x}]$ the truncated mean of some multivariate normal P (or the mean under truncation).

The following lemma, adapted from established work on truncated statistics [Das+18], establishes a strong convexity property crucial for proving the convergence of our estimation algorithms in Chapter 6.

Lemma 1 (Strong convexity with truncation adapted [Das+18, Lemma 4])**.** *Let $\boldsymbol{H}_\ell$ be the Hessian of the negative log likelihood function $\bar{\ell}(\boldsymbol{v})$, with the presence of arbitrary truncation S such that $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{I}_d;S) \geqslant \beta$ for some $\beta \in (0,1]$. Then it holds that*

$$\boldsymbol{H}_\ell(\boldsymbol{v}) \succeq \frac{1}{2^{13}}\left(\frac{\beta}{C}\right)^4 \cdot \min\left\{\frac{1}{4}, \frac{1}{16\|\boldsymbol{\mu}\|_2^2+1}\right\}\cdot\boldsymbol{I}_d,$$

where C is a universal constant.

Let $\mathcal{D} = \{P_v | v \in \mathbb{S}_d\}$ denote a family of distributions constructed in the following manner: Fix a one dimensional distribution A and pick a unit d-dimensional vector $v \in \mathbb{S}_d$ uniformly at random. P_v is a copy of A in the direction of v and standard normal in directions orthogonal to v.

Proposition 2.4.1 (Sample complexity lower bound for high-dimensional testing [DKS16, Proposition 7.1])**.** *Let A be a distribution on $\mathbb{R}$ such that A has mean 0 and $\chi^2(A,\mathcal{N}(0,1))$ is finite. Then, there is no algorithm that, for any d, given $N < d/(8\chi^2(A,\mathcal{N}(0,1)))$ samples from a distribution D over $\mathbb{R}^n$ which is either $\mathcal{N}(0,1)$ or $P_v \in \mathcal{D}$, correctly distinguihes between the two cases with probability $2/3$.*

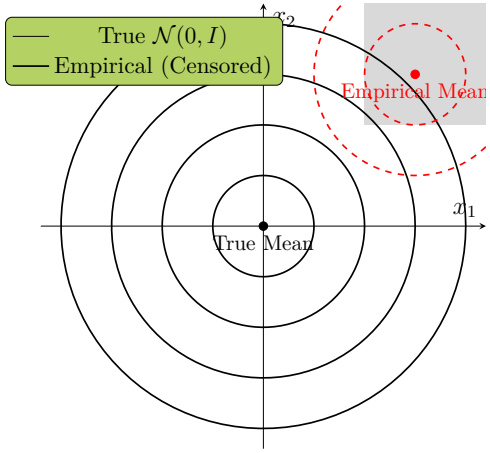


Figure 2.1: The effect of censoring on 2d Gaussian

2.4.2 Censoring

The censoring models defined in this section are the focus of Chapter 7, where we develop algorithms for learning high-dimensional Gaussians from censored data. Fig. 2.1 illustrates the effect of censoring on an underlying $\mathcal{N}(0, I)$ distribution. When the true sample distribution is restricted to an observed region (shaded area), the resulting observed data is biased. Consequently, the Empirical Mean (red dot) is systematically shifted away from the True Mean (origin), highlighting the primary challenge in estimating the true distribution parameters from censored data.

Definition 2.4.2 (Censored Gaussian Model)**.** *Let $Y \sim \mathcal{N}(\mu^*, \Sigma^*)$ be a multivariate Gaussian in $\mathbb{R}^d$ with unknown mean $\mu^* \in \mathbb{R}^d$ and unknown covariance matrix $\Sigma^* \in \mathbb{R}^{d\times d}$. Instead of observing full samples $y \in \mathbb{R}^d$, one observes censored samples represented by pairs (A, x), where $A \subseteq [d]$ is the set of observed coordinates and $x = y_A$ is the observed subvector. The censoring mechanism is governed by a known function $S : \mathbb{R}^d \to 2^{[d]}$, such that $A = S(y)$.*

Definition 2.4.3 (Missingness Model)**.** *A missingness model defines the probabilistic mechanism by which subsets of coordinates in $y \sim \mathcal{N}(\mu^*, \Sigma^*)$ are observed. Formally, the observed index set $A = S(y) \subseteq [d]$ is determined by a stochastic or deterministic function of y. The observed data consist only of $x = y_A$, the projection of y onto the coordinates in A.*

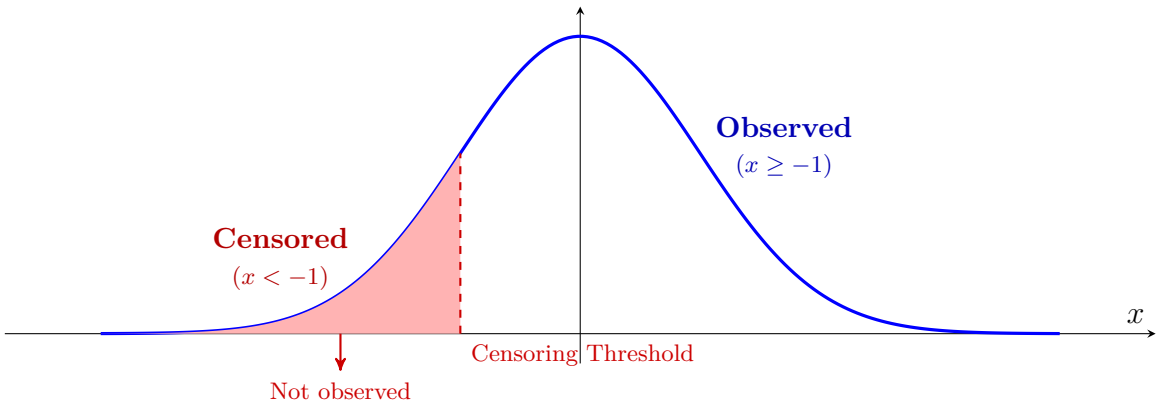


Figure 2.2: Self-censoring on 1d Gaussian

This concept is further illustrated in a simpler one-dimensional setting in Fig. 2.2, which depicts a self-censoring model. Here, a portion of the distribution (shaded red, $x < 1$) falls below a specific threshold and is not observed, while the remainder ($x \geq 1$) is observed. This visualizes how the observation status of a variable depends directly on its own value.

Definition 2.4.4 (Self-Censoring Model)**.** *In the self-censoring model, each coordinate $i \in [d]$ is observed or censored based on whether $y \in S_i$, where $S_i \subseteq \mathbb{R}^d$ is a known subset of the input space. The observation condition for coordinate i is:*

$$x_i = y_i \quad \textit{if and only if} \quad y \in S_i.$$

Assumption: For every pair $i, j \in [d]$, the joint observation probability $\mathbb{P}[i, j \in S(y)] \geq \alpha$ for some constant $\alpha > 0$.

The complexity increases with the linear-thresholding model, which is illustrated in Fig. 2.3. In this two-dimensional example, the resulting observed region is defined by the intersection of multiple linear constraints ($v1^T y = b1$ and $v2^T y <= b2$). This scenario highlights how censoring a variable depends on linear combinations of the entire vector, necessitating the following more complex definition.

Definition 2.4.5 (Linear-Thresholding Model)**.** *In the linear-thresholding model, each coordinate is observed if a linear inequality is satisfied. Formally, there exist known vectors $v_i \in \mathbb{R}^d$ and thresholds $b_i \in \mathbb{R}$ such that:*

$$S(y) = \{i \in [d] \mid v_i^\top y \leq b_i\}.$$

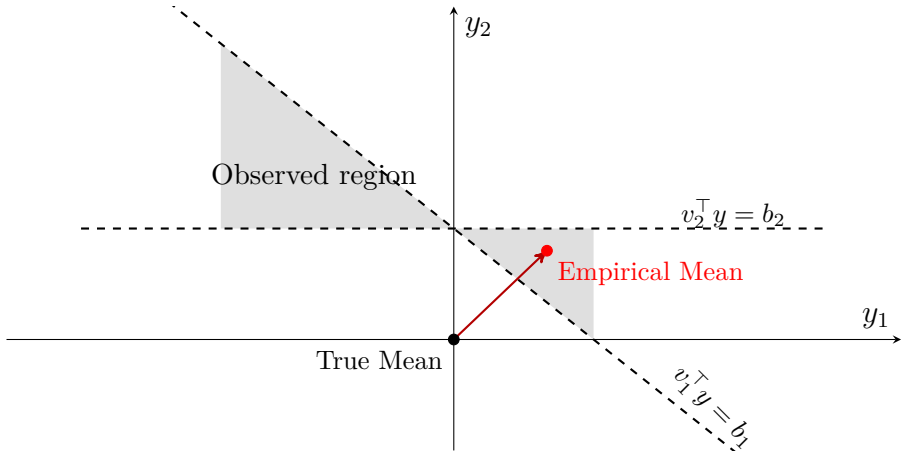


Figure 2.3: Two-dimensional linear-thresholding censoring model

This model allows for more structured and dependent forms of censoring, as the observation set A depends on linear projections of the entire vector y.

The following assumptions is for the learning guarantees under the linear-thresholding model.

Assumption 2.4.6 (Anchor Set)**.** *There exists a fixed subset $\mathcal{A} \subseteq [d]$ of coordinates such that $\mathcal{A} \subseteq S(y)$ for all $y \in \mathbb{R}^d$. That is, the coordinates in $\mathcal{A}$ are always observed.*

2.4.3 Outlier

This section introduces concepts from Algorithmic Information Theory (AIT), which provides a principled, mechanism-based framework for characterizing outliers in Chapter 8. We start with the fundamental definitions of complexity and deficiency.

Definition 2.4.7 (Kolmogorov Complexity)**.** *The* prefix Kolmogorov complexity *$K(x)$ of a finite object x is the length of the shortest prefix-free program that outputs x on a fixed universal Turing machine.*

Definition 2.4.8 (Randomness Deficiency)**.** *Given a probability distribution P, the* randomness deficiency *of an object x with respect to P is defined as*

$$\delta(x \mid P) = \log \frac{1}{P(x)} - K(x).$$

This measures how much less random x is compared to what P would suggest.

The randomness deficiency quantifies how much less random an observation x is compared to what the distribution P would suggest. We now apply this concept within a causal framework.

Definition 2.4.9 (Independence of Mechanisms Principle (IM))**.** *The* independence of mechanisms *principle asserts that the causal mechanisms* $P(X_i \mid \mathrm{PA}_i)$ *are algorithmically independent: knowledge of one mechanism provides no information about others. Formally, for any* $i \neq j$*, we have*

$$K(P(X_i \mid \mathrm{PA}_i) \mid P(X_j \mid \mathrm{PA}_j)) \approx K(P(X_i \mid \mathrm{PA}_i)).$$

Definition 2.4.10 (Outlier (Informal))**.** *Given a probability distribution* P *over a sample space* $\mathcal{X}$*, a data point* $x \in \mathcal{X}$ *is called an* outlier *if it exhibits significant deviation from the typical behavior expected under* P*. Or equivalently, if its* randomness deficiency

$$\delta(x \mid P) = \log \frac{1}{P(x)} - K(x)$$

is large, where $K(x)$ *denotes the Kolmogorov complexity of* x*.*

Definition 2.4.11 (Outlier Score)**.** *Given a sample* x_i *and its mechanism* $P(X_i \mid \mathrm{PA}_i)$*, the* outlier score *is defined as*

$$s_i(x_i) = \delta(x_i \mid P(X_i \mid \mathrm{PA}_i)).$$

The total score for a sample $x = (x_1, \ldots, x_n)$ *is*

$$s(x) = \sum_{i=1}^{n} s_i(x_i).$$

The utility of AIT is demonstrated by the following decomposition and conservation principles, which hold approximately under the independence of mechanisms.

Proposition 2.4.12 (Decomposability of Randomness Deficiency [EWJ25])**.** *Under the independence of mechanisms principle (IM), the total randomness deficiency decomposes across the nodes in the causal graph:*

$$\delta(x_1, \ldots, x_n \mid P) = \sum_{i=1}^{n} \delta(x_i \mid P(X_i \mid \mathrm{PA}_i)) + O(\log K(P)).$$

Theorem 2.4.13 (Weak Outlier Conservation [EWJ25] (informal))**.** *Let $x = (x_1, \ldots, x_n)$ be a sample from a CBN and P be the joint distribution. If x_j is not an outlier under its local mechanism, then its presence cannot cause x_i to become a strong outlier under another mechanism. That is, weak local deviations cannot amplify into global anomalies.*

Part I

Learning, Testing, and Inference from Structured Model

Chapter 3

Identifiability of Linear AMP Chain Graph Models

3.1 Introduction

Probabilistic graphical models offer architectures for modeling and representing uncertainties in decision making. From a computational standpoint, graphical representations enable efficient algorithms for inference, e.g., message passing, loopy belief propagation, and other variational inference methods [KFL01]. They have found applications in a wide range of domains, e.g., image processing, natural language processing and computational biology; see [KF09b; WJ08a] and references therein for examples.

A typical application of graphical models is to encode causal information. An influential article from [Pea95] elucidated how *Bayesian networks* can be used to represent causal processes and allow identification of causal effects. The graphical structure of a Bayesian network is a directed acyclic graph (DAG). Each node has a functional dependency on its parents, as determined by the graph. A popular way to substantiate Bayesian networks is as a *linear structural equation model (SEM)* where variables that correspond to nodes in the graph are a linear function of their parents' values plus additive independent noise (often Gaussian) [Bol89; Spi+00]. [Hoy+08a] defined the more general *additive noise model* where each node is an arbitrary function of its parents with an additive independent noise.

While Bayesian networks offer a clear conceptual way to model the causal structure of a system, they are in practice very hard to infer from data, as they require knowledge of how every single variable is generated. In applications involving hun-

dreds of variables (e.g., in computational biology), this requirement is unreasonable, particularly because at the end, we may only be interested in causal effects on a few target variables. Furthermore, in SEMs modeled by Bayesian networks, the noise terms of different variables must be independent whereas in real-world systems, correlations can arise for various reasons (e.g., latent confounders). An interesting middle ground is the notion of *chain graphs* [LW89]. Here, the variable set is partitioned into *chain components*, and there is a DAG on these chain components. The variables inside each chain component, however, are connected by undirected edges, not directed ones. See Fig. 3.1 for an illustration. Thus, chain graph models interpolate between directed (causal) models and undirected (probabilistic) models.

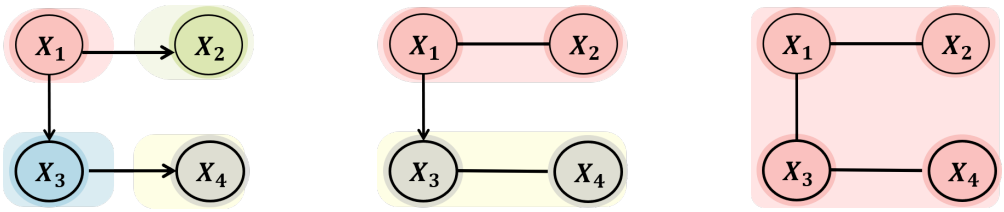


Figure 3.1: Chain graphs. Each shaded region is a maximal chain component.

There are several prevalent interpretations of chain graph models, namely the Lauritzen-Wermuth-Frydenberg (LWF) [LW89; Fry90], Alternative Markov Property or Andersson-Madigan-Perlman (AMP) [AMP01], and Multivariate Regression (MVR) [CW93]. They differ in the conditional independence relations implied by the graphical structure. In this work, we restrict ourselves to the linear AMP model, which is very natural from a generative viewpoint. Let $\mathcal{C}$ be an AMP chain graph[1] on n nodes. Suppose the nodes are partitioned into chain components $\{\tau\}$. Then, we say that a random variable $X \in \mathbb{R}^n$ is *generated by* $\mathcal{C}$ if for every chain component τ:

$$X_\tau = M_\tau X_{\mathrm{Pa}(\tau)} + Z_\tau \tag{3.1}$$

where X_τ is X restricted to τ, $\mathrm{Pa}(\tau) = \{v : \exists u \in \tau, v \to_{\mathcal{C}} u\}$, M_τ is a matrix satisfying:

$$(M_\tau)_{uv} \neq 0 \implies v \to_{\mathcal{C}} u,$$

[1]See Notations and Preliminaries section for formal definitions.

and Z_τ is independent from $X_{\mathrm{Pa}(\tau)}$ and is a multivariate Gaussian drawn from $N(\mathbf{0}, \Sigma_\tau)$ where Σ_τ satisfies:

$$(\Sigma_\tau^{-1})_{uv} \neq 0 \implies u \text{ —}_{\mathcal{C}}\ v$$

The last condition ensures that $N(0, \Sigma_\tau)$ is Markovian with respect to the undirected induced subgraph $\mathcal{C}_\tau$ on τ. One may also consider the additive noise AMP formulation where each

$$X_\tau = f_\tau(X_{\mathrm{Pa}(\tau)}) + Z_\tau, \tag{3.2}$$

the noise Z_τ is as above, and the function f_τ is arbitrary, provided it satisfies the directed graph structure:

$$\frac{(\partial f_\tau)_u}{\partial X_v} \neq 0 \implies v \to_{\mathcal{C}} u.$$

The directed edges of the AMP chain graph form a Bayesian network structure on the chain components, while for each τ, the undirected induced subgraph $\mathcal{C}_\tau$ describes a Gaussian Markov model for $X_\tau \mid X_{\mathrm{Pa}(\tau)}$.

In this chapter, we focus on the question of *identifiability* of chain graph models. That is, given knowledge of the distribution of X, can we recover the AMP chain graph $\mathcal{C}$ generating X? Moreover, can we recover $\mathcal{C}$ in polynomial time? For Bayesian networks[2], the study of identifiability has received sustained attention for more than two decades. In general, the problem is computationally hard [Chi96], but by making faithfulness or related assumptions, many sets of researchers (e.g., [Spi+00; Chi02b; ZS16; RU18]) have shown that the underlying DAG can be recovered up to its Markov equivalence class. This is quite unsatisfactory as the faithfulness assumption becomes too restrictive in the presence of finite sample error and the DAG is not uniquely identifiable. In a different line of work, [PB14a] showed that $\mathcal{C}$ is exactly identifiable for linear Gaussian SEMs if all the noise terms have equal variance. [GH17b; GH18a] and [PK20] established identifiability conditions for linear SEMs even with unknown heterogeneous error variances. Most recently, [Par20b] extended these conditions to additive noise models, while [GDA20a] further generalized to arbitrary Bayesian networks. See also [Ebe17] and [GZS19] for other perspectives.

[2]For Gaussian graphical models, identifiability reduces to finding the inverse of the covariance matrix.

We extend these identifiability conditions from DAGs to linear AMP chain graph models. Our main contributions are:

(i) **Additive noise AMP with known chain component decomposition**: We give a general class of identifiability conditions (generalizing the equal variance condition for linear SEMs) that imply identifiability of the DAG on a known collection of chain components. For instance, the DAG is identifiable if the determinant of the conditional covariance of a chain component τ given τ's parents[3] is the same for all τ. More generally, it is sufficient for this determinant to be monotonically non-decreasing with respect to a topological order on the chain components. The same is true if the trace or the permanent satisfies the monotonicity condition.

(ii) **AMP with unknown chain component decomposition**: We give an identifiability condition for recovering the chain components as well as the DAG for the standard AMP chain graph model. Informally, the requirement is quite natural: the variables in each chain component should be tightly correlated, while as a whole, each chain component should have large variance conditioned on its parents. More formally, the conditions are that:

   (a) If S is a proper subset of a chain component τ:

$$\det(\mathrm{Cov}(X_S \mid X_{\tau \setminus S}, X_{\mathrm{Pa}(\tau)})) < 1$$

   (b) $\det(\mathrm{Cov}(X_\tau \mid X_{\mathrm{Pa}(\tau)}))$ is greater than 1 and equal for all chain components τ. (Again, similar to (i) above, one can relax "equal" to "monotonically non-decreasing".)

In our conditions, the determinant of the covariance matrix of Gaussians plays a central role, and this is for good reason. If $X \sim N(0, \Sigma)$ is an n-dimensional Gaussian, then $\det(\Sigma)$ is the *generalized variance* of X and is related to its *differential entropy*. Namely, the differential entropy of X is $\frac{1}{2}(\log\det(\Sigma) + n\log(2\pi e))$; see, e.g., [KSG08; Yu15]. So, one can interpret condition (a) above as: If S is a proper subset of τ, its differential entropy conditioned on $\tau \setminus S$ and τ's parents is

[3]Note that if the generating equation is $X_\tau = B \cdot X_{Pa(\tau)} + Z_\tau$, where $Z_\tau = (\mathbf{0}, \Sigma_\tau)$ is the noise, then $\mathrm{Cov}(X_\tau | X_{Pa(\tau)}) = \Sigma_\tau$.

smaller than a threshold. Similarly, the first part of condition (b) can be restated as: If S equals τ, the differential entropy of S conditioned on its parents is larger than a threshold.

These identifiability conditions come with polynomial time algorithms. Notably, our algorithm for recovering the chain components in (ii) above involves a non-trivial submodular function minimization, in contrast to the more straightforward algorithms known for identifying linear SEMs and Bayesian networks [Par20b; GDA20a] under analogous conditions. The conditions in (i) and (ii) that determinants of residual covariances are equal is especially relevant where each chain component corresponds to the same physical system (e.g., in time series data).

3.2 Technical Overview

In this section, we describe some of the intuition behind our identifiability conditions.

Figure 3.2: Chain graph identifiability: how to determine which of these graphs is generating a given joint distribution $P(X_1, X_2, X_3)$?

Known chain components. Consider Fig. 3.2 which shows two chain graphs $\mathcal{C}_1$ and $\mathcal{C}_2$; the question is to determine which of these graphs is generating a given joint distribution (X_1, X_2, X_3). In $\mathcal{C}_1$, let $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}(0, \Sigma_1)$, and $X_3 = \beta_1 X_1 + Z$, where $\beta_1 \neq 0$ and $Z \sim \mathcal{N}(0, \sigma^2)$. In $\mathcal{C}_2$, let $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} \beta_2 \\ 0 \end{pmatrix} X_3 + Z$ where $\beta_2 \neq 0$, $Z \sim \mathcal{N}(0, \Sigma_2)$ and $X_3 \sim \mathcal{N}(0, \sigma^2)$. Assume $\mathrm{Det}(\Sigma_1) = \mathrm{Det}(\Sigma_2) = \sigma^2$, so that in both models, the determinant of the covariance of each chain component conditioned on its parents is σ^2.

We claim one can distinguish between $\mathcal{C}_1$ and $\mathcal{C}_2$ based on the generated distribution. Our algorithm first finds the chain component τ minimizing $\det(\mathrm{Cov}(X_\tau))$. Note that for $\mathcal{C}_1$, using the independence of Z: $\mathrm{Cov}(X_3) = \beta_1^2 \mathrm{Cov}(X_1) + \mathrm{Cov}(Z) \succ \mathrm{Cov}(Z)$, assuming[4] $\mathrm{Cov}(X_1) \succ 0$. Hence, $\det(\mathrm{Cov}(X_3)) > \det(Z) = \sigma^2 = \det(\mathrm{Cov}(X_{12}))$. On the other hand for $\mathcal{C}_2$, $\det(\mathrm{Cov}(X_{12})) > \sigma^2 = \det(\mathrm{Cov}(X_3))$. Thus, the chain component with the smallest determinant of the covariance can be identified as the first in a topological ordering. This can be understood as the uncertainty level of the parents is less than its children. Once the first chain component is known, we can select the second by choosing the one that minimizes the determinant of its covariance conditioned on the first chain component, and so on. It suffices to find the topological order because as described in [GDA20a], one can identify the directed edges by standard variable selection methods.

Note that the only property we used of the determinant is that $\det(A + B) > \det(A)$ if B is strictly positive definite. This property holds not only for the determinant but for many natural matrix functions. For example for any i, the diagonal entries $(A + B)_{ii} > A_{ii}$ when A and B are positive definite. Carrying out the same logic as above but now using projection to diagonal entries instead of determinants implies that the chain component DAG is identifiable when all the individual variables have equal variance, extending the result of [PB14a] to chain graphs. In fact, there is a large class of functions called "generalized matrix functions" that satisfy the desired super-additivity condition and hence result in identifiability conditions for the DAG on chain components.

Unknown chain components. Consider again $\mathcal{C}_1$ from Fig. 3.2, but suppose now that we do not have the chain component partitioning. Let (X_1, X_2, X_3) be generated as described above. In addition to imposing the condition that $\det(\Sigma_1) = \sigma^2$, we now also require that: (i) $\det(\mathrm{Cov}(X_1|X_2))$ and $\det(\mathrm{Cov}(X_2 \mid X_1))$ are[5] strictly less than 1, and (ii) σ^2 is strictly greater than 1.

[4] In this work, we make the assumption everywhere that all covariance matrices are strictly positive definite.

[5] $\det(\mathrm{Cov}(X_2 \mid X_1))$ is well defined, since (X_1, X_2) are jointly Gaussian, and hence, for any choice of x_1, $\mathrm{Cov}(X_2 \mid X_1 = x_1)$ is the same.

Now, we can show that

$$\det(\mathrm{Cov}(X_{12})) = \min_{S \subseteq \{1,2,3\}} \det(\mathrm{Cov}(X_S)).$$

Observe that $\det(\mathrm{Cov}(X_3)) > \det(\mathrm{Cov}(X_{12}))$ already follows from the earlier discussion. We now compare $\det(\mathrm{Cov}(X_{12}))$ to $\det(\mathrm{Cov}(X_1))$ and $\det(\mathrm{Cov}(X_2))$. We use the fact that:

$$\det(\mathrm{Cov}(X_{12})) = \det(\mathrm{Cov}(X_1)) \cdot \det(\mathrm{Cov}(X_2 \mid X_1)).$$

This follows from standard facts about multivariate Gaussians. From our assumption $\det(\mathrm{Cov}(X_2 \mid X_1)) < 1$, we get that $\det(\mathrm{Cov}(X_1)) > \det(\mathrm{Cov}(X_{12}))$. The same holds for $\det(\mathrm{Cov}(X_2))$. Finally, we need to show that $\det(\mathrm{Cov}(X_{123})) > \det(\mathrm{Cov}(X_{12}))$. Again, we can invoke the above fact:

$$\det(\mathrm{Cov}(X_{123})) = \det(\mathrm{Cov}(X_{12})) \cdot \det(\mathrm{Cov}(X_3 \mid X_{12})).$$

Our conclusion follows from the assumption $\sigma^2 > 1$.

For a general chain graph, it similarly follows that the non-empty set S minimizing $\det(\mathrm{Cov}(X_S))$ is the topologically smallest. We can identify the next component by conditioning on the components already discovered, which results in a Gaussian on the rest, and then finding a non-empty subset with conditional covariance matrix of smallest determinant. This algorithm can be implemented efficiently. The reason is that for any positive definite $n \times n$-matrix M, the function $F(S) = \log\det(M[S,S])$, where $M[S,S]$ is the submatrix on rows and columns indexed by $S \subseteq [n]$, is *submodular*. F, as noted earlier, corresponds to the differential entropy of a Gaussian vector with covariance M, which is a submodular function, plus an additional modular term. The problem of submodular function minimization has a long and rich history, beginning with the seminal works of [GLS81; GLS12] and continuing to the current day [IFF01; Sch00; LSW15; DVZ18; Jia21]. Thus, we can invoke any of these known polynomial-time algorithms for submodular function minimization to recover the chain components in topological order.

3.3 Related Work

3.3.1 Learning DAG Models.

The literature on learning pure DAG models is vast. One popular approach is to exploit the constraints imposed by Markov structure, e.g., the PC algorithm and its variants, like Fast Causal Inference (FCI), Really Fast Causal Inference (RFCI) and Cyclic Causal Discovery (CCD) [SGS01; Spi+00; Ric13; Col+11; TS13; HD13; CM14] under different assumptions. Another important class of algorithms aims to maximize a score function over the space of DAG's, such as Greedy Equivalence Search (GES) [Chi02b; Ram+17; NHM+18] and a recent line of work that formulates score maximization as a continuous optimization problem (e.g., [Zhe+18; Zhe+20; WGY20]). This latest direction has resulted in algorithms that learn the DAG structure with deep learning methods (e.g., [Yu+19; Lac+20; Wan+20]).

3.3.2 DAG Identifiability.

A probability distribution may be Markov with respect to many Bayes networks; so for exact identifiability, one needs to impose more structural constraints on the DAG model. For Structural Equation Models (SEM's), identifiability can be established by leveraging asymmetries between variable pairs [Shi+06a; Moo+16], restricting SEMs to having additive noise, such as linear non-Gaussian acyclic model (LiNGAM) [Shi+06a], general additive noise models [Pet+14], Post-nonlinear model (PNL) [Zha+16], or equal and unknown error variance [PB14a; GH17b; Ebe17; GH18a; CDW19; GZS19; PK20; Par20b; GDA20a].

3.3.3 Partially Directed Acyclic Graph (PDAG)

An acyclic graph containing both directed and undirected edges is a *Partially Directed Acyclic Graph (PDAG)* [VP90], also called *chain graph (C)*. When the entire PDAG is undirected, it forms a single chain component, also known as *undirected graphical models (UGs)*, *Markov random fields* or *Markov networks*. When the entire PDAG is directed, each node is a chain component, also known as *Directed Acyclic Graphical Models (DAGs)*.

3.3.4 Markov Equivalence Class

Chain graphs are commonly used to represent equivalence classes of DAGs. This happens when we targets at learning a DAG model from purely observational data, and generally cannot identify the unique DAG representing the underlying data generation mechanism, but can potentially obtain a *Markov equivalence class*. [VP90] proves that two DAGs are Markov equivalent if and only if they have the same v-structures and the same skeleton. Moreover, [AMP97] shows that a Markov equivalence class can be represented uniquely by an *essential graph*. This kind of representation also named as *patterns* [SM95], *maximally oriented graphs*[Mee95], or *completed PDAG (CPDAG)* [Chi02a]. A Markov equivalence class is used to represent a set of DAGs.

3.3.5 Chain Graph Interpretations

While DAGs works well in representing the asymmetric cause and effect relationships. However, the representation falls short if we have both symmetric and asymmetric relations simultaneously in a system. To further facilitate the characterizing of both symmetric and asymmetric relationships together, there exist three main different interpretations of chain graphs in the literature: the Lauritzen-Wermuth-Frydenberg (LWF) [LW89; Fry90], Alternative Markov Property or Andersson-Madigan-Perlman (AMP) [AMP01], and Multivariate Regression (MVR) [CW93]. The interpretation of directed edges is quite clear, the undirected edge can have many different interpretations. Depending on the representation of the type of edges.

3.3.6 Three Types of Chain Graph Models

Specifically, in LWF chain graph, undirected edges represents causal effects due to interference [SP15; STA17; BMS20]. Besides, MVR models [CW93; JV18] are equivalent to the acyclic directed mixed graphs without semi-directed cycles. Different from the undirected edges in LWF, MVR CGs actually contains bi-directed edges, represent one or more hidden common causes between the variables connected by it. For example, $X \longleftrightarrow Y$ can be repaleced by $X \leftarrow H \rightarrow Y$. In the AMP CG model [AMP01], it preserves some component-wise characteristics of DAGs, and

thus can be seen as a more direct generalization of DAG Markov properties than the LWF Markov properties of CGs. The undirected edges in AMPs can be seen as UGs.

In short, nodes in LWF and AMP CGs are connected to each other by undirected edges. In MVR CGs, nodes are connected by bidirected edges. The chain components are then themselves connected to each other by directed edges. LWF, AMP and MVR CGs are just suited to different problems, similarly to the way that DAGs and UGs are different from each other.

3.3.7 Learning AMP Chain Graph Models.

AMP chain graphs, our focus in this work, have been less widely studied than pure DAG models and more in the statistics literature than computer science. Informally speaking, [Peñ15] showed that any AMP model can be viewed as arising from a DAG causal model subject to selection bias. [LPM01] introduced a pathwise separation criterion to characterize conditional independence relations in AMP chain graphs. [Rov05; SR09; Peñ17a] studied the equivalence classes of chain graph models, and [Peñ18] provided a factorization for positive distributions that are Markov with respect to an AMP chain graph. [Drt+09] showed that the AMP conditional independence relations may lead to non-smooth models for discrete variables. [Peñ14b; Peñ16] investigated extensions to the AMP model, e.g., the marginal AMP model (MAMP) that is a common generalization of AMP and MVR. When the chain graph structure is known, [DE06] proposed an algorithm for maximum likelihood estimation of the model parameters. [Peñ12; Peñ14a; PG16] proposed PC-LIKE, a constraint based algorithm under faithfulness assumptions for learning the structure of AMP and MAMP models. Peña also designed a score-based algorithm for AMP model structure learning similar to the work on additive noise models [Peñ17b] and an algorithm based on answer set programming [Peñ16]. Recently, [JVJ20] solved the problem of efficiently finding minimal separating sets in AMP chain graphs and obtained a new decomposition-based structure learning algorithm called LCD-AMP.

3.4 Identifiability with Known Chain Component Decomposition

In this section, we give a general class of conditions which are sufficient to ensure that the DAG structure of the chain graph is identifiable from data generated by it. Here, the chain component decomposition $\mathcal{D}$ is already known to the algorithm. $\mathcal{D}$ consists of t disjoint maximal chain components that partition the variable set.

We formulate our results for general AMP chain graph models. They will immediately imply the conditions for additive noise AMP models mentioned in Section 2.3.

Algorithm 1: Learning the topological order of a chain graph with chain component decomposition $\mathcal{D}$ of size t

Input : $\mathcal{A}$, $P \leftarrow \emptyset$, $i \leftarrow 0$
while $|\mathcal{A}| \neq t$ **do**
  $\tau_i \leftarrow \arg\min_{\tau \in \mathcal{C} \setminus \mathcal{A}} d_{|\tau|}\left(\mathbb{E}\left[\mathrm{Cov}(X_\tau \mid X_P)\right]\right)$;
  $\mathcal{A} \leftarrow \mathcal{A} \cup \{\tau_i\}$;
  $P \leftarrow P \cup \tau_i$;
  $i \leftarrow i + 1$;

Theorem 3.4.1. *Suppose the random variable X is generated by an AMP-CG $\mathcal{C}$ with known chain component decomposition $\mathcal{D}$. Then, $\mathcal{C}$ is identifiable from P if there exists a topological ordering π of $\mathcal{C}$ and a positive and super-additive family $\{d_n : \mathbb{C}_n \to \mathbb{R}\}$ such that:*

$$d_{|\tau|}\left(\mathop{\mathbb{E}}_{X_{\mathrm{Pa}(\tau)}} \mathop{\mathrm{Cov}}_{X_\tau}(X_\tau \mid X_{\mathrm{Pa}(\tau)})\right) \leq d_{|\tau'|}\left(\mathop{\mathbb{E}}_{X_{\mathrm{Pa}(\tau')}} \mathop{\mathrm{Cov}}_{X_{\tau'}}(X_{\tau'} \mid X_{\mathrm{Pa}(\tau')})\right) \tag{3.3}$$

for any two chain components τ, τ' where $\tau \prec_\pi \tau'$.

The following corollary is immediate.

Corollary 3.4.2. *Suppose X corresponds to an additive noise model generated by a chain graph $\mathcal{C}$, i.e.:*

$$X_\tau = f_\tau(X_{\mathrm{Pa}(\tau)}) + Z_\tau,$$

where the noise term Z_τ is independent of $X_{\mathrm{Pa}(\tau)}$, for all chain components τ of $\mathcal{D}$.

Then, given the chain component decomposition, a topological ordering of $\mathcal{D}$ is identifiable from X if there exists a topological ordering π of $\mathcal{D}$ such that

$$\det(\mathrm{Cov}(Z_\tau)) \leq \det(\mathrm{Cov}(Z_{\tau'}))$$

for all chain components $\tau \prec_\pi \tau'$.

Non-parametric algorithm. We give a finite-sample version algorithm using the determinant as d_n. One can estimate $\det(\mathbb{E}[\mathrm{Cov}(X_\tau \mid X_P)])$ by (i) using a non-parametric regressor $\widehat{F}_{\tau,P}(X_P)$ to estimate $\mathbb{E}[X_\tau \mid X_P]$ with n_1 samples, and (ii) using a plug-in estimator on n_2 samples:

$$\det\left(\frac{1}{n_2}\sum_{i=1}^{n_2}\left((X_\tau^{(i)})^{\otimes 2} - \widehat{F}_{\tau,P}^{\otimes 2}(X_P^{(i)})\right)\right)$$

Using standard non-parametric regularity conditions, we can lower bound the probability of the algorithm recovering the true topological order. Note that the result does not depend on the particular choice of the estimator $\widehat{F}_{\tau,P}$ as long as it is asymptotically consistent. Due to space constraints, a detailed statement is deferred to the appendix.

3.5 General Identifiability

In this section, we establish identifiability conditions for recovering both the chain components as well as the DAG structure of chain graphs from the generated probability distribution. Here, by identifiability, we mean that the partitioning into chain components and the topological order on the chain components are uniquely specified. The exact set of directed and undirected edges can then be recovered using standard variable selection methods (as described in Appendix A of [GDA20a]).

Theorem 3.5.1. *Suppose the random variable X is generated by an AMP-CG $\mathcal{C}$ with unknown structure. Then, $\mathcal{C}$ is identifiable from X if the following three conditions hold:*

(i) For all chain components τ and all non-empty proper subsets $S \subset \tau$:

$$\det(\mathrm{Cov}(X_s \mid X_{\tau\setminus s}, X_{\mathrm{Pa}(\tau)})) < 1.$$

Algorithm 2: Infinite sample algorithm for learning the topological order of a chain graph with unknown chain components.

Input : $P \leftarrow \emptyset$, $i \leftarrow 1$

$\tau_1 = \arg\min_{S \subseteq V,\, S \neq \emptyset} \det(\mathrm{Cov}(X_S))$;

$P \leftarrow P \cup \tau_1$;

while $V \setminus P \neq \emptyset$ **do**

  $\tau_i \leftarrow \arg\min_{S \subseteq V \setminus P,\, S \neq \emptyset} \det(\mathrm{Cov}(X_S \mid X_P))$;

  $P \leftarrow P \cup \tau_i$;

  $i \leftarrow i + 1$;

return the topological sort $(\tau_1, \ldots, \tau_i)$;

(ii) For all chain components τ:

$$\det(\mathrm{Cov}(X_\tau \mid X_{\mathrm{Pa}(\tau)})) > 1.$$

(iii) There is a topological order π on the chain components such that for all $\tau \preceq_\pi \tau'$:

$$\det(\mathrm{Cov}(X_\tau \mid X_{\mathrm{Pa}(\tau)})) \leq \det(\mathrm{Cov}(X_{\tau'} \mid X_{\mathrm{Pa}(\tau')})).$$

Informally speaking, for any subset S, given its complementary set and parents union of τ in $\mathcal{C}$, we require the variables in each chain component to be tightly correlated. Besides, given the union of the parents of chain components τ, we require the clustered variables in each chain component to have large generalized variance. The third condition is the same one imposed in the identifiability with known chain component decomposition section.

There is a geometric way to view the conditions in Theorem 3.5.1, which substantiates the intuition that they require each chain component to cluster together while having large variance as a whole. Recall that for any matrix M, $\det(M)$ corresponds to the volume of the parallelepiped spanned by the rows of M. Let the chain components be denoted $\tau_1, \ldots, \tau_k$ in a topological order. For $i = 1, \ldots, k$, let M_i denote the covariance matrix of $X_{\tau_i} \mid X_{\tau_1 \cup \cdots \cup \tau_{i-1}}$, and let M denote the full covariance matrix, $\mathrm{Cov}(X_{\tau_1 \cup \cdots \cup \tau_k})$. From fact 2.2.9,

$$\det(M) = \det(M_1) \cdots \det(M_k). \tag{3.4}$$

Let V_i denote the set of row vectors of M_i, and we identify V_i with the parallelepiped it spans. Due to Eq. (3.4), we can view each V_i as residing in a subspace

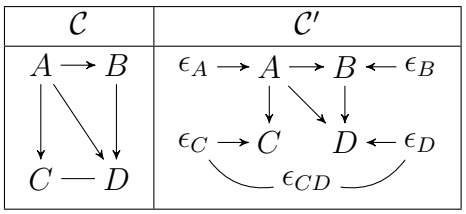


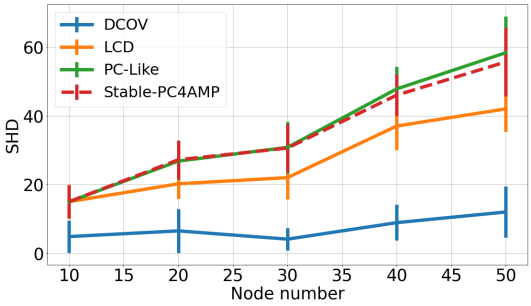
Figure 3.3: Synthetic data generation. Undirected edges correspond to correlated noise.

orthogonal to the spans of other V_j's, so that their volumes just multiply with each other. (Alternatively, construct a block diagonal matrix M' where the i'th block on the diagonal is M_i; clearly, $\det(M) = \det(M')$.) In this language, Condition (ii) in Theorem 3.5.1 says that the volume of each V_i is more than 1, and condition (iii) says that the volumes are non-decreasing with i. Condition (i) says that for any V_i, the volume of any sub-parallelepiped is larger than the volume of the whole. Intuitively, this means that the vectors in V_i form very small angles with each other, so that the volumes keep decreasing as more vectors are added.

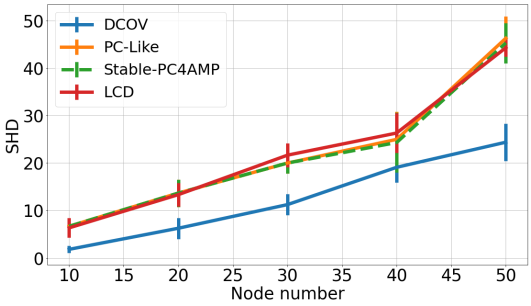
Computational Efficiency. It is known that Algorithm 2 can be implemented in polynomial time. This is because the optimization problems in lines 3 and 5 of the pseudocode correspond to submodular function minimization, as explained earlier. A slight non-triviality is that the optimization is over all non-empty sets instead of over all sets. However, it is well-known how to reduce this to unconstrained minimization (e.g., see Section 4.1 of [GR20]).

3.6 Experiments

In this section, we compare the performance of Algorithm 1 and Algorithm 2 on synthetic datasets to state-of-the-art methods for AMP chain graph structure learning. Recall that as we showed in Theorem 3.4.1, the DAG on the chain components of an AMP chain graph is identifiable if (3.3) is satisfied for a positive and super-additive family d_τ. Here, we let d_τ be the determinant operator, and hence dub our algorithm as Determinant of Covariance (DCOV).



Algorithm 1 (known chain components)



Algorithm 2 (unknown chain components)

Figure 3.4: SHD performance (lower is better)

3.6.1 Synthetic Data Generation.

To generate the chain graph $\mathcal{C}$, in our first step, an undirected graph $\mathcal{G}$ with n nodes is generated by using the Erdős Rényi (ER) model with an expected neighbor size $s = 2$ and then symmetrizing. Given the number of chain components c, we split the interval $[1, n]$ into c equal-length sub-intervals $[I_1, \ldots, I_c]$ so that variable sets for each sub-interval forms chain components $\tau_1, \ldots, \tau_c$. Meanwhile, for any (i, j) pair, we set $\mathcal{C}_{i,j} = 0$ if $\exists i \in I_\ell, j \in I_m, \ell > m$. Given the binary adjacency matrix $\mathcal{C}$, we generate the matrix M of edge weights by $M_{i,j} \sim U(-1.5, -0.5] \cup U[0.5, 1.5)$ if $\mathcal{C}_{i,j} \neq 0$ and $M_{i,j} = 0$ otherwise.

The observational i.i.d. data $X_\tau = M_\tau X_{\mathrm{Pa}(\tau)} + Z_\tau$ is generated with a sample size $n = 1000$ and a variable size $d \in \{10, 20, 30, 40, 50\}$. Z_τ is an independent multivariate Gaussian drawn from $N(0, \Sigma_\tau)$ where Σ_τ is generated randomly with $\det(\Sigma_\tau) = 1$, satisfying the assumption of Corollary 3.4.2. Figure 3.3 illustrates how the synthetic AMP chain graph data is generated.

3.6.2 Baseline Algorithms.

We compare our DCOV method against the PC-Like [Peñ12; Peñ14a; PG16], Stable-PC4AMP [JVJ20], and LCD algorithm (Learn Chain Graphs via Decomposition) [MXG08]. We use default parameters among those baseline algorithms in order to avoid skewing the results in favour of any particular algorithm as a result of hyperparameter tuning[6]. All the baseline algorithms above are implemented

[6]The implementation of baseline algorithms is available at https://github.com/majavid/AMPCGs2019.

using R-packages (licensed under GPL-2 or GPL-3) such as `ggm` [Mar+06], `pcalg` [Kal+12], `mgcv` [WW15], `np` [RH20], and `lcd` [MXG08]. We use `rpy2` [Gau12] to access R-packages from Python and ensure that all algorithms can be compared in the same environment. The results are averaged over 20 independent repetitions. The experiments were conducted on an Intel Core i7-9750H 2.60GHz CPU.

3.6.3 Implementation of DCOV.

We implement Algorithm 2 using the Matlab toolbox `Submodular Function Optimization` [Kra10]. Each iteration of Algorithm 1 and Algorithm 2 estimates the conditional covariance of the remaining chain components using the finite-sample algorithm mentioned earlier. Like [GDA20a], we run a *gam* regression to estimate conditional expectations. We set the p-value with significance level of 0.001 for determining the parents of the node.

3.6.4 Performance Evaluation Metrics.

We evaluate the performance of the proposed algorithms in terms of the four measurements, namely, true positive rate (TPR), false positive rate (FPR), accuracy (ACC), and structural hamming distance (SHD) that are commonly used in [JVJ20; CM14; MXG08].

3.6.5 Agnostic learning

When the chain components are unknown, our theoretical results treat the case that the data is realized following the conditions in Theorem 3.5.1. A straightforward question is: how the algorithm performs when the condition is violated? To answer this question, we conduct agnostic learning experiments by showing the experiment results based on the following conditions:

1. *Chain graph experiments:* Take the opposite condition from Theorem 3.5.1, where $\det(\mathrm{Cov}(X_\tau \mid X_{\mathrm{Pa}(\tau)})) \leq 1$.

2. *DAG experiments:* Evaluate the performance of our algorithms on synthetic Directed Acyclic Graph (DAG) data;

a) The variance for each node is equal and > 1;

b) The variance for each node is uniformly in $[0.5, 1.5]$;

3.6.6 Summary of Experiment Results.

As shown in Fig. 3.4, DCOV, under both known and unknown chain component conditions, shows superior performance compared with all other baselines by wide margins. This is because the identifiability condition we proposed provides a correctness guarantee for the recovery of chain-graph structures. Surprisingly, in the DAG structure learning task, if the condition in Theorem 3.5.1 holds, our unknown chain graph structure learning algorithm (Algorithm 2) can correctly identify the special one-node chain component structures. It also shows superior performance over all other baseline methods. Besides, in our chain graph agnostic learning experiments, when node number increases, although SHD is still lower than other baseline algorithms, the ACC, TPR, and FPR performances are relatively worse. Furthermore, we also conduct agnostic learning experiments on DAG structures. Since our proposed condition does not hold in this case, in the worst condition, Algorithm 2 can wrongly treat all the nodes in a DAG graph as one chain component. This leads to the highest SHD and FPR, and lower ACC performance in our experiment results. We also evaluate the performance of Algorithm 1 on four real Gaussian Bayesian networks from R package `bnlearn` [Scu09]. The **ECOLI70** graph provided by [SS05] contains 46 nodes and 70 edges. The **MAGIC-NIAB** graph from [Scu+14] contains 44 nodes and 66 edges. The **MAGIC-IRRI** graph contains 64 nodes and 102 edges. The experimental details are available in the supplementary material. One limitation of this work is the lack of real datasets that can be modeled by chain graphs.

3.7 Conclusion

In this chapter, we address the problem of recovering linear AMP chain graph in polynomial time from observational data, and we proposed two algorithms for both known and unknown chain components to handle the problem. In our experiments, we implemented our algorithms over both known and unknown chain components. As future work, we are also interested in exploring a score-based approach for chain

graph structure learning from observational data.

Chapter 4

Optimal estimation of Gaussian (poly) trees

4.1 Introduction

Graphical models are a classical statistical tool for efficiently modeling data with rich, combinatorial structure. Directed acyclic graphs (DAGs) are widely used to capture causal relationships among complex systems. Probabilistic graphical models defined on DAGs, known as Bayesian networks [Pea+00], have found broad applications in various disciplines, from biology [MS07; Zha+13; AE10], social science [GK08], knowledge representation [VLP08], data mining [Hec97], recommendation systems [HLH12], legal decision making [Tha04], and more. When this structure is known in advance, it is straightforward to exploit this structure for inference tasks, among other things [WJ08b]. When this structure is unknown, it is must first be learned from data, which is the difficult problem of *structure learning* in graphical models. First, observational data only reveal the Markov equivalence class, captured by a completed partially directed acyclic graph (CPDAG) [AMP97]. Classical approaches to learning a CPDAG from data include the PC algorithm [SG91; KB07] and GES [Chi02c; NHM+18]. Moreover, it is also known that the general problem of learning DAGs from observational data is an NP-complete problem [Chi96; CHM04; CDW19], although a few polynomial-time algorithms have been proposed for special cases [GH17c; CDW19; Par20a; GDA20b].

An important unresolved problem in this direction is to characterize the sample complexity of structure learning, or the minimum number of samples required to

learn the graph from data. The past decade has produced a broad literature on this problem, mainly focused on *undirected* graphical models (i.e. Markov random fields [Wai19; WWR10; SW12]). By comparison, much less is known about DAGs. In this paper, we study in detail the simplest unresolved DAG model, namely, directed Gaussian trees. Perhaps surprisingly, despite its simplicity, and unlike in the undirected case, the optimal sample complexity of learning directed Gaussian trees has remained an open problem. Suppose we are given sample access to a Gaussian distribution $P = \mathbb{N}(0, \Sigma)$, where the goal is to learn a DAG G that represents P. While we defer formal definitions to Chapter 2, we can broadly summarize three different problems to be addressed here at the outset:

1. (Non-realizable setting) When P is an arbitrary Gaussian (i.e. not representable by any tree), how many samples are required to learn a tree-structured distribution Q that is optimally close to P?

2. (Realizable setting) When P itself is tree-structured, how samples are required to learn a tree-structured distribution Q that is optimally close to P?

3. (Faithful setting) When P is faithful to some tree T, how samples are required to learn T itself (i.e. the tree structure) up to Markov equivalence?

It is well-known that each of these problems is solvable—in principle—under different assumptions. For example, the celebrated Chow-Liu algorithm solves the first two problems, however, whether or not this can be improved with a more efficient algorithm is unknown. The same goes for the third setting: The famous PC and GES algorithms can find a faithful DAG (even without the tree assumption), however, their optimality remains unresolved. One of our main contributions is to study all three problems in a single unified setting, allowing for apples-to-apples comparisons of the assumptions required, and the resulting (optimal) sample complexity for each.

Although faithfulness can be a strong assumption in practice, we emphasize that to the best of our knowledge, no optimality results under this assumption are known. Thus, our analysis presents a possible first foray in this direction. Previous work has shown that faithfulness is notoriously challenging to analyze [e.g. Uhl+13; GTA23].

4.1.1 Our Contributions

We are given n i.i.d. samples $X = (X^{(1)}, \ldots, X^{(n)}) \in \mathbb{R}^{n \times d}$ from an unknown Gaussian P. We consider two distinct but canonical problems: *Distribution learning* and *structure learning*. The difference between these two problems lies in the error metric: In distribution learning, we seek to learn P in KL-divergence, with no respect for underlying structure (i.e. there may be no structure at all), whereas in structure learning, we assume *a priori* the existence of a tree T and seek to learn T exactly, with no respect for the distribution P. Structure learning is known to require restrictive assumptions, and thus part of our effort is to illustrate how different assumptions lead to different conclusions and sample complexities. With this in mind, our results consider three progressively stronger assumptions on P: Non-realizable, realizable, and faithful.

Below, we outline our main contributions at a high-level, while deferring precise statements and problem formulations to Section 4.2 and Section 4.3.

Non-realizable Setting Without making additional assumptions on P, we show that[1]

$$n = \widetilde{\Theta}\left(\frac{d^2}{\varepsilon^2}\right) \tag{4.1}$$

are necessary and sufficient to learn (with probability at least 2/3) a tree-structured distribution that is ε-close to the closest tree-structured distribution for P.

Realizable Setting When P itself is Markov to a tree T (i.e. it is *tree-structured*), then

$$n = \widetilde{\Theta}\left(\frac{d}{\varepsilon}\right) \tag{4.2}$$

are necessary and sufficient to learn (with probability at least 2/3) a tree-structured distribution that is ε-close to P itself.

Faithful Polytrees Switching our goal from learning the closest tree-structured distribution to structure learning, we additionally assume that P is faithful to some

[1] $\widetilde{\Theta}$ is used to ignore potential log factors.

polytree T. We show that the optimal sample complexity of learning $\overline{T}$, the CPDAG of T, is

$$n = \Theta\left(\frac{\log d}{c^2}\right), \tag{4.3}$$

where c is a faithfulness parameter defined in Equation 4.3.2.

Clearly, and unsurprisingly, realizable distribution learning is easier than the non-realizable case. A more interesting question is how to compare these to structure learning. In Section 4.5, we conclude with a discussion and comparison of these two cases, with some intriguing directions for future work.

4.1.2 Other Related Work

Learning Bayesian Networks Structure learning of Bayesian networks has been extensively studied, and the reader may consult one of several overviews for more details and background [SGS00; Pea+00; KF09a; Mur12; PJS17; Maa+18; SU22]. Classical approaches assume faithfulness, a condition that permits learning of the Markov equivalence class, such as constraint-based methods [SG91; FNP13] and score-based approaches [Chi02c; NHM+18]. A different strand of research has explored a range of alternative distributional assumptions that allow for effective learning such as non-gaussianity [Shi+06b; Shi14; WD20], non-linearity [Hoy+08b; ZH09] or equal error variances [PB14b; GH17c; GH18b; CDW19; GDA20b].

When it comes to the tree-structured graphical model of a distribution, the classical Chow-Liu algorithm [CL68] can recover the skeleton of a non-degenerate polytree in the equivalence class. Furthermore, [CW73] demonstrate that as the number of samples approaches infinity, the Chow-Liu algorithm is *consistent*. One of the first papers to consider the problem of learning polytrees was [REB87], after which [Das99b] showed that learning polytrees is NP-hard in general. [Sre03] has shown that the related problem of finding the maximum likelihood graphical model with bounded treewidth is also NP-hard. Subsequent research by Tan et al. [TAW10; TAW11] investigated the recovery difficulty of trees and forests, while Liu et al. [Liu+11] adopted a nonparametric approach using kernel density estimates. The Chow-Liu algorithm has also been applied for learning latent locally tree-like graphs [AV13].

Sample Complexity of Structure Learning Early work to consider the sample complexity problem for Bayesian networks includes [FY96; ZMD12]. More recently, for distribution learning over finite alphabets, [DP20; DP21] showed that d-variable tree-structured Ising models can be learned computationally-efficiently to within total variation distance ε from an optimal $O(d \log d/\varepsilon^2)$ samples. Around the same time, [Bha+21] derived explicit sample complexity bounds for the Chow-Liu algorithm of $\widetilde{O}(d\varepsilon^{-1})$ for trees on d vertices, and $d^2\varepsilon^{-2}$ samples for a general distribution P. [Cho+23] further extend [Bha+21] into d-polytree when the underlying graph skeleton is known.

The literature on structure learning is comparatively deeper; however, it has traditionally forgone concerns about optimality and lower bounds. As this is our main focus, we focus here on prior work on optimal algorithms. [GH17a] first established lower bounds for a range of DAG models, after which [GTA22] showed that a variant of the algorithm from [CDW19] achieves optimal sample complexity of $n \asymp q \log(d/q)$ for equal variance DAGs [PB14b; LB14], where q is the maximum number of parents and d is the number of nodes. To the best of our knowledge, optimality results and lower bounds in the faithful setting are missing, one exception is the sub-problem of neighbourhood selection [GTA23], and one of our main contributions is to partially fill this gap. We mention prior work that considers consistency and upper bounds under faithfulness [KB07; NHM+18; REB+18], relaxation and improvement on classical methods [Chi20; MGM21; LAR22], and recent progress on learning polytrees [GA21; ATB21; TMD22; Jak+22].

Learning polytrees is among the easiest tasks in learning DAGs and has received attention in [CL68; KS01; RP13; Nie+14]. The crucial advantage of such networks is that they allow for a more efficient solution of the inference task [Pea88; GH02]. The complexity of polytree learning has been studied in several works [SMS13; Gas+15; GKM21]. A recent work in [GA21] shows that learning polytrees is more manageable than general DAG models, for which they establish clear conditions for the identifiability and learnability of nonparametric polytrees in polynomial time. Some other earlier works such as reconstruction of evolutionary trees can be found in [Bun71; CH91; Cha96; DMR11]. Besides, latent tree model is a class of latent variable models in which the graph may be a forest has received considerable attention [Cho+11; TAW11; SXP11; Ana+11; PSP11; MRS13; Son+14; Drt+17]. Specifi-

cally, [Ana+11] shows that the structure of multivariate latent tree models can be learned with a sample complexity depends solely on intrinsic spectral properties of the distribution. (also see survey paper [Mou+13] for more details). [Ana+12] proved that a $\mathrm{poly}(d, r)$ sample and computational requirements serves as a good approximation of a r-component mixture of d-variate graphical models.

Furthermore, developing a (conditional) independence tester with respect to mutual information with $o(1/\varepsilon^2)$ sample complexity was posed as an open problem in [Can+18]. In [Can+18], they have shown that both Ising model *Goodness-of-fit Testing* and Ising model *Independece Tesing* can be solved from $\mathrm{poly}(d, 1/\varepsilon)$ samples in polynomial time. More details related to the *distribution property testing* can be found in [Rub12; Can20; Gol17; BY22].

4.2 Learning Tree-structured Gaussians

We begin by studying the sample complexity for learning tree-structured Gaussian distributions. For any $\varepsilon > 0$, we would like to devise an algorithm taking samples drawn from a Gaussian P that returns a directed tree $\widehat{T} \in \mathcal{T}$ and a distribution $P_{\widehat{T}}$ that is Markov to $\widehat{T}$ such that

$$D_{\mathrm{KL}}(P \parallel P_{\widehat{T}}) \leq \min_{T \in \mathcal{T}} D_{\mathrm{KL}}(P \parallel P_T) + \varepsilon\,,$$

We seek to achieve this goal with a minimal number of samples.

Notably, for any $T \in \mathcal{T}$, $D_{\mathrm{KL}}(P \parallel P_T)$ can be expressed as

$$-\sum_{i=1}^{d} I(X_i; X_{\mathrm{pa}(i)}) - H(X) + \sum_{i=1}^{d} H(X_i), \tag{4.4}$$

where H is the entropy function and I is the mutual information.

4.2.1 Distribution Learning Upper Bounds

The classical Chow-Liu algorithm [CL68] builds the maximum weight spanning tree where the weight of the "potential" edge between nodes j and k is the estimated mutual information $\widehat{I}(X_j, X_k)$ from data. Although its return is an undirected graph, we modify the output to be any directed tree whose skeleton matches the undirected graph with light abuse of notation. This is because any $T \in \mathcal{T}$ with the

same skeleton will share the same P_T, which is the target of distribution learning analyzed in the sequel.

Algorithm 3: Modified `Chow-Liu` algorithm

Input: n i.i.d. samples $(X_1^{(i)}, \ldots, X_d^{(i)})$

1. For each $j = 1, \ldots, d$:
   a) $\widehat{\sigma}_j^2 \leftarrow \frac{1}{n} \sum_{i=1}^n (X_j^{(i)})^2$
2. For each pair $(j, k), 1 \le j < k \le d$:
   a) $\widehat{\rho}_{jk} \leftarrow \frac{1}{n} \sum_{i=1}^n X_j^{(i)} X_k^{(i)}$
3. For each pair $(j, k), 1 \le j < k \le d$:
   a) $\widehat{I}(X_j; X_k) \leftarrow -\frac{1}{2} \log\left(1 - \frac{\widehat{\rho}_{jk}^2}{\widehat{\sigma}_j^2 \widehat{\sigma}_k^2}\right)$ which is same as $\frac{1}{2} \log(1 + \frac{\widehat{\beta}_{jk}^2 \widehat{\sigma}_j^2}{\widehat{\sigma}_{k|j}})$ defined in Section B.2.2
4. $G \leftarrow$ the weighted complete undirected graph on $[d]$ whose edge weight for (j, k) is $\widehat{I}(X_j; X_k)$
5. $\widehat{S} \leftarrow$ the maximum weighted spanning tree of G
6. $\widehat{T} \leftarrow$ any directed tree with skeleton to be $\widehat{S}$

Output: A directed tree $\widehat{T}$

Our first result gives an upper bound on the sample complexity for distribution learning in the non-realizable setting:

Theorem 4.2.1. *Let P be a Gaussian distribution. Given n i.i.d. samples from P, for any $\varepsilon, \delta > 0$, if $n \gtrsim \frac{d^2}{\varepsilon^2} \log \frac{d}{\delta}$, then $\widehat{T}$ returned by Algorithm 5 satisfies*

$$D_{\mathrm{KL}}(P \parallel P_{\widehat{T}}) \le \min_{T \in \mathcal{T}} D_{\mathrm{KL}}(P \parallel P_T) + \varepsilon,$$

with probability at least $1 - \delta$.

When P is Markov to a tree (i.e. it is tree-structured), then the sample complexity improves:

Theorem 4.2.2. *Let T^* be a directed tree and P_{T^*} be a T^*-structured Gaussian. Given n i.i.d. samples from P_{T^*}, for any $\varepsilon, \delta > 0$, if $n \gtrsim \frac{d}{\varepsilon} \log \frac{d}{\delta}$, then $\widehat{T}$ returned by Algorithm 5 satisfies*

$$D_{\mathrm{KL}}(P_{T^*} \parallel P_{\widehat{T}}) \le \varepsilon,$$

with probability at least $1 - \delta$.

Remark: We can also obtain a sample-efficient algorithm for bounded-degree gaussian *polytrees*, using the guarantees of the estimator $\widehat{I}$, assuming that the skeleton is known. We defer the description of this result to Appendix B.2.5.

4.2.2 Distribution Learning Lower Bounds

The main idea of our proof is to reduce a distribution testing problem to our problem. Intuitively, the distribution testing problem is defined as follows. Suppose $R^{(1)}$ and $R^{(2)}$ are two distributions whose $D_{\mathrm{KL}}(R^{(1)} \parallel R^{(2)})$ is small. We are given n i.i.d. samples drawn from a distribution P where P is a m-variate distribution and each coordinate is distributed as either $R^{(1)}$ or $R^{(2)}$ uniformly and indepedently. Our task is to determine which of $R^{(1)}$ or $R^{(2)}$ the samples are drawn from correctly for at least $m/2$ coordinates. The formal definition will be presented in problem B.2.4. When $D_{\mathrm{KL}}(R^{(1)} \parallel R^{(2)})$ is sufficiently small, one should expect that n needs to be large enough to solve this problem with probability $2/3$. Hence, we construct the $(R^{(1)}, R^{(2)})$ pairs for the non-realizable and realizable case accordingly.

Theorem 4.2.3. *Suppose P is an unknown Gaussian distribution. Given n i.i.d. samples drawn from P. For any small $\varepsilon > 0$, if $n = o(d^2/\varepsilon^2)$, then for any estimator $\widehat{T}$, the maximum probability of achieving the required accuracy over a hard family of distribution $\mathcal{P}$ is bounded, such that:*

$$\inf_{\widehat{T}} \sup_{P \in \mathcal{P}} \Pr\left(D_{\mathrm{KL}}(P \parallel P_{\widehat{T}}) \leq \min_{T \in \mathcal{T}} D_{\mathrm{KL}}(P \parallel P_T) + \varepsilon \right) \leq 2/3$$

Theorem 4.2.4. *Suppose P is an unknown Gaussian distribution such that there exists a directed tree T^* that P is T^*-structured, i.e. $P = P_{T^*}$. Given n i.i.d. samples drawn from P. For any small $\varepsilon > 0$, if $n = o(d/\varepsilon)$, then for any estimator $\widehat{T}$, the maximum probability of success over a hard family of distribution $\mathcal{P}$ is bounded, such that*

$$\inf_{\widehat{T}} \sup_{P \in \mathcal{P}} \Pr\left(D_{\mathrm{KL}}(P \parallel P_{\widehat{T}}) \leq \varepsilon \right) \leq 2/3$$

4.3 Optimal Faithful Tree Learning

In the preceding section, we learned a tree-structured distribution under the KL distance, without concern for the learned tree structure. This viewpoint primarily pertains to *distribution learning*. This section adopts an different approach, emphasizing the aspect of *structure learning*. Specifically, we assume the underlying graph structure is indeed a tree, more generally, a polytree. We introduce an estimator based on the classic PC algorithm [SG91] and analyze its sample complexity under faithfulness. Crucially, we provide a matching lower bound to conclude the minimax optimality of the algorithm, which offers insights into the difficulty of structure learning under faithfulness.

4.3.1 Tree-Faithfulness

As alluded to in Chapter 2, the tree structure allows us to relax the usual notion of faithfulness:

Definition 4.3.1 (Tree-faithfulness)**.** *We say distribution P is tree-faithful to a polytree T if*

1. *For any two nodes connected $X_j - X_k$, we have $X_k \not\perp\!\!\!\perp X_j \,|\, X_\ell$ for all $\ell \in V \cup \{\emptyset\} \setminus \{k, j\}$;*
2. *For any v-structure $X_k \to X_\ell \leftarrow X_j$, we have $X_k \not\perp\!\!\!\perp X_j \,|\, X_\ell$.*

Tree-faithfulness comprises two components, each corresponding to adjacency-faithfulness and orientation-faithfulness respectively in restricted faithfulness (cf. Definition 2.3.3). In comparison to adjacency-faithfulness, tree-faithfulness solely requires conditional dependence for neighbouring nodes with conditioning sets of size at most one. Likewise, compared to orientation faithfulness, tree-faithfulness only needs conditional dependence for v-structures given the the collider. Let $\rho(X_j, X_k \,|\, X_\ell)$ be the conditional correlation coefficient between X_k and X_j given X_ℓ. As usual, in order to establish uniform, finite-sample results, we need the following concept of c-strong tree-faithfulness:

Definition 4.3.2 (c-strong tree-faithfulness)**.** *We say that P is c-strong tree-faithful to a polytree T if*

Algorithm 4: `PC-Tree` algorithm

Input: n i.i.d. samples $(X_1^{(i)}, \ldots, X_d^{(i)})$

1. Let $\widehat{E} = \emptyset$.
2. For each pair (j, k), $0 \leq j < k \leq d$:
   a) For all $\ell \in [d] \cup \{\emptyset\} \setminus \{j, k\}$:
      i. Test $H_0 : X_j \perp\!\!\!\perp X_k \,|\, X_\ell$ vs. $H_1 : X_j \not\perp\!\!\!\perp X_k \,|\, X_\ell$, store the results.
   b) If all tests reject, then $\widehat{E} \leftarrow \widehat{E} \cup \{j - k\}$.
   c) Else (if some test accepts), let $S(j, k) = \{\ell \in [d] \cup \{\emptyset\} \setminus \{j, k\} : X_j \perp\!\!\!\perp X_k \,|\, X_\ell\}$.

Output: $\widehat{T} = ([d], \widehat{E})$, separation set S.

1. *For any two nodes connected $X_j - X_k$, we have $\rho(X_k, X_j \,|\, X_\ell) \geq c$ for $\ell \in V \cup \{\emptyset\} \setminus \{k, j\}$;*
2. *For any v-structure $X_k \to X_\ell \leftarrow X_j$, we have $\rho(X_k, X_j \,|\, X_\ell) \geq c$.*

Under strong tree-faithfulness, we can now establish how the sample complexity depends on both the dimension d and the signal strength c.

4.3.2 Structure Learning Upper Bounds

We develop the `PC-Tree` algorithm for learning polytrees as a modification to the classic PC algorithm, outlined in Algorithm 4, effectively identifying the polytree's skeleton. An important by-product is the separation set resulted from the CI testing, which is used to obtain the CPDAG by applying an `ORIENT` step (Algorithm 14) as in the original PC algorithm.

In contrast to the original PC algorithm, `PC-Tree` distinguishes itself in two key aspects. Firstly, when assessing the presence of an edge between any two nodes, instead of exploring all potential conditioning sets, `PC-Tree` simplifies the process by exclusively testing marginal independence and conditional independence given only one other node. Furthermore, a notable departure from the original PC algorithm is that `PC-Tree` combines marginal independence tests and conditional independence tests, as opposed to ignoring the latter once marginal independence is established. `PC-Tree` will rely on sample (conditional) correlation coefficient for all

the (conditional) independence tests when running the algorithm, see more details in Appendix B.3.1.

Now we are ready to provide the sample complexity of `PC-Tree` in the following theorem, whose proof is postponed to Appendix B.3.2 and B.3.3.

Theorem 4.3.3. *For any $T \in \widetilde{\mathcal{T}}$, assuming P is c-strong tree-faithful to T, applying Algorithm 4 with sample correlation for CI testing, if the sample size*

$$n \gtrsim \frac{1}{c^2}\left(\log d + \log(1/\delta)\right),$$

then $\Pr(\widehat{T} = \mathrm{sk}(T)) \geq 1 - \delta$, *and* $\Pr(\textsc{Orient}(\widehat{T}, S) = \overline{T}) \geq 1 - \delta$

We may compare this upper bound $(\log d)/c^2$ with some of existing results on structure learning. Compared to learning equal variance general DAGs [GTA22] with optimal rates being $q \log(d/q)$, tree structure simplifies the problem by removing the factor of in-degree q. As against recovering undirected graph in MRF [MVL20], whose optimal sample complexity is $(s \log d)/\kappa^2$, we are able to improve the rate by the maximum degree s. Moreover, considering directed trees $T \in \mathcal{T} \subset \widetilde{\mathcal{T}}$, Lemma 18 shows c to be a constant under mild assumption on the parametrization of Equation 8.9, which assures possible concern of dependence on c.

4.3.3 Structure Learning Lower Bounds

Having provided the sample complexity upper bound, we continue to derive a matching lower bound:

Theorem 4.3.4. *Assuming c-strong tree-faithfulness, and $c^2 \leq 1/5$, $d \geq 4$, if the sample size is bounded as*

$$n \leq \frac{1 - 2\delta}{8} \times \frac{\log d}{c^2},$$

then for any estimator $\widehat{T}$ for $\overline{T}$,

$$\inf_{\widehat{T}} \sup_{\substack{T \in \widetilde{\mathcal{T}} \\ P \text{ is } c\text{-strong} \\ \text{tree-faithful to } T}} \Pr(\widehat{T} \neq \overline{T}) \geq \delta - \frac{\log 2}{\log d}.$$

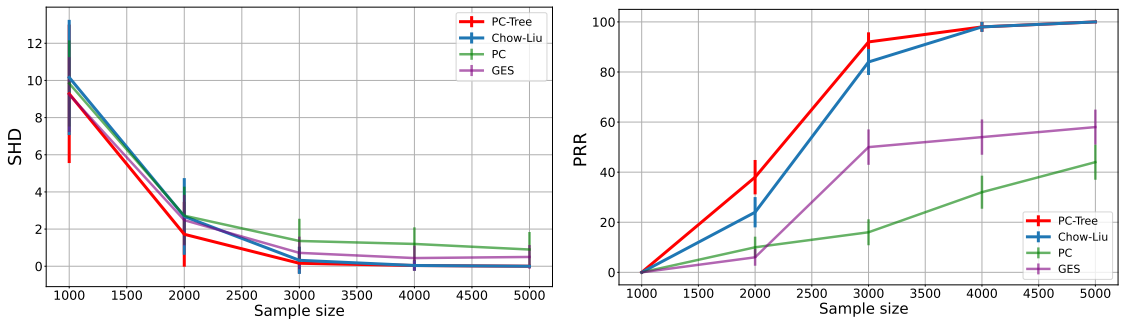
The lower bound in Theorem 4.3.4 implies the optimal sample complexity is $\Theta(\log d/c^2)$, where the dependence on $1/c^2$ term characterizes the hardness from

"how (Tree-)faithful" the distribution is; and $\log d$ term comes from the cardinality of all polytrees, which is much smaller compared to number of all DAGs.

To prove this lower bound, we employee Fano's inequality [Yu97] and consider a subclass of $\mathcal{T}$ to exploit the property that any node in directed tree has at most one parent. This subclass of directed trees has large enough cardinality by Cayley's formula of undirected trees. With the parametrization of edge weights appropriately calibrated, we show the KL divergence between the distributions consistent with any two instances from the subclass is well controlled, which leads to the final lower bound. The detailed proof can be found in Appendix B.3.4.

Remark. The optimality results in this section also extend to directed tree, polyforest and Markov chain. Since the lower bound is constructed using directed trees, the optimality applies. For polyforest, which is essentially polytree but allows for disconnected component, `PC-Tree` algorithm is able to identify the correct skeleton. On the other hand, polytree is a subclass of polyforest, thus the lower bound in Theorem 4.3.4 applies. For Markov chain, the algorithm is modified to dismiss marginal independence test, and the lower bound construction considers all Markov chains with the same way of parametrization as in Theorem 4.3.4. All these graphical models share the optimal sample complexity $\Theta(\log d/c^2)$.

4.4 Experiments



(a) SHD comparison (b) PRR comparison

Figure 4.1: Performance comparison for PC-Tree, Chow-Liu, PC, and GES algorithms evaluated on SHD and PRR. The red, blue, green, and purple lines represent PC-Tree, Chow-Liu, PC, and GES, respectively.

We conduct experiments to verify our findings in structure learning. While the main theorems assume Gaussian distributions, the robustness of our PC-Tree algorithm against non-Gaussian noise (Uniform and Laplace distributions) is empirically analyzed in Appendix B.4 and Appendix B.2. For brevity, we report here only the most difficult setting with $d = 100$ nodes; full details on the experiments and additional setups, e.g. when noise η_k is not Gaussian, can be found in Appendix C.3. We simulated random directed trees and synthetic data via Equation 8.9. We compare the performance of PC-Tree, Chow-Liu to PC and GES as classical baselines when only faithfulness assumed. Though Chow-Liu algorithm aims for distribution learning, it also estimates the skeleton as a byproduct. Therefore, to make fair comparison, we evaluate them by the accuracy of skeleton of the outputs (of PC-Tree, PC and GES). The results on average Structural Hamming Distance (SHD) and the Precise Recovery Rate (PRR) are reported in Fig. 4.1, where PRR measures the relative frequency of exact recovery. From the figure, we can see PC-Tree algorithm does perform the best, especially the significantly better result on PRR over the baselines, which is the main metric we are concerned with and have established optimality for. The competitive performance of Chow-Liu is also noticeable, for which we have not analyzed under the goal of structure learning, and we conjecture a similar sample complexity is shared with PC-Tree.

4.5 Comparison and Discussion

The literature on distribution learning and structure learning have largely evolved separate from one another. An interesting aspect of our results is that they consider both problems in a unified setting, allowing for an explicit comparison of these problems.

First, it is clear that the non-realizable setting should not be compared to structure learning, since in the former setting there is no structure to speak of. In the realizable setting, however, it is reasonable to ask for a comparison. Comparing Equation 4.2 and Equation 4.3, it is easy to see that there is a phase transition when $\varepsilon \asymp dc^2$. Focusing on directed trees for an apple-to-apple comparison, if the SEM parameters, e.g. β_k, σ_k^2 in Equation 8.9 are bounded, then strong tree-faithfulness holds with $c \asymp 1$, see Lemma 18. In this case, the optimal sample complexity for

structure learning is $\log d$ and $(d \log d)/\varepsilon$ for distribution learning, which has an additional factor of d/ε. Thus, as long as $\varepsilon = o(d)$, which is typical, structure learning is easier than distribution learning.

Another interesting scenario arises when $\varepsilon \ll dc^2$: Here, distribution learning is harder, however, we might hope to learn the structure of T "for free" by first learning the distribution to within KL accuracy ε. This is because, as ε goes to zero, $\widehat{P}$ converges to P, which implies we can use $\widehat{P}$ directly to estimate partial correlations for structure learning. Then the question boils down to whether there exists a good estimator of the structure that exploits $\widehat{P}$ when $\varepsilon \ll dc^2$. Lemma 19 shows that as long as the estimator is agnostic to $\widehat{P}$ (in the sense that it treats $\widehat{P}$ as a black-box), then we must have at least $\varepsilon \ll c^2$. Thus, there is a regime $c^2 \ll \varepsilon \ll dc^2$ where distribution learning does not automatically imply structure learning, at least in general. It remains as an interesting open question how small ε must be for $\widehat{P}$ to be efficiently used for structure learning, or whether or not there exist *specific* estimators $\widehat{P}$ that can be used for structure learning when $c^2 \ll \varepsilon \ll dc^2$.

Extending these results beyond the Gaussians we consider here (and finite alphabets as in previous work) is a promising direction for future research. Especially interesting would be bounds in a non-parametric setting.

Chapter 5

Testing Mutual Information Optimally in Linear Models

5.1 Introduction

Independence is a fundamental concept in statistics and many related fields, underlying various statistical analyses and methodologies. Conditional independence (CI) testing is a critical problem with widespread applications, such as economics [SGS01; SW07], machine learning [Sen+17; Shi+21], graphical models [GP93; Lin+14; Bor10], causal discovery [Zha+12; Zha+17; DSZ22] and causal inference [Spi10; Pea10; CD17]. These applications extend to numerous domains, including medical [MH59], genetics [LW08; Gio+14], and finance [WH18], where understanding the relationships between variables is crucial.

While testing independence and estimating dependence asymptotically are well-established practices, with roots tracing back to correlation analysis [Pea20; Sti89], more measure of dependence have been developed and studied (see [SW81; LG96; DM01; ST03; SW07; SW08; Son09; GS10; Hua10; BT14; WH18] and references therein). Among these, mutual information stands out as one of the attractive measures of dependence [Bel62; Ste+02; GVG15; GG21]. The theoretical results in these works are asymptotic, while the finite sample performance of their proposed testers is evaluated through simulations. However, the field of independence testing with optimal sample complexity guarantees remains relatively underexplored. Moreover, a significant body of research has concentrated on independence testing under Gaussianity assumption [Mor+09; KSG04; Wan+24]. In this scenario, indepen-

dence testing simplifies to determining whether certain partial correlations between variables are zero. The Gaussian assumption offers a simplified route to CI testing, as partial correlations are relatively straightforward to estimate. However, in non-Gaussian models, this approach can yield deceptive conclusions since variables may remain conditionally dependent despite exhibiting zero partial correlation.

This work addresses these gaps by developing an optimal nonparametric mutual information tester among the simplest linear structural equation models with unknown additive noise. Perhaps surprisingly, despite the simplicity of linear models, the optimal sample complexity of testing nonparametric mutual information in linear models has remained an open problem.

The major challenge is that without knowing the distribution of X, Y under some mild assumption, can we still distinguish the null hypothesis $X \perp\!\!\!\perp Y$ from alternative $X \not\perp\!\!\!\perp Y$ with optimal sample complexity?

5.1.1 Problem Definition

We now formally define our problem. Let $\mathcal{G}$ be the class of distributions on $\mathbb{R}$ whose pdf is twice diffierentiable, zero-mean, log-concave and sub-Gaussian. Let $\mathcal{P}$ be the class of distributions on $\mathbb{R}^2$ such that the random variable satisfies

$$\begin{aligned} X &= \eta_X \\ Y &= \alpha X + \eta_Y \end{aligned} \qquad \text{or} \qquad \begin{aligned} Y &= \eta_Y \\ X &= \alpha Y + \eta_X \end{aligned} \tag{5.1}$$

where α is an unknown coefficient and η_X and η_Y are independent random variables in $\mathcal{G}$.

Suppose we are given n samples drawn from a distribution in $\mathcal{P}$. The goal is to determine whether X and Y are independent by using these n samples. We assume that the mutual information between X and Y, $I(X;Y)$, is bounded below by ε if $I(X;Y) \neq 0$. Hence, we would like to minimize the number of samples, n, in order to achieve this goal in terms of ε.

Furthermore, we define the following model. A tree is an undirected graph in which any two nodes are connected by exactly one path. A directed tree is a directed graph in which, for some root node u, and any other node v, there is exactly one directed path from u to v. Let $\mathcal{T}$ be the set of directed trees on d nodes. For any directed tree $T \in \mathcal{T}$, a skeleton of T is a tree whose edge set is same as in T by

removing the direction. For any directed tree $T \in \mathcal{T}$ and any node v in T, a node u in T is the parent node of v if there is a directed edge $u \to v$ and a node w in T is a descendant of V if there is a directed path from v to w. For any $T \in \mathcal{T}$, let $\mathcal{P}_T$ be the set of distributions on $\mathbb{R}^d$ such that the random variable $(X_1, \ldots, X_d)$ satisfies

$$X_i = \begin{cases} \eta_i & \text{if } i \text{ is the root node} \\ \alpha_i X_{\mathrm{pa}(i)} + \eta_i & \text{if } i \text{ is not the root node} \end{cases} \tag{5.2}$$

where $\mathrm{pa}(i)$ is the parent node of i in T, α_i is unknown coefficient and $\{\eta_1, \ldots, \eta_d\}$ is a set of independent random variables in $\mathcal{G}$. Recall that $\mathcal{G}$ is the class of distributions on $\mathbb{R}$ whose pdf is twice diffierentiable, zero-mean, log-concave and sub-Gaussian.

Suppose we are given n samples drawn from a distribution in $\mathcal{P}_{T^*}$ for some unknown $T^* \in \mathcal{T}$. Our goal is to design an algorithm to find out a skeleton of T^* by using these n samples. We assume that, for any three nodes X, Y, Z in T^* where Z is a descendant of Y, the conditional mutual information of Y and Z given X, $I(Y; Z \mid X)$, is bounded below by ε if $I(Y; Z \mid X) \neq 0$. Hence, we would like to minimize the number of samples, n, in order to achieve this goal in terms of ε.

5.1.2 Our Contributions

We summarize our contributions in this paper as follows:

- We propose a new mutual information tester that takes $\frac{1}{\varepsilon}$ samples drawn from the model in Equation 5.1 to determine whether X and Y are independent. One may notice that if we manage to estimate the mutual information within a good error margin, then we will be able distinguish these two cases. However, we would like to emphasize that we are *not* estimating the mutual information $I(X; Y)$. Our novelty lies in bypassing the estimation of mutual information for this testing task. (Theorem 5.2.2)

- We deploy this mutual information tester in the Chow-Liu algorithm. Specifically, we build a weighted complete graph by using this mutual information tester as weights and apply the Chow-Liu algorithm to recover the underlying tree structure when we are given samples drawn from the model in Equation 5.2. Our novelty lies in applying the Chow-Liu algorithm without using the estimate mutual information as weights. We then show that one

only need $\frac{1}{\varepsilon}\log d$ samples to achieve this goal by extending our anylsis for our mutual information tester to a new conditional mutual information tester. (Theorem 5.3.3)

5.1.3 Other Related Work

The main technical contribution of this work is a new *mutual information tester* which can test independence without knowing the distribution. This falls in the general framework of *distribution property testing* [HK07; Rub12; Gol17; Eve+17; Can20; BY22].

It is noteworthy that conditional independence tests are well understood for discrete variables. For example, [Can+18] show that for three random variables X, Y, Z each pover Σ, testing if $I(X;Y \mid Z)$ is 0 or $\geq \varepsilon$ is possible with $\tilde{O}(|\Sigma|^3/\varepsilon)$ samples. Another solved case is that of linear models with Gaussian noise [Wan+24] (see also Chapter 4). In view of recent results on conditional independence testing based on generalized covariance measure (GCM) in [SP20], one cannot hope to design non-trivial tests. Conditional independence cannot be tested without additional assumptions about the distribution. Developing a (conditional) independence tester with respect to mutual information with $o(1/\varepsilon^2)$ sample complexity was posed as an open problem in [Can+18]. In [Can+18], they have shown that both Ising model *Goodness-of-fit Testing* and Ising model *Independece Tesing* can be solved from $\text{poly}(d, 1/\varepsilon)$ samples in polynomial time.

Furthermore, for distribution learning over finite alphabets, [DP20; DP21] showed that d-variable tree-structured Ising models can be learned computationally-efficiently from an optimal $O(d\log d/\varepsilon^2)$ samples. [Bha+21] derived explicit sample complexity bounds for the Chow-Liu algorithm of $\widetilde{O}(d\varepsilon^{-1})$ for trees on d vertices, and $d^2\varepsilon^{-2}$ samples for a general distribution P, and [Cho+23] further extend [Bha+21] into d-polytree. Beside, for structure learning over finite samples, we mention prior work that considers consistency and upper bounds under faithfulness [KB07; NHM+18; REB+18], and recent progress on learning polytrees [GA21; ATB21; TMD22; Jak+22].

5.2 Mutual Information Tester

In this section, we introduce our new mutual information tester. Our new mutual information tester takes $O(1/\varepsilon)$ samples of the form in Equation 5.1 as input and returns whether X and Y are independent. To achieve this, we first show the mutual information of X and Y, $I(X;Y)$, of the form in Equation 5.1 is at most α^2 (Theorem 5.2.1). With this and the assumption that $I(X;Y) \geq \varepsilon$ if $I(X;Y) \neq 0$, it is intuitively equivalent to approximating α up to $\sqrt{\varepsilon}$ error. By the standard concentration bound, we only need $(1/\sqrt{\varepsilon})^2 = 1/\varepsilon$ samples to achieve this error. Therefore, by our mutual information tester, we can distinguish whether X and Y are independent using $1/\varepsilon$ samples (Theorem 5.2.2). The detailed proofs are deferred to Appendix C.

For the purpose of analysis, WLOG, we assume

$$\begin{aligned} X &= \eta_X \\ Y &= \alpha X + \eta_Y. \end{aligned} \tag{5.3}$$

We first have the following notations. For any random variable R, let σ_R^2 be $\mathbf{E}(R^2)$, e.g. we have

$$\sigma_X^2 = \mathbf{E}(X^2), \quad \sigma_Y^2 = \mathbf{E}(Y^2), \quad \sigma_{\eta_X}^2 (= \sigma_X^2) = \mathbf{E}(\eta_X^2) \quad \text{and} \quad \sigma_{\eta_Y}^2 = \mathbf{E}(\eta_Y^2).$$

For any random variables R, S, let ρ_{RS} be $\mathbf{E}(RS)$, e.g. we have $\rho_{XY} = \mathbf{E}(XY)$. Hence, it is easy to see that $\alpha = \rho_{XY}/\sigma_X^2$. Suppose $(X^{(1)}, Y^{(1)}), \ldots, (X^{(n)}, Y^{(n)})$ are i.i.d. samples of (X, Y) of the form in Equation 5.1. We use $\widehat{\cdot}$ to denote the empirical version of the term unless otherwise specified, e.g. we have

$$\begin{aligned} &\widehat{\sigma}_X^2 = \widehat{\sigma}_{\eta_X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X^{(i)})^2, \quad \widehat{\sigma}_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y^{(i)})^2, \quad \widehat{\rho}_{XY} = \frac{1}{n}\sum_{i=1}^{n} X^{(i)}Y^{(i)}, \\ &\widehat{\alpha} = \frac{\widehat{\rho}_{XY}}{\widehat{\sigma}_X^2} \quad \text{and} \quad \widehat{\sigma}_{\eta_Y}^2 = \widehat{\sigma}_Y^2 - \widehat{\alpha}^2\widehat{\sigma}_X^2. \end{aligned}$$

Theorem 5.2.1. *Let (X, Y) be the random variable of the form in Equation 5.1. Assume that α is bounded above by a constant, i.e. $\alpha = O(1)$. Then, we have*

$$I(X;Y) \leq O\Big(\sigma_X^2\alpha^2\Big).$$

To prove Theorem 5.2.1, the main idea is to express the mutual information $I(X;Y)$ as a function of α. Recall that, WLOG, we assume Equation 5.3. By a straightforward calculation and the definition of $I(X;Y)$, we can express

$$I(X;Y) = -\int_{-\infty}^{\infty} f_Y(y,\alpha)\log f_Y(y,\alpha)dy + \int_{-\infty}^{\infty} f_{\eta_Y}(y)\log f_{\eta_Y}(y)dy$$

where $f_Y(\cdot,\alpha)$ is the pdf of Y and f_{η_Y} is the pdf of η_Y. Let F be

$$F(\alpha) := -\int_{-\infty}^{\infty} f_Y(y,\alpha)\log f_Y(y,\alpha)dy.$$

Note that when $\alpha = 0$, we have $Y = \eta_Y$ which means that

$$F(0) = -\int_{-\infty}^{\infty} f_{\eta_Y}(y)\log f_{\eta_Y}(y)dy \quad \text{and further implies} \quad I(X;Y) = F(\alpha) - F(0).$$

Since the pdfs involved are in $\mathcal{G}$, we can expand Taylor expansion for F at $\alpha = 0$. If we expand the Taylor expansion for F at $\alpha = 0$, we have

$$I(X;Y) \approx \frac{\partial F(0)}{\partial \alpha}\alpha + \frac{1}{2}\frac{\partial^2 F(0)}{\partial \alpha^2}\alpha^2.$$

We can indeed show that the first derivative at $\alpha = 0$, $\frac{\partial F(0)}{\partial \alpha}$, is zero and the second derivative, $\frac{\partial^2 F(0)}{\partial \alpha^2}$, is bounded. Then, the desired result follows.

With Theorem 5.2.1, we are now ready to define our mutual information tester. Define $\tilde{I}(X;Y)$ to be

$$\tilde{I}(X;Y) := -\log(1 - \frac{\widehat{\rho}^2_{XY}}{\widehat{\sigma}^2_X\widehat{\sigma}^2_Y}). \tag{5.4}$$

We would like to emphasize that $\tilde{I}(X;Y)$ is *not* an estimator of $I(X;Y)$. It is simply a tester for distinguish whether $I(X;Y) = 0$ or not. One may notice that another candidate for the mutual information tester can simply be $\frac{\widehat{\rho}^2_{XY}}{\widehat{\sigma}^2_X\widehat{\sigma}^2_Y}$. Indeed, when $a = \frac{\widehat{\rho}^2_{XY}}{\widehat{\sigma}^2_X\widehat{\sigma}^2_Y}$ is small, we have $-\log(1-a) \approx a$ which suggests that they are basically the same. However, we preview that, in the next section, we establish a connection between mutual information and condition mutual information via the chain rule, i.e.

$$I(X;Y) - I(X;Z) = I(X;Y \mid Z) - I(X;Z \mid Y).$$

The advantage of the current version is that it satisfies the "empirical version" of the chain rule, i.e.

$$\tilde{I}(X;Y) - \tilde{I}(X;Z) = \tilde{I}(X;Y \mid Z) - \tilde{I}(X;Z \mid Y) \quad \text{see Section 5.3 and Lemma 34 for the detail.}$$

Theorem 5.2.2 (Mutual Information Tester)**.** *Suppose we are given n i.i.d. samples of (X, Y) of the form in Equation 5.1. For any sufficiently small $\varepsilon, \delta > 0$ that $I(X;Y) \geq \varepsilon$ when $I(X;Y) \neq 0$, if $n = \Omega(\frac{1}{\varepsilon}\log\frac{1}{\delta})$, there exists a constant C such that the estimator defined in Equation 5.4 satisfies the following with probability $1-\delta$:*

- *If $I(X;Y) = 0$, then $\tilde{I}(X;Y) \leq \frac{\varepsilon}{100C}$.*
- *If $I(X;Y) \geq \varepsilon$, then $\tilde{I}(X;Y) \geq \frac{\varepsilon}{50C}$.*

To prove Theorem 5.2.2, we first rewrite our mutual information tester $\tilde{I}(X;Y)$ as

$$\tilde{I}(X;Y) = -\log\Big(1 - \frac{\widehat{\rho}^2_{XY}}{\widehat{\sigma}^2_X\widehat{\sigma}^2_Y}\Big) = \log\Big(1 + \frac{\widehat{\alpha}^2\widehat{\sigma}^2_X}{\widehat{\sigma}^2_{\eta_Y}}\Big).$$

Here, recall that

$$\widehat{\alpha} = \frac{\widehat{\rho_{XY}}}{\widehat{\sigma}^2_X} \quad \text{and} \quad \widehat{\sigma}^2_{\eta_Y} = \widehat{\sigma}^2_Y - \widehat{\alpha}^2\widehat{\sigma}^2_X$$

and the $\widehat{\cdot}$ mark simply means the empirical version of the term. Furthermore, when the term $\frac{\widehat{\alpha}^2\widehat{\sigma}^2_X}{\widehat{\sigma}^2_{\eta_Y}}$ is bounded, we have

$$\tilde{I}(X;Y) = \log\Big(1 + \frac{\widehat{\alpha}^2\widehat{\sigma}^2_X}{\widehat{\sigma}^2_{\eta_Y}}\Big) \approx \Theta(\frac{\widehat{\alpha}^2\widehat{\sigma}^2_X}{\widehat{\sigma}^2_{\eta_Y}}).$$

Now, we analyze our mutual information tester $\tilde{I}(X;Y)$ in two cases. When $I(X;Y) = 0$, it suggests that the term $\alpha = 0$ and hence, by the concentration bound, $\frac{\widehat{\alpha}^2\widehat{\sigma}^2_X}{\widehat{\sigma}^2_{\eta_Y}}$ is small, say $O(\varepsilon)$, which further implies $\tilde{I}(X;Y)$ is also smaller than $O(\varepsilon)$. When $I(X;Y) \geq \varepsilon$, by Theorem 5.2.1, it suggests that the term $\frac{\alpha^2\sigma^2_X}{\sigma^2_{\eta_Y}}$ is $\Omega(\varepsilon)$ away from 0 and hence, by the concentration bound, so is $\frac{\widehat{\alpha}^2\widehat{\sigma}^2_X}{\widehat{\sigma}^2_{\eta_Y}}$ which further implies $\tilde{I}(X;Y)$ is also $\Omega(\varepsilon)$ away from 0.

5.3 Application to Structure Learning

In this section, we extend our analysis for our mutual information tester to our conditional mutual information tester under certain conditions and deploy it in the Chow-Liu algorithm. Our approach is to build a weighted complete graph by

using our mutual information tester define in Equation 5.4 as the weight and apply the modified Chow-Liu algorithm (Algorithm 5) to return the maximum spanning tree. To show that this outputted tree is the skeleton of T^* (Theorem 5.3.3), we argue that intuitively if we do not recover an edge in T^*, the difference between the weight of the outputted tree and T^* is at least $\Omega(\varepsilon)$ caused by $I(Y;Z \mid X)$ for some three nodes X, Y, Z in T^* where Z is a descendant of Y. It turns out that we can modify our analysis for the mutual information tester in Section 5.2 and design a conditional mutual information tester to distinguish this ε error (Theorem 5.3.2). The detailed proofs are deferred to Appendix C.

As mentioned above, we often encounter $I(Y;Z \mid X)$ for some three nodes X, Y, Z in T^*. Note that one can always write X, Y, Z as follows.

$$\begin{aligned} X &= \xi_X \\ Y &= \beta X + \xi_Y \\ Z &= \lambda X + \gamma Y + \xi_Z \end{aligned} \tag{5.5}$$

for some coefficients β, λ, γ and some random variables ξ_X, ξ_Y, ξ_Z where $\mathbf{E}(X\xi_Y) = \mathbf{E}(X\xi_Z) = \mathbf{E}(Y\xi_Z) = 0$. It is easy to check that ξ_X, ξ_Y, ξ_Z is a linear combination of η_i defined in Equation 5.2. Also, it is known that twice differentiable, zero-mean, log-concave and sub-Gaussian are closed under linear combinations, i.e. $\xi_X, \xi_Y, \xi_Z \in \mathcal{G}$.

Theorem 5.3.1. *Let (X, Y, Z) be the random variable of the form in Equation 5.5 such that $\lambda = 0$ and Y and ξ_Z are independent. Assume that γ is bounded above by a constant, i.e. $\gamma = O(1)$. Then, we have*

$$I(Y;Z \mid X) \leq O\Big(\sigma_{\xi_Y}^2 \gamma^2\Big).$$

To prove Theorem 5.3.1, the idea is similar to the proof of Theorem 5.2.1 which is to express $I(Y;Z \mid X)$ as a function of γ. Recall that we write (X, Y, Z) in the form of Equation 5.5. Note that we assume $\lambda = 0$ and Y and η_Z are independent which is crucial in the detail analysis. By a straightforward calculation and the definition of $I(Y;Z \mid X)$, we can express

$I(Y;Z \mid X) = F(\gamma) - F(0)$ for some function F by s similar argument in Theorem 5.2.1.

Since the pdfs involved are in $\mathcal{G}$, we can expand the Taylor expansion for F at $\gamma = 0$. Hence, we have

$$I(Y;Z \mid X) \approx \frac{\partial F(0)}{\partial \gamma}\gamma + \frac{1}{2}\frac{\partial^2 F(0)}{\partial \gamma^2}\gamma^2.$$

We can indeed show that the first derivative at $\gamma = 0$, $\frac{\partial F(0)}{\partial \gamma}$, is zero and the second derivative, $\frac{\partial^2 F(0)}{\partial \gamma^2}$, is bounded. Then, the desired result follows.

With Theorem 5.3.1, we are now ready to define our conditional mutual information tester. Define $\tilde{I}(Y;Z \mid X)$ to be

$$\tilde{I}(Y;Z \mid X) := \log\left(1 + \frac{\widehat{\gamma}^2\widehat{\sigma}^2_{\xi_Y}}{\widehat{\sigma}^2_{\xi_Z}}\right). \tag{5.6}$$

Recall that we use $\widehat{\cdot}$ to denote the empirical version of the term. We again would like to emphasize that $\tilde{I}(Y;Z \mid X)$ is *not* an estimator of $I(Y;Z \mid X)$. It is simply a proxy for distinguish whether $I(Y;Z \mid X) = 0$ or not.

Theorem 5.3.2 (Conditional Mutual Information Tester)**.** *Suppose we are given n i.i.d. samples of (X,Y,Z) of the form in Equation 5.5. For any sufficiently small $\varepsilon, \delta > 0$ that $I(Y;Z \mid X) \geq \varepsilon$ when $I(Y;Z \mid X) \neq 0$, if $n = \Omega(\frac{1}{\varepsilon}\log\frac{1}{\delta})$, there exists a constant C such that the estimator defined in Equation 5.6 satisfies the following with probability $1-\delta$:*

- *If $I(Y;Z \mid X) = 0$, then $\tilde{I}(Y;Z \mid X) \leq \frac{\varepsilon}{100C}$.*
- *If $I(Y;Z \mid X) \geq \varepsilon$, $\lambda = 0$ and Y and ξ_Z are independent, then $\tilde{I}(Y;Z \mid X) \geq \frac{\varepsilon}{50C}$.*

To prove Theorem 5.3.2, the idea is similar to the proof of Theorem 5.2.2. We first observe that when the term $\frac{\widehat{\gamma}^2\widehat{\sigma}^2_{\xi_Y}}{\widehat{\sigma}^2_{\xi_Z}}$ is bounded we have

$$\tilde{I}(Y;Z \mid x) = \log\left(1 + \frac{\widehat{\gamma}^2\widehat{\sigma}^2_{\xi_Y}}{\widehat{\sigma}^2_{\xi_Z}}\right) \approx \Theta(\frac{\widehat{\gamma}^2\widehat{\sigma}^2_{\xi_Y}}{\widehat{\sigma}^2_{\xi_Z}}).$$

Now, we analyze our mutual information tester $\tilde{I}(Y;Z \mid X)$ in two cases. When $I(Y;Z \mid X) = 0$, it suggests that the term $\gamma = 0$ and hence, by the concentration bound, $\frac{\widehat{\gamma}^2\widehat{\sigma}^2_{\xi_Y}}{\widehat{\sigma}^2_{\xi_Z}}$ is small, say $O(\varepsilon)$, which further implies $\tilde{I}(Y;Z \mid X)$ is also smaller than $O(\varepsilon)$. When $I(Y;Z \mid X) \geq \varepsilon$, by Theorem 5.3.1, it suggests that the term $\frac{\gamma^2\sigma^2_{\xi_Y}}{\sigma^2_{\xi_Z}}$ is $\Omega(\varepsilon)$ away from 0 and hence, by the concentration bound, so is $\frac{\widehat{\gamma}^2\widehat{\sigma}^2_{\xi_Y}}{\widehat{\sigma}^2_{\xi_Z}}$ which further implies $\tilde{I}(Y;Z \mid X)$ is also $\Omega(\varepsilon)$ away from 0.

Algorithm 5: Modified `Chow-Liu` algorithm

Input: n i.i.d. samples $(X_1^{(i)}, \ldots, X_d^{(i)})$

1. We use j and X_j interchangeably in the subscript of $\widehat{\sigma}$ and $\widehat{\rho}$
2. For each $j = 1, \ldots, d$:
   - $\widehat{\sigma}_j^2 \leftarrow \frac{1}{n}\sum_{i=1}^n (X_j^{(i)})^2$
3. For each pair $(j,k), 1 \le j < k \le d$:
   - $\widehat{\rho}_{jk} \leftarrow \frac{1}{n}\sum_{i=1}^n X_j^{(i)} X_k^{(i)}$
4. For each pair $(j,k), 1 \le j < k \le d$:
   - $\tilde{I}(X_j; X_k) \leftarrow -\log\left(1 - \frac{\widehat{\rho}_{jk}^2}{\widehat{\sigma}_j^2 \widehat{\sigma}_k^2}\right)$
5. $G \leftarrow$ the complete undirected graph on $[d]$ whose edge weight for (j,k) is $\tilde{I}(X_j; X_k)$
6. $\widehat{T} \leftarrow$ the maximum spanning tree of G

Output: A tree $\widehat{T}$

Theorem 5.3.3. *Let $T^* \in \mathcal{T}$ be a directed tree. Suppose we are given n i.i.d. samples of $(X_1, \ldots, X_d)$ of the form in Equation 5.2. For any sufficiently small $\varepsilon, \delta > 0$ that $I(Y; Z \mid X) \ge \varepsilon$ when $I(Y; Z \mid X) \ne 0$ for any three nodes X, Y, Z in T^*, if $n = \Omega(\frac{1}{\varepsilon}\log\frac{d}{\delta})$, the tree outputted by Algorithm 5, $\widehat{T}$, is equal to the skeleton of T^* with probability $1 - \delta$.*

To prove Theorem 5.3.3, we analyze the difference in the edges sets of T^* and $\widehat{T}$. For the purpose of exposition, we assume that there is exactly one edge misidentified. Let $X \to Y$ be the edge in T^* and $W - Z$ be the edge in $\widehat{T}$ such that the order of them is $W - X \to Y \rightsquigarrow Z$. Here, $-$ means an undirected path in T^*, $\to$ means a directed edge in T^* and $\rightsquigarrow$ means a directed path in T^*. As mentioned before, the advantage of the specific form of our mutual information tester defined in Equation 5.4 is that it satisfies the "empirical version" of the chain rule, i.e.

$$\begin{aligned}
\tilde{I}(X;Y) - \tilde{I}(W,Z) &= \tilde{I}(X;Y) - \tilde{I}(X;Z) + \tilde{I}(X;Z) - \tilde{I}(W,Z) \\
&= (\tilde{I}(X;Y \mid Z) - \tilde{I}(X;Z \mid Y)) + (\tilde{I}(X;Z \mid W) - \tilde{I}(W;Z \mid X)) \\
&= (\tilde{I}(X;Y \mid Z) + \tilde{I}(X;Z \mid W)) - (\tilde{I}(X;Z \mid Y) + \tilde{I}(W;Z \mid X))
\end{aligned}$$

By the definition of the output of Algorithm 5, we have

$$\tilde{I}(X,Y) - \tilde{I}(W;Z) \leq 0.$$

which implies

$$\tilde{I}(X;Y \mid Z) + \tilde{I}(X;Z \mid W) \leq \tilde{I}(X;Z \mid Y) + \tilde{I}(W;Z \mid X) \tag{5.7}$$

Recall that we have the order $W - X \to Y \rightsquigarrow Z$ and hence

$$\begin{cases} I(X;Z \mid Y), I(W;Z \mid X) = 0 \\ I(X;Y \mid Z), I(X;Z \mid W) \geq \varepsilon \end{cases} \underset{\text{by Theorem 5.3.2}}{\Rightarrow} \begin{cases} \tilde{I}(X;Z \mid Y), \tilde{I}(W;Z \mid X) \leq \frac{\varepsilon}{50C} \\ \tilde{I}(X;Y \mid Z), \tilde{I}(X;Z \mid W) \geq \frac{\varepsilon}{25C} \end{cases}$$

By plugging them into Equation 5.7, we immediately have a contradiction and hence $\widehat{T}$ must be equal to the skeleton of T^*.

5.4 Experiments

We conducted experiments to verify our findings on the mutual information tester and its application for structure learning with the Chow-Liu algorithm in Algorithm 5. More details on the experiments can be found in Appendix C.3.

Mutual Information Tester For independence tests, we compared our mutual information tester with other mutual-information-based independence tests discussed in related work, including the KNN-based estimator [PS11] with Von Mises entropy estimation (KNNVM), and the recent VM-CI estimator from [JGK23], which is a nonparametric Von Mises estimator for the entropy (KDEVM) of multivariate distributions built on a kernel density estimator.

We conducted the experiments using synthetic data sampled from different distributions belonging to the distribution class $\mathcal{G}$. For brevity, we report here only the performance of our mutual information tester versus other baseline algorithms when given n samples drawn from $\mathcal{P}$ in Fig. 5.1, such as Gaussian, uniform, Beta, and Laplace distributions. For each value of n, we ran 1000 experiments: the first half with $X \perp\!\!\!\perp Y$ and the second half with $I(X,Y) \geq \varepsilon$.

These results illustrate that our mutual information tester demonstrates optimal performance compared to other methods. Specifically, for both Gaussian and

uniform distributions, the KDEVM algorithm struggles to distinguish between the null and alternative hypotheses using a constant threshold. Although the KNNVM algorithm can distinguish the two cases with a constant threshold, the gap between them is much smaller than with our method. Furthermore, even for Beta and Laplace distributions, where all baseline methods can distinguish between the two cases using a constant threshold, our tester still outperforms the others by producing a larger gap between the cases.

Structure Learning Additionally, by extending our mutual information tester to a conditional mutual information tester, we also conducted experiments in structure learning. Details are as follows:

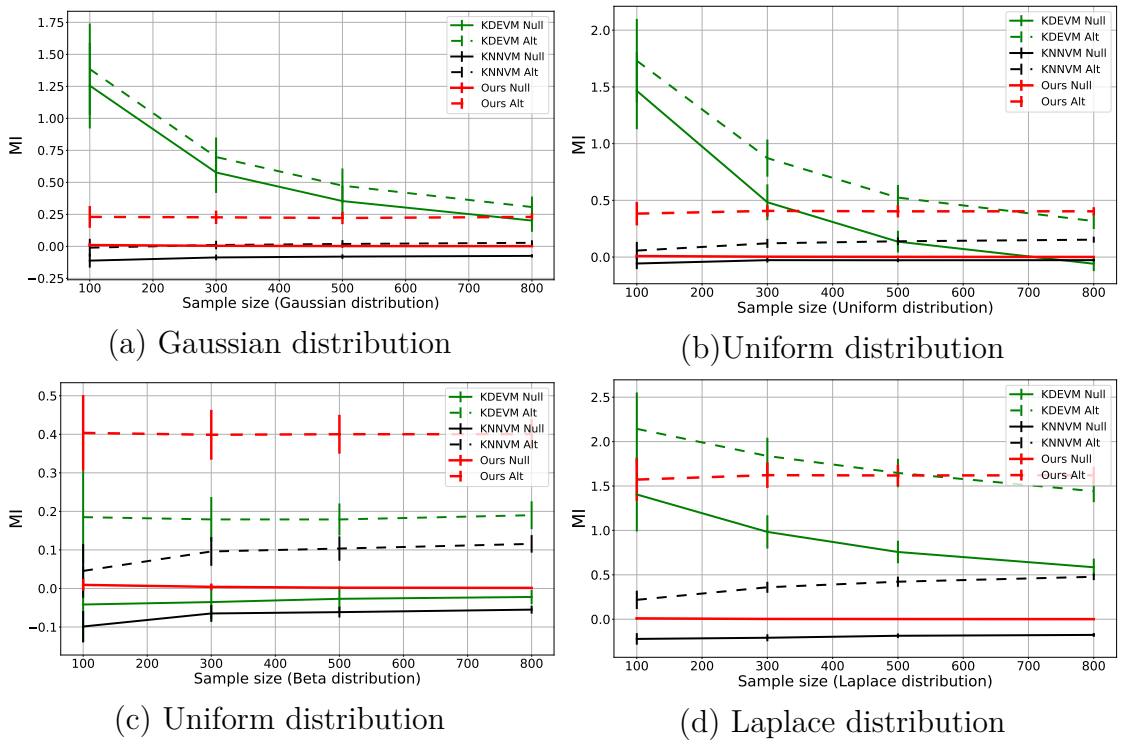
Synthetic Data Generation We generate trees using package `networkx`. We consider number of nodes $d = 100$. To generate the data as in Equation 5.3, we uniformly sample α_k from the interval $(-0.5, 0.1] \cup [0.1, 0.5)$ as our coefficient weight. For sample size $n = \{100, 500, 1000, 1500, 2000\}$, we generate our i.i.d. samples $X \in \mathbb{R}^{n \times d}$ according to Equation 5.3, where $\eta \sim \mathcal{G}$. Specifically, we present experiments where $\eta \sim \mathbb{N}(0, 1)$ is Gaussian distribution, or $\eta \sim \text{Laplace}(0, 1)$ is Laplace distribution.

Baselines We have employed two baseline algorithms: the PC algorithm has been executed using the Python package `Causal-learn`, while the GES algorithm has been implemented with `py-tetrad`.

Evaluation For each experiment setup, we report the average (over 50 random instantiations) Structural Hamming Distance (SHD) between the ground truth and our estimated graph skeleton, and the Precise Recovery Rate (PRR), which is the frequency of exact recovery of the tree skeleton. All experiments were conduced on an Intel Core i7-12800H 2.40GHz CPU.

The results on average Structural Hamming Distance (SHD) and Precise Recovery Rate (PRR) are reported in Fig. 5.2, where PRR measures the relative frequency of exact recovery. From the figure, we can see that the Chow-Liu algorithm performs the best for both Gaussian and Laplace distributions, especially achieving

significantly better PRR results than the baselines.



(a) Gaussian distribution

(b)Uniform distribution

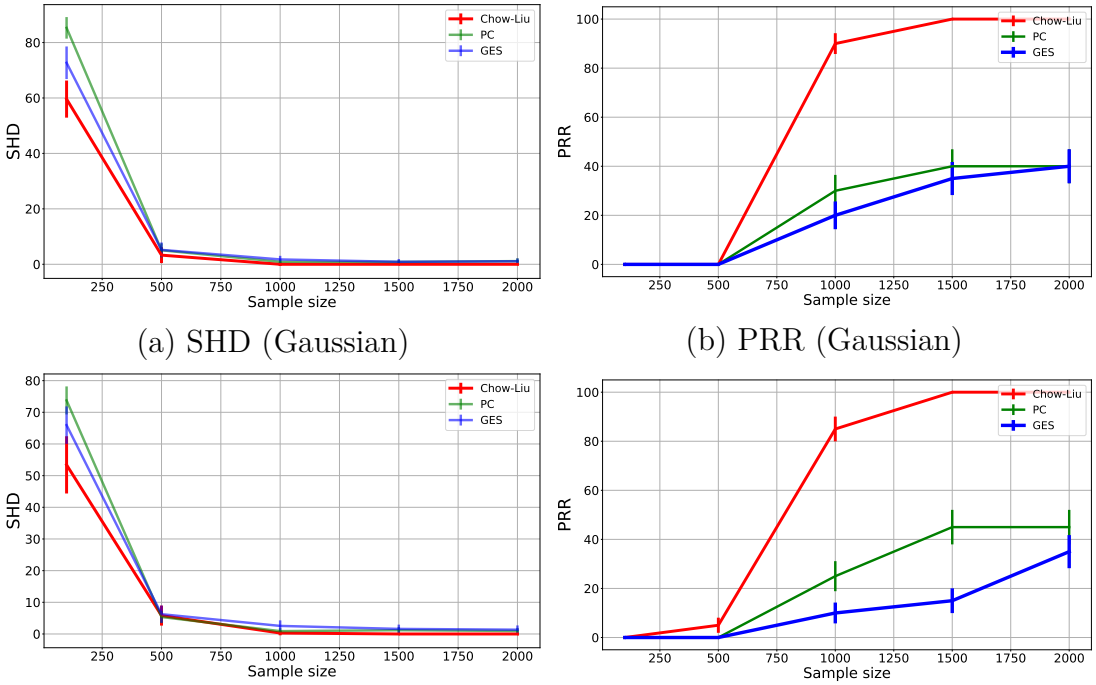
(c) Uniform distribution

(d) Laplace distribution

Figure 5.1: MI tester on null (solid line) and alternative (dashed line) hypotheses. The red, green, and black lines represent our methods, KDEVM, and KNNVM, respectively.

5.5 Conclusion and Future Work

This chapter first considered a linear model between two random variables X and Y with only mild assumptions on the noise. We propose a novel mutual information tester using only $\frac{1}{\varepsilon}$ (which is optimal) samples drawn from this model to determine whether X and Y are independent given that the mutual information of X and Y, $I(X;Y)$, is at least ε if $I(X;Y) \neq 0$. We bypass the standard way of estimating the mutual information and manage to accomplish this testing task. We then consider the model where the data are generated from a distribution who obeys a directed tree structure. Our task is to identify this tree structure and we deploy our new mutual information tester in the Chow-Liu algorithm. Then, our result suggests that one only need $\frac{1}{\varepsilon} \log d$ samples to accomplish this goal. Finally, our findings are demonstrated through empirical experiments.



(a) SHD (Gaussian) (b) PRR (Gaussian)

(c) SHD (Laplace) (d) PRR (Laplace)

Figure 5.2: Performance comparison for Chow-Liu, PC, and GES algorithms evaluated on SHD and PRR. The red, blue, and green lines represent Chow-Liu, PC, and GES, respectively.

For the future directions, we would like to extend our current setting to the following:

- Consider a more general relation between X and Y instead of linear models.
- Apply our methods to more general underlying graphical structures such as LiNGAM models.
- Use mutual information testing to determine the direction of the edge in the underlying graph.

Part II

Learning, Testing, and Inference from Biased Data

Chapter 6

Gaussian Mean Testing from Truncation

6.1 Introduction

The Gaussian mean testing problem, which originated in the context of signal processing under the name of signal detection, asks the following: given independent observations from a high-dimensional vector subject to random white noise, how to detect whether the underlying signal has large magnitude, or is non-significant? This can be seen as a hypothesis testing version of the so-called *Gaussian location model* (GLM) question from information theory and signal processing, where the objective is to *detect* a signal instead of *learning* it.

Mean testing has recently seen a surge of interest from the machine learning and theoretical computer science (and, specifically, *distribution testing*) communities, focusing on efficient algorithms with finite-sample guarantees, i.e., requiring as few observations (samples) as possible. This culminated in simple, sample-optimal algorithms for this task under an array of settings, including relaxing the assumption on the random noise [Can+21; DKP22], considering it in the distributed, communication-limited setting [ACT20; SVZ23], or requiring robustness to adversarial corruptions of the observations [Can+23].

In this chapter, we consider a different variant, and focus on the *truncated samples* setting. Truncation happens when some observations fail to be observed or recorded, e.g., due to limitations in the sensing equipment or, in the case of social studies or surveys, when a subset of respondents systematically withhold

their response. A typical example is when asking insurance customers for some sensitive medical information, as people with at-risk factors may decide to opt out of the survey entirely for fear of having their insurance premiums go up. Truncated samples (and the related notion of censored data) have a rich history in Statistics, and a host of applications in medical science, social studies, and Economics, to name a few (see, e.g., [Coh91a]); and, following [Das+18], has recently been the focus of a line of work on efficient truncated statistics, whereby one seeks to develop efficient algorithms to efficiently estimate the parameters of a population given truncated samples: we elaborate on this in Section 6.1.2.

Despite the existence of these two lines of work – one on Gaussian mean testing, and the other on learning parameters from truncated samples, to the best of our knowledge there has not been any study of the very natural related question of *Gaussian mean testing from truncated samples.* In this chapter, we address this question, and show that the complexity of the testing task changes drastically (and quite surprisingly) depending on the truncation set itself, and whether we have some *a priori* information about it. In order to present our results and discuss their implications, we start by formally defining the problem:

Problem formulation. Let $\boldsymbol{\mu} \in \mathbb{R}^d, \boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ be an unknown vector and covariance matrix, respectively, and $S \subseteq \mathbb{R}^d$, the *truncation set*, be a subset of measure at least $1 - \varepsilon$ under the spherical normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $0 \leq \varepsilon < 1$. We define the S-truncated Gaussian distribution, denoted $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; S)$, as the normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ conditioned on taking values on the subset S. We suppose that samples, $\mathbf{X} = \{\boldsymbol{x^{(1)}}, \ldots, \boldsymbol{x^{(n)}}\}$, from an unknown d-variate normal $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are only revealed if they fall into some subset $S \in \mathbb{R}^d$; otherwise the samples are hidden and their count in proportion to the revealed samples is also hidden. We will make no assumptions about S, except that its measure ε with respect to the unknown distribution is non-trivial, say $\varepsilon = 1\%$: that is, one should think of ε as a small (positive) constant. We will focus on the case of *spherical* covariance matrices (before truncation), that is, where $\boldsymbol{\Sigma} = \boldsymbol{I_d}$: this corresponds to the signal detection problem alluded to before, where a signal is observed through random white noise.

Given n i.i.d. samples $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots$ from a truncated Gaussian distribution P on $\mathbb{R}^d$ (with unknown vector $\boldsymbol{\mu}$ and truncation set S) and $\alpha \in (0, 1]$ an accuracy,

the task is to distinguish between the following cases:

- **(Completeness)** if $P = \mathcal{N}(0, \boldsymbol{I_d}; S)$, the algorithm must output "ACCEPT" with probability at least 2/3;
- **(Soundness)** if $P = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I_d}; S)$ for some $\boldsymbol{\mu}$ with $\|\boldsymbol{\mu}\|_2 \geq \alpha$, the algorithm must output "REJECT" with probability at least 2/3.

The objective is to minimize the *sample complexity* of the algorithm, i.e., the number of samples n required to achieve the task, over all possible vectors $\boldsymbol{\mu}$ and truncation sets S. Note that the complexity of the task might vary, depending on the parameter regime and the information available about S: namely, (1) the relation between truncated mass ε and desired accuracy α, and (2) whether the set S is unknown to the algorithm or known (either provided explicitly, or as a membership oracle.[1]

6.1.1 Our contributions

We establish upper and lower bounds on the sample complexity of the problem, and show it undergoes a stark transition as α and ε vary, when the truncation set is unknown to the algorithm. Specifically, we show the following, where, for ease of exposition, we focus on the dependence on the dimension d and treat ε, α as constants:

- When $\varepsilon\sqrt{\log 1/\varepsilon} < \alpha$, i.e., the accuracy parameter is significantly larger than the truncated probability mass, then the simple testing algorithm designed for the *non-truncated* version of the problem works, achieving the optimal sample complexity $\Theta(\sqrt{d})$ (Theorem 6.2.1).
- When $\varepsilon < \alpha < \varepsilon\sqrt{\log 1/\varepsilon}$, there is a sudden phase transition: we provide an information-theoretic lower bound showing that *any* algorithm requires $\Omega(d)$ samples (Lemma 4). Combined with an $O(d)$ upper bound obtained by *learning* the unknown mean vector $\boldsymbol{\mu}$, our results show that in this regime *testing suddenly becomes as hard as learning.*
- When $\alpha < \varepsilon$, it follows from [Das+18] that the testing task becomes information-theoretically impossible, regardless of sample complexity.

[1] A membership oracle for a set S is a procedure which, on any input x, indicates whether $x \in S$.

	$\varepsilon < \frac{\alpha}{\sqrt{\log \frac{1}{\alpha}}}$	$\frac{\alpha}{\sqrt{\log \frac{1}{\alpha}}} \leq \varepsilon < \alpha$	$\alpha \leq \varepsilon$
Unknown	$\Theta(\sqrt{d})$	$\Theta(d)$	∞
Known	$\Theta(\sqrt{d})$	$\Theta(\sqrt{d})$	$\Theta(\sqrt{d})$

Table 6.1: Mean testing sample complexity for small enough constant ε and α.

In contrast, we show that when the truncation set is known, a different (yet still relatively simple) algorithm, based on the gradient of the maximum likelihood estimator, achieves the optimal sample complexity $O(\sqrt{d})$, *across all parameter ranges* (Theorem 6.3.1).

6.1.2 Related Works

We here discuss the literature and previous related work.

Learning from Truncated or Censored Samples Distribution learning under censored, truncated mechanisms has had a long history. Censoring happens when the events can be detected, but the measurements (the values) are completely unknown, while truncation occurs when an object falling outside some subset are not observed, and their count in proportion to the observed samples is also not known, see [DV55; Coh57; Dix60; HS90; Coh91b; BS99; CCS13; CSV17] for an overview of the related works in estimating the censored or truncated normal or other type of distributions. [Pea02; PL08; Lee14] used the method of moments, while [Fis31] used the maximum likelihood approach for the distribution learning from truncated samples. Since then, [Das+18; Das+19; DRZ20] developed computationally and statistically efficient algorithms under the assumption that the truncation set is known. Furthermore, [WDS19] considered the problem of estimating the parameters of a d-dimensional rectified Gaussian distribution from i.i.d. samples. This can be seen as a special case of the self-censoring truncation, where the truncation happens due to the ReLU generative model.

Robust mean estimation Robust statistics [HR11] considers statistical inference problems under the setting where samples observed could be contaminated in various ways. For robust estimation, the usual goal is to obtain accurate estimation of parameters for parametric families such as Gaussian distributions under

ε-contamination, where ε is the maximum fraction of samples ($\varepsilon \cdot n$ out of n) allowed to be contaminated. This problem has been extensively studied in recent years (see the book of [DK23], and references therein). There are algorithms and lower bounds with different characteristics under different contamination models (time complexity and accuracy trade-off) [HL19; Bla+22; Dia+19]. [HLZ20] studies (nonparametric) robust mean estimation: distributions with finite covariance. Notably, using algorithms developed through robust mean estimation (also called learning) for Gaussian under some strong contamination model, we can reduce our testing under truncation problem via the standard learning-to-test argument, which will give us a sample complexity upper bound of $\mathcal{O}(d/\alpha^2)$.

Robust mean testing Gaussian mean testing has been studied and well known to have a sample complexity of $\Theta(\sqrt{d}/\alpha^2)$ [DKS17; DKP22]. Recently, [Can+23] studied the Gaussian mean testing problem under two contamination models: *oblivious contamination model* and *strong contamination model* – both yield improved sample complexity than their learning counterparts. In the oblivious contamination model, an adversary could remove ε fraction of original samples from P *without* observing them and replacing them with samples from a different distribution. In this model, [Can+23] prove a near-optimal sample complexity bound of $\tilde{\Theta}\left(\max\left(\frac{\sqrt{d}}{\alpha^2}, \frac{d\varepsilon^3}{\alpha^4}, \min\left(\frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}, \frac{d\varepsilon}{\alpha^2}\right)\right)\right)$.

In the strong contamination model, where the adversary could first observe the values of original samples from P, then pick ε fraction of them and replace with arbitrary values, [Can+23] gives the near-optimal sample complexity bound of $\tilde{\Theta}\left(\max\left(\frac{\sqrt{d}}{\alpha^2}, \frac{d\varepsilon^2}{\alpha^2}\right)\right)$.[2]

Indeed, truncation can be viewed as a special form of contamination model, and a strictly weaker form of contamination than the strong contamination model considered in [Can+23]. Yet, it is somewhat orthogonal (neither stronger or weaker) to the oblivious contamination model. We remark that our paper covers the full parameter regime in terms of the relation between ε and α, while [Can+23, Theorem 7.1] has a limitation in the $\alpha \geq \varepsilon \cdot \mathrm{polylog}(d, \frac{1}{\varepsilon}, \frac{1}{\alpha})$. Under the regime, $\alpha \ll \varepsilon \cdot \sqrt{\log \frac{1}{\varepsilon}}$, there is separation in sample complexity: $\Theta(\frac{\sqrt{d}}{\alpha^2})$ v.s. $\tilde{\Theta}\left(\max\left(\frac{\sqrt{d}}{\alpha^2}, \frac{d\varepsilon^2}{\alpha^2}\right)\right)$ between the truncation model and strong contamination model.

[2] We use $\tilde{(\cdot)}$ to hide the polylogarithmic factors.

6.2 Testing under Unknown Truncation

When the truncation set is unknown, we will focus on three possible regimes depending on the relation between the accuracy and truncation parameter:

- $\varepsilon \cdot \sqrt{\log(1/\varepsilon)} \ll \alpha$: in this case, the truncation size is much smaller than the required accuracy, meaning the change in the empirical mean after truncation is negligible (at most $\varepsilon \cdot \sqrt{\log(1/\varepsilon)}$. Therefore, applying the standard mean tester [DKP22] Algorithm 6 with a sample complexity of $\mathcal{O}(\sqrt{d}/\alpha^2)$ is sufficient.

- $\varepsilon \ll \alpha \ll \varepsilon \cdot \sqrt{\log(1/\varepsilon)}$: Here, the truncation size is close to the accuracy threshold. An adversarial truncation (knowing the true mean) can select a truncation set that shifts the truncated mean by at least $\Omega(\varepsilon \cdot \sqrt{\log(1/\varepsilon)})$. In this regime, we establish a lower bound of $\Omega(d/\varepsilon)$, indicating a transition in sample complexity from $\Theta(\sqrt{d})$ to $\Theta(d)$.

- $\alpha \ll \varepsilon$: When the truncation size exceeds the accuracy threshold, it has been shown that testing becomes information theoretically unfeasible [Das+18, Lemma 12].

Our contribution are in the first two regimes and we will elaborate on in the following subsections.

6.2.1 When Truncation Size is Much Smaller Than Accuracy $\varepsilon\sqrt{\log 1/\varepsilon} \lesssim \alpha$

In this subsection, we present Theorem 6.2.1. Given that the change in the expectation after truncation is minimal, it is sufficient to bound the change in both the mean and variance of the truncated normal distribution (as outlined in Lemma 3). We then apply the tester and analysis from [DKP22, Theorem 1.1]. As a result, it is sufficient to apply the standard mean tester in Algorithm 6 with a sample complexity of $\mathcal{O}(\sqrt{d}/\alpha^2)$.

Theorem 6.2.1. *There exists an algorithm (Algorithm 6) that, given i.i.d. samples from truncated Gaussian distribution P with an unknown support set $S \subset \mathbb{R}^d$, can*

distinguish the following two cases based on the truncation mass parameter $\varepsilon \in (0,1)$ *and the accuracy parameter* $\alpha > 0$*:*

- ***(Completeness)*** *If* P *is a truncated Gaussian distribution* $\mathcal{N}(\mathbf{0}, \mathbf{I}_d; S)$ *and the truncation mass satisfies* $1 - \mathcal{N}(0, \mathbf{I}_d; S) \leqslant \varepsilon$*, the algorithm will output "ACCEPT" with probability at least* $2/3$*.*
- ***(Soundness)*** *If* P *is a truncated Gaussian distribution* $\mathcal{N}(\mu, \mathbf{I}_d; S)$ *where* $\|\mu\|_2 \geqslant \alpha \geqslant c_1 \cdot \varepsilon\sqrt{\log \frac{1}{\varepsilon}}$ *for some constant* $c_1 > 0$ *and the truncation mass satisfies* $1 - \mathcal{N}(0, \mathbf{I}_d; S) \leqslant \varepsilon$*, the algorithm will output "REJECT" with probability at least* $2/3$*.*

The algorithm requires $\mathcal{O}\left(\frac{\sqrt{d}}{\alpha^2}\right)$ *samples from* P*.*

Algorithm 6: `GaussianMeanTester` [DKP22]

Input : Sample access to distribution P on $\mathbb{R}^d$ and $\alpha > 0$
Output : "ACCEPT" if $P = \mathcal{N}(\mathbf{0}, \mathbf{I}_d, S)$;
"REJECT" if $P = \mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d, S)$ and $\|\boldsymbol{\mu}\|_2 \geq \alpha$;
both with probability at least $2/3$

Set $n = \mathcal{O}(\sqrt{d}/\alpha^2)$;
Sample $2n$ i.i.d. points from P and denote them by $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$;
Define $Z = \frac{1}{n^2}\left(\sum_{i=1}^n X_i\right)^\top \left(\sum_{i=1}^n Y_i\right)$;
if $|Z| \leq \mathcal{O}(\sqrt{d}/n)$ **then**
  return "ACCEPT";
else
  return "REJECT";

We now provide the proof sketch of Theorem 6.2.1. Given $2n$ i.i.d. samples from a d-variate truncated normal $P \sim \mathcal{N}(\mu, \mathbf{I}_d, S)$, let the sample set be $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}, \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}\}$, where $\mathbf{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$, $\mathbf{Y} = \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}\}$. The measure of S under the non-truncated distribution $\mathcal{N}(\mu, \mathbf{I}_d)$ is at least $1 - \varepsilon$, where $0 \leq \varepsilon < 1$. Define the empirical means of the sample sets in $\mathbb{R}^d$ as

$$\bar{\mathbf{X}} := \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i, \qquad \bar{\mathbf{Y}} := \frac{1}{n}\sum_{i=1}^n \mathbf{Y}_i.$$

Our core test statistic is the inner product of these two empirical means:

$$Z = \langle \bar{\mathbf{X}}, \bar{\mathbf{Y}} \rangle \tag{6.1}$$

Let $\mu_S = \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mu,\mathbf{I}_d,S)}[\mathbf{x}]$ denote the mean of the truncated distribution, and let $\mathbf{\Sigma}_S = \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mu,\mathbf{I}_d,S)}[(\mathbf{x}-\mu_S)\cdot(\mathbf{x}-\mu_S)^T]$ be the covariance matrix under truncation.

Lemma 2. *For the random variable Z defined in Eq. (6.1), obtained from two independent sets of n samples (i.e. $2n$ total samples) from P, the following holds:*

$$\mathbb{E}[Z] = \langle \mathbb{E}[\bar{\mathbf{X}}], \mathbb{E}[\bar{\mathbf{Y}}] \rangle = \|\mu_S\|_2^2 \tag{6.2}$$

$$\mathrm{Var}[Z] \leqslant \frac{\|\mathbf{\Sigma}_S\|_F^2}{n^2} + \frac{2}{n}\|\mathbf{\Sigma}_S\|_F\|\mu_S\|_2^2 \tag{6.3}$$

Lemma 3 (Truncated vs non-truncated parameters)**.** *Let $\mu_S, \mathbf{\Sigma}_S$ be the mean and covariance of the truncated Gaussian $\mathcal{N}(\mu, \mathbf{I}_d; S)$ with a measure of at least $1-\varepsilon$. Then the following holds:*

$$\|\mu_S - \mu\|_2 \leq \mathcal{O}(\varepsilon \cdot \sqrt{\log(1/\varepsilon)}) \text{ and } \|\mathbf{\Sigma}_S - \mathbf{I}_d\|_F \leqslant \mathcal{O}(\sqrt{d}).$$

Using Lemma 2 and Lemma 3, we compute the expectation and variance of Z. In the completeness case, the quantity $|Z - \|\mu\|_2^2|$ is small, with $\mathbb{E}[Z] < \mathcal{O}(\alpha^2)$ and $\mathrm{Var}[Z] \lesssim \alpha^4$. In the soundness case, We can lower bound the expectation of μ_S for $\mathcal{N}(\mu, \mathbf{I_d}, S)$, where $\|\mu\|_2 \geqslant \alpha$, and show that $\mathbb{E}[Z] \geqslant \Omega(\alpha^2)$, and $\mathrm{Var}[Z] \lesssim \mathbb{E}^2[Z]$. This provides a clear separation between the two cases.

6.2.2 When Truncation Size is Near Accuracy $\varepsilon \lesssim \alpha \lesssim \varepsilon\sqrt{\log 1/\varepsilon}$

As the truncation mass ε approaches to α, the null and alternative hypothesis may overlap due to the non-negligible truncation size. This overlap occurs because it becomes possible to choose truncation regions that can substantially alter μ_S by an amount comparable to α, rendering the standard algorithm ineffective. Surprisingly, it presents a much greater challenge for our testing problem, where the sample complexity escalates to $\Omega(d)$, matching that of the existing robust learning algorithms [DK23, Proposition 1.20].

Theorem 6.2.2. *The sample complexity for truncated mean testing when $\varepsilon \lesssim \alpha \lesssim \varepsilon \cdot \sqrt{\log \frac{1}{\varepsilon}}$ is $\Theta(d)$.*

Mean Testing Lower Bound We now show the main idea of our lower bound proof. Intuitively, the hard instance constructed in Lemma 4 does exactly this: it modifies the mean by α and selects a random unit vector v to define its direction in $\mathbb{R}^d$, thereby forming a d-variate truncated normal distribution in the soundness case. This family of hard instances will be difficult to distinguish from $\mathcal{N}(0, \mathbf{I}_d)$, the standard multivariate normal distribution without truncation. We can establish a $\Omega(d)$ sample complexity bound using lower bound machinery developed in [DKS16, Proposition 7.1]. This indicates that any tester will require a sufficient number of samples to estimate the hidden direction v before being able to differentiate between the null and alternative hypothesis.

Lemma 4 (Sample Complexity Lower Bound for Mean Testing with Unknown Truncation When $\varepsilon \lesssim \alpha \lesssim \varepsilon\sqrt{\log(1/\varepsilon)}$)**.** *No algorithm can distinguish between* $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ *and a family of truncated normal distribution of the form:* $\mathcal{N}(\mathbf{v}, \mathbf{I}_d, S)$ *with measure* ε *on the truncation set* $\bar{S} = \mathbb{R}^d \backslash S$*, for any* $\varepsilon < 1$ *and some* $\|\mathbf{v}\|_2 = \alpha = \Theta(\varepsilon\sqrt{\log(1/\varepsilon)})$*, using fewer than* $\Omega\left(d/\varepsilon\right)$ *samples with a probability greater than* $2/3$*.*

The complete proof is provided in Appendix D.3. Below, we present a sketch of the proof for Lemma 4. We begin by constructing a one-dimensional truncated normal distribution $A = \mathcal{N}(\alpha, 1, S)$, where the truncated mass is ε. This means $\Pr_{x\sim\mathcal{N}(\alpha,1)}[x \in S] = 1 - \varepsilon$. We can determine the $1 - \varepsilon$ quantile as:

$$b = \alpha + \sqrt{2}\,\mathrm{erf}^{-1}(1 - 2\varepsilon).$$

which defines the truncation set as $S = (-\infty, b]$.

Let $\alpha(\varepsilon) = \alpha = \Theta\left(\varepsilon\sqrt{\log\frac{1}{\varepsilon}}\right)$. For any ε, we can find a constant $c_2 = \Theta(1)$ such that $\mathbb{E}[A] = 0$:

$$\mathbb{E}_{X\sim A}[X] = \alpha - \frac{\exp\left(-\frac{1}{2}\left(b - \alpha\right)^2\right)}{\sqrt{2\pi}(1 - \varepsilon)} = 0,$$

which is equivalent to:

$$\frac{\exp(-(\mathrm{erf}^{-1}(1 - 2\varepsilon))^2)}{\sqrt{2\pi}(1 - \varepsilon)} = \Theta\left(\varepsilon\sqrt{\log\frac{1}{\varepsilon}}\right) = \alpha.$$

Next, we compute an upper bound on the chi-squared divergence between the truncated distribution A and the standard normal distribution $\mathcal{N}(0, 1)$. We find that

$$\chi^2(A, \mathcal{N}(0, 1)) \leq \mathcal{O}(\varepsilon + \alpha^2),$$

We now apply Proposition 2.4.1 [DKS16, Proposition 7.1], and obtain a lower bound of

$$\Omega\left(\frac{d}{\varepsilon+\alpha^2}\right)=\Omega\left(\frac{d}{\varepsilon}\right).$$

Mean Testing Upper Bound We apply the standard learning-to-test approach: first we estimate the pre-truncation mean of the truncated normal using $\mathcal{O}(d/\alpha^2)$ samples, following [DK23, Proposition 1.20]. This gives an estimate $\hat{\mu}$ that is within α of the true mean before truncation. If $\hat{\mu}$ is sufficiently close to zero, we return "ACCEPT". Otherwise, return "REJECT".

6.3 Testing under known truncation

In this section, we demonstrate in Theorem 6.3.1 that when the truncation set is known, an alternative yet straightforward algorithm, which leverages the gradient of the maximum likelihood estimator, achieves the optimal sample complexity of $\mathcal{O}(\sqrt{d})$ across all parameter regimes. As a result, it is sufficient to apply Algorithm 7 with a sample complexity of $\mathcal{O}(\sqrt{d}/\alpha^2)$.

Algorithm 7: `GaussianMeanTester` with known truncation

Input : Sample access to truncated normal P on $\mathbb{R}^d$, threshold $\alpha>0$, and oracle access to its support set S

Output : "ACCEPT" if $P=\mathcal{N}(\mathbf{0},\mathbf{I}_d,S)$; "REJECT" if $P=\mathcal{N}(\mu,\mathbf{I}_d,S)$ and $\|\mu\|_2\geq\alpha$; both with probability at least 2/3

Compute $\boldsymbol{\mu}'_S=\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d,S)}[\mathbf{x}]$;
Set $n=\mathcal{O}(\sqrt{d}/\alpha^2)$;
Sample $2n$ i.i.d. points from P and denote them by $X_1,\ldots,X_n$ and $Y_1,\ldots,Y_n$;
$Z_1=\left(\frac{1}{n}\sum_{i=1}^n X_i-\boldsymbol{\mu}'_S\right)^\top\left(\frac{1}{n}\sum_{i=1}^n Y_i-\boldsymbol{\mu}'_S\right)$;
if $|Z_1|\leq\mathcal{O}(\alpha^2)$ **then**
    return "ACCEPT";
else
    return "REJECT";

The algorithm works as follows: Given the support S, it first calculates the truncated mean for the standard multivariate normal, denoted as μ'_S. Next, it

draws $2n$ i.i.d. samples from the truncated normal distribution P with unknown mean. The algorithm then computes the statistic:

$$Z_1 = \left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu'_S\right)^\top \left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu'_S\right).$$

The algorithm will return "ACCEPT" if $|Z_1| \leq \mathcal{O}(\alpha^2)$ and "REJECT" otherwise.

The proof of Theorem 6.3.1 relies on the following two lemmas.

Lemma 5. *Let Z_1 be the statistics in Algorithm 7 Line 4, and $\mu'_S = \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d,S)}[\mathbf{x}]$ (truncated mean under zero mean). Let μ_S be the truncated mean of the unknown Gaussian P, we can show that*

$$\mathbb{E}[Z_1] = \|\mu_S - \mu'_S\|_2^2.$$

$$\mathrm{Var}[Z_1] \leqslant O(\alpha^4 + \alpha^2 \cdot \|\mu_S - \mu'_S\|_2^2).$$

Lemma 6 (Gap of Mean under Truncation)**.** *Let $\mathbb{E}_{\mathbf{y}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d;S)}[\mathbf{y}] = \mu'_S$ and $\mathbb{E}_{\mathbf{y}\sim\mathcal{N}(\mu'',\mathbf{I}_d;S)}[\mathbf{x}] = \mu''_S$, where $\|\mu''\|_2^2 \geq \alpha^2$. Additionally, assume that $\mathcal{N}(\mu'', \mathbf{I}_d; S) \geq 1 - \beta$ for some constant β. Then, it holds that*

$$\|\mu'_S - \mu''_S\|_2^2 \geq \Omega(\alpha^2).$$

Proof sketch. Consider the negative log-likelihood function, $\bar{\ell}(\mathbf{0})$, with the mean set to $\mathbf{0}$ as the input parameter. This function is defined for a population drawn from a truncated normal distribution with an unknown mean μ. From Equation 2.4, we can express the gradient of the negative log-likelihood with respect to the mean evaluated at $\mathbf{0}$, as follows:

$$\nabla\bar{\ell}(\mathbf{0}) = -\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mu,\mathbf{I}_d,S)}[\mathbf{x}] + \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d,S)}[\mathbf{z}] = \mu_S - \mu'_S.$$

Likewise, when evaluating the gradient at μ, we have

$$\nabla\bar{\ell}(\mu) = -\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mu,\mathbf{I}_d,S)}[\mathbf{x}] + \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mu,\mathbf{I}_d,S)}[\mathbf{z}] = \mathbf{0}.$$

So, $\nabla\bar{\ell}(\mathbf{0})$ represents the difference between the truncated mean of the underlying distribution and that of the distribution with mean $\mathbf{0}$. From Lemma 1, we know that $\bar{\ell}(\cdot)$ is λ_0-strongly convex, and λ_0 is a constant if β is a constant. Therefore,

by leveraging the properties of strong convexity and applying the CauchySchwarz inequality, we obtain the following result:

$$\sqrt{\|\mu - \mathbf{0}\|_2^2 \cdot \|\nabla \bar{l}(\mu) - \nabla \bar{l}(\mathbf{0})\|_2^2}$$
$$\geqslant \langle \nabla \bar{l}(\mu) - \nabla \bar{l}(\mathbf{0}), \mu - \mathbf{0} \rangle \geqslant \frac{\lambda_0}{2} \|\mu\|_2^2$$

By simplifying the expression and substituting μ with any $\|\mu''\|_2^2 \geqslant \alpha^2$, we can show that:

$$\|\mu''_S - \mu'_S\|_2^2 \geqslant \Omega(\alpha^2).$$

□

Theorem 6.3.1 (Known truncation tester)**.** *There exists an algorithm (Algorithm 7) that takes i.i.d. samples from truncated normal Gaussian P and given oracle access to $S \subset \mathbb{R}^d$, the effective support of P, distinguishing the cases for parameters (mass of truncation) $0 < \varepsilon \leqslant 1 - \beta$, where β is a constant and (accuracy) $\frac{1}{4} \geqslant \alpha > 0$:*

- ***(Completeness)*** *P is a truncated Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_d, S)$ and $1 - \mathcal{N}(\mathbf{0}, \mathbf{I}_d; S) \leqslant \varepsilon$. In this case, the algorithm will output yes with probability at least $2/3$.*
- ***(Soundness)*** *P is a truncated Gaussian distribution $\mathcal{N}(\mu, \mathbf{I}_d, S)$ where $\|\mu\|_2 \geqslant \alpha$ and $1 - \mathcal{N}(\mathbf{0}, \mathbf{I}_d; S) \leqslant \varepsilon$. In this case, the algorithm will output no with probability at least $2/3$.*

The algorithm will take $\mathcal{O}\left(\frac{\sqrt{d}}{\alpha^2}\right)$ samples from P.

Proof sketch. Using Lemma 5 and Lemma 6, we apply Chebyshev inequality in the two cases:

1. **Completeness**: We know that $\mathbb{E}[Z_1] = 0$ and $\mathrm{Var}[Z_1] \leq O(\alpha^4)$. Thus, by Chebyshev's inequality, with probability at least $2/3$ using,
$$Z_1 \leq \mathcal{O}(\alpha^2).$$
2. **Soundness**: Let the non-truncated mean be μ'' (and $\|\mu''\|_2^2 \geq \alpha^2$) with $\|\mu''\|_2^2 \geq \alpha^2$. Here, $\mathbb{E}[Z_1] = \|\mu''_S - \mu'_S\|_2^2$ and $\mathrm{Var}[Z_1] \leq O(\alpha^4 + \alpha^2 \|\mu''_S - \mu'_S\|_2^2)$. Applying Chebyshev's inequality, with probability at least $2/3$, we have
$$Z_1 \geq \|\mu''_S - \mu'_S\|_2^2 - O(\alpha^2 + \alpha \|\mu''_S - \mu'_S\|) \geq \Omega(\alpha^2).$$

□

6.4 Conclusion and Future Work

In this chapter, we highlight the critical interplay between truncation mass ε and accuracy α in determining the sample complexity required for Gaussian-Mean-Testing in both known and unknown truncation regimes.

- **Unknown Truncation**: For $\varepsilon < \alpha/\sqrt{\log(1/\alpha)}$, we establish the tight sample complexity of $\Theta(\sqrt{d})$, indicating the effectiveness of testing under mild truncation. However, as ε approaches α, the sample complexity sharply jumps to $\Theta(d)$, indicating a much more challenging testing regime (where testing brings no sample complexity savings over learning). Furthermore, when $\alpha \leq \varepsilon$, the sample complexity becomes infinite, as testing becomes information-theoretically unfeasible.

- **Known Truncation**: In contrast, when the truncation is known, the sample complexity remains $\Theta(\sqrt{d})$ across all parameter ranges, even when $\varepsilon > \alpha$. Thus, having prior knowledge of truncation can facilitate efficient testing regardless of the relationship between α and ε.

Overall, this is the first work that provide valuable insights into the sample complexity for efficient Gaussian-Mean-Testing, emphasizing the importance of understanding truncation in designing algorithms for robust statistics.

In future work, we aim to generalize the soundness case by extending our analysis to any arbitrary (unknown) covariance matrix $\mathbf{\Sigma}$, beyond the identity-covariance case. Another avenue of research, inspired by the recent line of work on convex truncation [DNS23], is to explore whether structural assumptions on the truncation set (whether known or unknown), for instance convexity or rotational symmetry, could enable significantly more sample-efficient algorithms for the task.

Chapter 7

Learning High-dimensional Gaussians from Censored Data

7.1 Introduction

Missing data is a quite prevalent factor contributing to bias in statistical inference. It arises from various causes, such as limitations in instruments leading to unreliable data, incomplete data collection resulting in missing relevant information, societal biases influencing the suppression of observations, behavioral biases leading to subjects dropping out of studies or avoiding survey questions, ethical, legal, or privacy considerations restricting the utilization of collected data, and other similar factors. Unfortunately training models without consideration of missing data can lead to models that incorporate biases in the training data and make incorrect predictions, which may in turn reinforce those biases when the models are deployed.

Since the early days of statistics, missing data has been a well-known challenge in statistical inference, which occurs in a variety of domains, such as biology, physics, clinical trial design, genetics, economics, survey research, and the social sciences. It has motivated a vast effort towards developing methodologies that are more robust to missing data. As example, we refer the reader to some of the early works in statistics [Gal98; Pea02; PL08; Lee14; Fis31], some standard references in statistics and econometrics [Tob58; Ame73; HW77; Hec79; HM98; LR19], works targeting missing data in specific domains [War92; BK96; Tro+01; ABM08; HK10], books overviewing this literature [Mad86; Bre+96; BC14], and finally some recent work in computer science [MPT13; Das+18; Das+19; DRZ20; Das+21a; Das+21b; KTZ19;

FKT20; Ple21].

The effect that data misssingness has on statistical inference depends heavily on the missingness model. In general, missingness models in which the value of some datapoint influences whether or not it will be missing from the dataset are harder to deal with compared to models in which this happens randomly.

Techniques that have been extensively researched in scenarios where missingness either does not depend on the data or only depends on the observed data are referred to as missing completely at random (MCAR) and missing at random (MAR) respectively [Rub76; Tsi06; LR19]. In problems where missing entries depend on the underlying values which are themselves censored, known as missing not at random (MNAR), is substantially more difficult and less explored [RG97; RR97; SRR99; SMP15; Ada+20]. The MNAR model is quite often relevant in practical applications. For example, the depression registry for mental health status is more likely to have missing questionnaires leading to the self-censoring missingness [Car+21]. Data are missing by design due to the limitations of measurement resources, or the treatment discontinuation when participants go off-control due to the lack of tolerability [Lit+12].

The goal of this work is to advance our understanding of density estimation in the non-asymptotic sample regime when data is missing not at random. In particular, we consider the standard task of high-dimensional Gaussian distribution estimation, albeit in settings where every sample of the Gaussian may have a subset of its coordinates censored and which subset this is depends on the sample itself. We consider two models for how the censoring may depend on the sample:

- Self-censoring model (see Section 7.1.1.1): in this model, a sample $\mathbf{y}$ is drawn from an underlying Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, and each coordinate y_i of this sample is censored (i.e. replaced with a '?') depending on whether or not it satisfies a coordinate-specific Boolean predicate, i.e. whether $S_i(y_i) = 1$ or not.

- Linear thresholding model (see Section 7.1.1.2): in this more challenging model, whether or not each coordinate is seen depends on whether the **whole** sample satisfies a coordinate-specific predicate.

Our goal in both cases is to identify conditions on the predicates and the under-

lying distribution under which their parameters can be estimated computationally and sample-efficiently in each of the aforedescribed models. Our work advances prior research on Gaussian estimation in the presence of MNAR data in the non-asymptotic sample regime along the following axes:

- Gaussian estimation under censoring (see e.g. the classical works of [Gal98; Pea02; PL08; Fis31] and the ensuing literature): Prior work on this problem in the non-asymptotic sample regime studies the "all-or-nothing setting," where either all coordinates or no-coordinate can be observed [Das+18]. They also require that some absolute constant fraction of the Gaussian can be observed. In comparison to this work, we allow heavily corrupted data where no such constant fraction exists where no coordinate is missing. However, we allow the predicate determining the censoring of coordinate i to either be very general but only dependent on this coordinate (self-censoring model), or depend on all coordinates but be simpler, namely a hyperplane (linear-thresholding model).

- Gaussian estimation under self-selection (see e.g. the classical work of [Roy51] and the ensuing literature): Prior work on this problem in the non-asymptotic sample regime [Che+22] studies specific selection mechanisms (in particular hiding all but the maximum coordinate of each sample) and also assumes independence among the coordinates. In comparison to this work, we allow correlations among coordinates and more general masking mechanisms. However, we focus on Gaussian distributed coordinates while they can accommodate non-parametric distributions.

7.1.1 Our Contributions

In this chapter, we are interested in recovering the "uncorrupted" Gaussian distribution given samples from a "corrupted" distribution (according to our missingness model). We use a population maximum likelihood approach as the estimation algorithm, and apply projected stochastic gradient descent on the likelihood function. We give theoretical proof of fast convergence in the parameter space.

A *missingness model* is defined by a function $\mathbb{S} : \mathbb{R}^d \to 2^{[d]}$. For an underlying d-dimensional vector $\mathbf{y}$, $\mathbb{S}(\mathbf{y})$ is interpreted as the set of coordinates of $\mathbf{y}$ that are

not missing. An observation is a pair $(A, \mathbf{x})$, where $A = \mathbb{S}(\mathbf{y})$ and $\mathbf{x} = \mathbf{y}_A$ for an underlying sample $\mathbf{y} \in \mathbb{R}^d$.

7.1.1.1 Distribution Learning Under The Self-Censoring Mechanism

Self-censoring Missingness Model. The Self-censoring mechanism is commonly encountered in practice. In this model the missingness of an outcome is affected by its underlying value. For example, smokers are not willing to report their smoking behavior in insurance applications. Voters holding particular beliefs may not disclose their political preferences in election surveys. The Self-censoring model is of significant interest because the model is: (i) conceptually well-motivated, and (ii) can be considered as a baseline for other more complex missingness models. We say that $\mathbb{S}$ is a *self-censoring missingness model* if there exist sets $S_1, \ldots, S_d$ such that $\mathbb{S}(\mathbf{y}) = \{i \in [d] : y_i \in S_i\}$.

Our result in this setting rests on the following hypothesis:

Assumption 7.1.1. *For any pair of coordinates $i, j \in [d]$: $\Pr_{\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)}[y_i \in S_i, y_j \in S_j] \geq \alpha$.*

Theorem 7.1.2. *Suppose we can observe samples from $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ censored through a self-censoring missingness model $\mathbb{S}$. If assumption 7.1.1 is satisfied for some constant value of the parameter α, there exists a polynomial-time algorithm that recovers estimated $\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*$ with arbitrary accuracy. Specifically, for all $\varepsilon > 0$, and given that the eigenvalues of $\boldsymbol{\Sigma}^*$ lie in the interval $[\lambda_{min}, \lambda_{max}]$, the algorithm uses $\tilde{\mathcal{O}}\left(\frac{d^2(\lambda_{max}/\lambda_{min})^2}{\alpha\varepsilon^2}\right)$ samples and produces estimates that satisfy the following:*

$$\left\|\boldsymbol{\Sigma}^{*-1/2}(\boldsymbol{\mu}^* - \hat{\boldsymbol{\mu}})\right\|_2 \leq \mathcal{O}(\varepsilon);$$
$$\text{and } \left\|\boldsymbol{I} - \boldsymbol{\Sigma}^{*-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{*-1/2}\right\|_F \leq \mathcal{O}(\varepsilon).$$

Note that the sample complexity is proportional to $1/\alpha$. Furthermore, under the above conditions, we have $\mathrm{d_{TV}}(\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*), \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})) \leq \mathcal{O}(\varepsilon)$.

7.1.1.2 Mean Estimation Under Linear Thresholding Missingness

We say $\mathbb{S}$ is a *linear thresholding missingness model* if there exist $\mathbf{v}_1, \ldots, \mathbf{v}_d \in \mathbb{R}^d$ and $b_1, \ldots, b_d \in \mathbb{R}$ such that $\mathbb{S}(\mathbf{y}) = \{i \in [d] : \mathbf{v}_i^T \mathbf{y} \leq b_i\}$. For instance, if for a

pair of coordinates x_i, x_j, we can only observe the maximum of the two, this can be modeled by a linear thresholding model where the i'th coordinate is observed if $x_j - x_i \leq 0$ and the j'th coordinate is observed if $x_i - x_j \leq 0$.

Our main algorithmic result rests on the following two data hypotheses:

Assumption 7.1.3. *There exist some $\alpha, \beta > 0$ such that for any set $A \subseteq [d]$ of size at most βd,*

$$\Pr_{\mathbf{y} \sim N(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})}[A \subseteq \mathbb{S}(\mathbf{y})] \geq \alpha.$$

Note that we only consider the case where βd is a positive integer without loss of generality.

Assumption 7.1.4. *(Informal) There exists an* anchoring set *of coordinates C such that (i) C is observed in every sample, and (ii) conditioned on the values at C, each missingness pattern occurs with probability 0 or at least γ.*

The second assumption states that the anchoring subset is compulsorily observed and values at these coordinates determine the missingness pattern (the set of coordinates observed) almost fully. This is in analogy to the anchor topic modeling used in natural language processing, which is a variation of probabilistic topic modeling that incorporates a set of predefined "anchor words" to guide the topic modeling process. Our anchored missingness is similar to the "anchor words" assumption. For instance, the anchoring subset might be a set of questions in a questionnaire that are mandatory to answer and whose values are very indicative of the respondant's behavior. We now informally state our main algorithmic result here:

Theorem 7.1.5. *For a known covariance matrix $\boldsymbol{\Sigma}$, suppose we can observe samples from $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})$ censored through a linear thresholding missingness model $\mathbb{S}$. If assumption 7.1.3 and assumption 7.1.4 are satisfied, there exists a polynomial-time algorithm that recovers estimated $\boldsymbol{\mu}^*$ with arbitrary accuracy. Specifically, for all $\varepsilon > 0$, the algorithm uses* $\mathrm{poly}(d, 1/\alpha, 1/\beta, 1/\gamma, \lambda_{\max}(\boldsymbol{\Sigma})/\lambda_{\min}(\boldsymbol{\Sigma}), 1/\varepsilon, \log(1/\delta))$ *samples and running time, and with probability at least $1 - \delta$, produces an estimate $\hat{\boldsymbol{\mu}}$ such that $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*\|_\Sigma \leq \varepsilon$.*

7.1.2 Our Techniques

Self Censoring In the self censoring model, we show that the problem can be reduced to solving truncation problems in each of the 2-dimensional subspaces spanned by pairs e_i, e_j of the basis vectors. This allows us to use a 2-dimensional version of the algorithm in [Das+18] as subroutine for our algorithm in order to extract the information about pairwise correlation of coordinates needed to reconstruct the true covariance matrix $\mathbf{\Sigma}^*$. The mean is reconstructed in a more straightforward way via solving 1-dimensional truncation problems for each coordinate.

In particular, to estimate the diagonal entries of $\mathbf{\Sigma}^*$, we use 1-dimensional subproblems and for each off-diagonal entry Σ_{ij} we solve the 2×2 subproblem on the coordinates i and j, but we only care about the off-diagonal entries of the result (ignoring the rest) and we use exactly that value as the estimate for the entry. To achieve the required guarantees, assuming the condition number of the true covariance matrix $\mathbf{\Sigma}^*$ is constant, one needs to run the subproblems with $\varepsilon' = c \cdot \varepsilon/d$ for some constant $c < 1$ (since the Frobenious distance can be at most a factor d larger than the maximum entry-wise difference), which would imply a sample complexity of $O(1/\varepsilon'^2)$ for each subproblem. Therefore, we will need $O(1/\varepsilon'^2) = O(d^2/\varepsilon^2)$ samples in which both coordinates i and j are present for each of the $O(d^2)$ possible coordinate pairs. According to our assumption 7.1.3, this happens with probability at least α for a particular pair of coordinates. Therefore, if we draw $O(\frac{d^2 \log(1/\delta)}{\alpha\varepsilon^2})$ samples, we get that for each pair of coordinates, we have the required amount of samples with probability at least $1 - \delta$. Since we need $\delta < \frac{1}{d^2}$ to apply a union bound, we have that $O(\frac{d^2 \log d}{\alpha\varepsilon^2}) = \tilde{O}(\frac{d^2}{\alpha\varepsilon^2})$ samples are sufficient for $\mathbf{\Sigma}^*$ with constant condition number.

Linear-thresholding Model For the linear thresholding model, the above reduction does not work because the problem can no longer be "decomposed" into 2-dimensional ones. The reason is that whether or not some pair of coordinates (x_i, x_j) appears can now be affected by the value of $\boldsymbol{x}$ in coordinates different than i and j. Therefore, we design a projected stochastic gradient descent (PSGD) algorithm that, given the covariance matrix $\mathbf{\Sigma}$ is known, yet arbitrary, it will give us an estimate of the true mean $\boldsymbol{\mu}^*$ of the original distribution, which can be arbitrarily

close to it with the right choice of parameters. The algorithm first uses an empirical estimator for the initialization of the estimate. We show that its distance to the true mean is bounded as a function of $\boldsymbol{\Sigma}$ and the parameters of assumption 7.1.3. Subsequently, we run a PSGD algorithm, whose projection step maintains this property. The gradient sampling step is non-trivial as a straightforward rejection sampling approach would run in exponential time. Therefore, we resort to a *Langevin Monte Carlo algorithm* which yields an approximately unbiased sample of the gradient. The projection set in this algorithm ensures that the centralized second moment of the gradient estimator is bounded, while its bias is also kept small. Combining this with our lower bound on the convexity parameter of the strongly convex likelihood function $\ell(\boldsymbol{\mu})$, we are able to show the result.

7.1.3 Related Work

High-dimensional distribution learning [Kea+94] initiated a systematic investigation of the computational complexity of distribution learning. Since then, there has been a large volume of works devoted to the parameter and distribution learning from a wide range of distributions in both low and high dimensions [Das99a; SK01; Cha+13; GHK15; Dia+19; Bak+22]. Broadly, this problem falls into the realm of robust statistics. Following the pioneering works by [Tuk60; Hub92], other recent works on high dimensional robust distribution learning can be found at [CSV17; Rek+17; Dia+18; Kho+19; Dia+20; Kan21]. We will be particularly interested in robustly estimating mean and covariance from high-dimensional data with partially-reliable data samples [Bar64; Sza80; Ste81; BM09; BS10; Pas+13; LRV16; DK19; Dia+19; Lei+20; Che+20; CMY20; HLZ20]. Settings similar to ours are studied in [Liu+21; HR21] regarding robust mean estimation with coordinate-level corruptions. In this paper, we obtain stronger guarantees for the mean estimation, yet incomparable to [Liu+21] due to their stronger corruption model.

Learning from truncated or censored samples Distribution learning under censored, truncated mechanisms has had a long history. Censoring happens when the events can be detected, but the measurements (the values) are completely unknown, while truncation occurs when an object falling outside some subset are not

observed, and their count in proportion to the observed samples is also not known, see [DV55; Coh57; Dix60; HS90; Coh91b; BS99; CCS13; CSV17] for an overview of the related works in estimating the censored or truncated normal or other type of distributions. [Das+18; Das+19; DRZ20] developed computationally and statistically efficient algorithms under the assumption that the truncation set is known. Furthermore, [WDS19] considered the problem of estimating the parameters of a d-dimensional rectified Gaussian distribution from i.i.d. samples. This can be seen as a special case of the self-censoring truncation, where the truncation happens due to the ReLU generative model. [SMP15] explored the identification and estimations conditions when data are missing not-at-random. While [Bha+20; NBS20; MST21] explored the necessary and sufficient graphical conditions to recover the full data distribution under no self-censoring condition.

Learning from general missingness More broadly, self-selection models fall under the literature of regression with MNAR in the outcomes [RR95; RRS98; TWS18]. Unlike self-censoring, this project doesn't restrict the form of the representation. Two most popular methods are the expectation-maximization algorithm [DLR77] and Gibbs sampling [GG84] under MAR. Despite the long history and the application of missing data models, most of the existing methods with regard to robust learning [RS01] are consistent in the asymptotic sample regime. For example, likelihood method [EB01], multiple imputation [All00], semiparametric estimation with influence function [RRS00], inverse probability weighted complete-case estimator [Woo07; SW13], and double/debiased machine learning [Che+18]. See textbook [Tch06; Tsi06; Van18] for more introductions and further applications in this field. Recently, there are several finite sample guarantees for the double robust estimator when data are MNAR [CNS18; CNS21] and high-dimensional [Qui22]. In addition to the works discussed, there has been significant research on detecting truncation [DNS23; De+24] and estimation under unknown truncation [KTZ19; DKS17].

7.2 Notations and Preliminaries

Throughout, let $d \geq 1$ denote the dimension of the underlying domain. For a d-dimensional vector $\mathbf{u}$ and a subset $A \subseteq [d]$, let $\mathbf{u}_A \in \mathbb{R}^{|A|}$ denote the restriction of

$\mathbf{u}$ to the coordinates in A. A *missingness model* is defined by a function $\mathbb{S} : \mathbb{R}^d \to 2^{[d]}$. Given a distribution $\mathcal{D}$ on $\mathbb{R}^d$, an *observation of $\mathcal{D}$ censored by $\mathbb{S}$* is a pair $(A, \mathbf{x}) \in 2^{[d]} \times \mathbb{R}^{|A|}$, generated by first sampling $\mathbf{y} \sim \mathcal{D}$ and then setting $A = \mathbb{S}(\mathbf{y})$ and $\mathbf{x} = \mathbf{y}_A$. The interpretation is that y_i is seen for every $i \in \mathbb{S}(\mathbf{y})$ while y_i is missing for every $i \notin \mathbb{S}(\mathbf{y})$. We denote the resulting distribution on pairs by $\mathcal{D}^{\mathbb{S}}$. If the density function of $\mathcal{D}$ is f, then the density function of $\mathcal{D}^{\mathbb{S}}$ is $f^{\mathbb{S}}$ defined as:

$$f^{\mathbb{S}}(A, \mathbf{x}) = \int_{\mathbf{y} \in \mathbb{R}^d} \mathbb{1}_{\{]}\mathbb{S}(\mathbf{y}) = A] \cdot \delta(\mathbf{y}_A - \mathbf{x}) f(\mathbf{y}) d\mathbf{y}. \tag{7.1}$$

Note that $\sum_{A \subseteq [d]} \int_{\mathbf{x} \in \mathbb{R}^{|A|}} f^{\mathbb{S}}(A, \mathbf{x}) = 1$, as desired. We say that $\mathbb{S}$ is a *self-censoring missingness model* if there exist sets $S_1, \ldots, S_d$ such that $\mathbb{S}(\mathbf{y}) = \{i \in [d] : y_i \in S_i\}$. We say $\mathbb{S}$ is a *linear threshold missingness model* if there exist $\mathbf{v}_1, \ldots, \mathbf{v}_d \in \mathbb{R}^d$ and $b_1, \ldots, b_d \in \mathbb{R}$ such that $\mathbb{S}(\mathbf{y}) = \{i \in [d] : \mathbf{v}_i^T \mathbf{y} \leq b_i\}$.

Fact 7.2.1. *Let $A = \{a_{ij}\}$ and $B = \{b_{ij}\}$ be two $d \times d$ matrices such that $\forall i, j$: $|a_{ij} - b_{ij}| \leq \delta$. Then, $\|A - B\|_F \leq \delta \cdot d$.*

7.3 Distribution Learning under Self-Censoring Missingness

The problem of learning a distribution from truncated samples was studied in [Das+18]. Their guarantee, as presented in Theorem 7.3.1, is given under the assumption that a fraction α of all the samples is fully observed across all dimensions.

Theorem 7.3.1 (adapted from [Das+18])**.** *Given oracle access to a measurable set T, whose measure under some unknown d-variate normal $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ is at least some constant $\alpha > 0$, and samples $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}$ from $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ that are truncated to this set, there exists a polynomial-time algorithm that recovers estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$. In particular, for all $\varepsilon > 0$, the algorithm uses $\tilde{O}(d^2/\varepsilon^2)$ truncated samples and queries to the oracle and produces estimates that satisfy the following with probability at least 99%.*

$$\|(\boldsymbol{\mu}^* - \hat{\boldsymbol{\mu}})\|_2 \leq \varepsilon \sqrt{\lambda_{max}}; \quad \text{and} \quad \|\boldsymbol{\Sigma}^* - \hat{\boldsymbol{\Sigma}}\|_F \leq \varepsilon \lambda_{max}.$$

This simplifies the problem because with enough samples, part of the shape of the Gaussian distribution can be observed, allowing for simultaneous estimation of the mean and covariance. In contrast, the self-censoring missingness only allows us to observe a subset of samples, making the estimation problem more challenging. The goal is to recover $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$ under minimal assumptions on the censoring mechanism. In this section, we present and analyze our algorithm for estimating the true mean and covariance of the multivariate normal distribution under self-censoring missingness.

The main idea behind our algorithm for self-censoring missingness is to use the solutions to 1-dimensional and 2-dimensional subproblems as subroutines and subsequently combine them appropriately to obtain the solution. These subproblems are either the restriction of our problem to a single coordinate or a pair of coordinates. assumption 7.1.1 guarantees the existence of sufficiently many samples for these problems and allows us to use the $1D$ and $2D$ versions of Algorithm 1 in [Das+18] as our `Univariate_SGD_truncation` and `Bivariate_SGD_truncation` estimator respectively.

Even though these subroutines can give us accurate estimates for each coordinate of the true mean and the correlations between pairs of sample coordinates it is unfortunately not straightforward to provide an estimate of the $d \times d$ covariance matrix satisfying our desired guarantees. We explain below how to get around this issue.

We reconstruct the covariance matrix by only considering pairs of coordinates. For each $i \neq j$, we apply the 2-dimensional version of the algorithm in [Das+18] (`Bivariate_SGD_truncation`) on the ith and j'th coordinates to obtain the 2×2-matrix $\hat{\boldsymbol{\Sigma}}^{ij}$. We will show that the $d \times d$ matrix $\hat{\Sigma}$ whose off diagonal entries $(\hat{\Sigma}_{ij})$ are given by the off diagonal entries $(\hat{\Sigma}^{ij}_{12})$ of the corresponding 2×2 matrix is a good approximation for the true Σ.

We are now ready to describe Algorithm 8 demonstrating our distribution learning approach under self-censoring missingness mechanism.

Mean Estimation We show in Lemma 7 the finite sample bound with a consistent mean estimation up to a bounded error of $\mathcal{O}(\varepsilon)$. The proof is deferred to the appendix.

Algorithm 8: [Truncation_PSGD] Mean and covariance recovery algorithm with oracle access that generates samples with incomplete data.

Input : Data $\mathbf{x} \in \mathbb{R}^{n \times d}$, where $n = \frac{1}{\alpha \varepsilon'^2}$

1 **for** $i \leftarrow 1$ **to** d **do**
2 $\hat{\mu}_i, \hat{\Sigma}_{ii} \leftarrow$ Uni_SGD_trunc$(\mathbf{x}_i, S_i)$;
3 **for** $i \leftarrow 1$ **to** $d-1$ **do**
4 **for** $j \leftarrow i+1$ **to** d **do**
5 $\hat{\mathbf{\Sigma}}_{12}^{ij} \leftarrow$ Biv_SGD_trunc$(\mathbf{x}_i, \mathbf{x}_j, S_i \times S_j)$;
6 $\hat{\boldsymbol{\mu}} \leftarrow [\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_d]$;
7 **for** $i \leftarrow 1$ **to** $d-1$ **do**
8 **for** $j \leftarrow i+1$ **to** d **do**
9 $\hat{\mathbf{\Sigma}}_{ij} \leftarrow \hat{\mathbf{\Sigma}}_{12}^{ij}; \qquad \hat{\mathbf{\Sigma}}_{ji} \leftarrow \hat{\mathbf{\Sigma}}_{12}^{ij};$
10 **return** $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}})$

Lemma 7. *Let $\mathcal{N}(\boldsymbol{\mu}^*, \mathbf{\Sigma}^*)$ be the normal distribution with mean $\boldsymbol{\mu}^*$ and covariance matrix $\mathbf{\Sigma}^*$. Suppose that assumption 7.1.1 holds for some constant $\alpha > 0$, and let $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_d)$ be the estimated mean from the censored Gaussian in Line 6 of Algorithm 1. For all $\varepsilon > 0$, using $\tilde{\mathcal{O}}(\frac{d^2}{\alpha \varepsilon^2})$ samples[1] we have that:*

$$\forall i \in [d] : |\mu_i^* - \hat{\mu}_i| \leq (\varepsilon/d)\sigma_i \leq (\varepsilon/d)\sqrt{\lambda_{max}(\mathbf{\Sigma})}$$

where σ_i denotes the standard deviation of coordinate i (i.e $\sigma_i = \sqrt{\Sigma_{ii}^*}$, where Σ_{ii}^* is the i-th diagonal entry of the covariance matrix $\mathbf{\Sigma}^*$).

Covariance Estimation In Lemma 8 below, we show that if for each pair of coordinates we are given enough samples in which this particular pair is seen, we are able to obtain an accurate estimation of $\mathbf{\Sigma}^*$. In particular, we will run the 2D version of the problem for each of the $\binom{d}{2}$ pairs of coordinates and require that the estimate has accuracy ε_2. By applying Theorem 7.3.1 for $d = 2$, and error $\delta = \frac{1}{100\binom{d}{2}}$, we conclude that $\tilde{\mathcal{O}}(1/\varepsilon_2^2)$ samples are sufficient to achieve 99% success probability via a union bound. We will show that ε_2 doesn't need to be too small.

Lemma 8. *Let $\hat{\mathbf{\Sigma}}$ be the matrix with entries $\hat{\Sigma}_{ij} = \hat{\Sigma}_{12}^{ij}$, where $\hat{\Sigma}_{12}^{ij}$ denotes the value of the off diagonal entries of the 2×2 matrix $\hat{\mathbf{\Sigma}}^{\boldsymbol{ij}}$. By $\hat{\mathbf{\Sigma}}^{ij}$ we denote the estimation of*

[1] We note that the $\tilde{\mathcal{O}}_\alpha$ notation here hides both $\log d$ and $\log(1/\delta)$ factors.

a 2×2 *covariance matrix that we get when we restrict the input data to coordinates* i *and* j*. Then the following holds: Using* $\tilde{\mathcal{O}}(\frac{d^2}{\alpha\varepsilon^2})$ *samples to get the above estimates,we have that:*

$$\|\boldsymbol{\Sigma}^* - \hat{\boldsymbol{\Sigma}}\|_F \leq \varepsilon\lambda_{max}$$

where λ_{max} *is the maximum eigenvalue of* Σ^*

Based on the above results, we summarize our main results under the self-censoring missingness mechanism in Theorem 7.1.2.

Theorem 7.1.2. *Suppose we can observe samples from* $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ *censored through a self-censoring missingness model* $\mathbb{S}$*. If assumption 7.1.1 is satisfied for some constant value of the parameter* α*, there exists a polynomial-time algorithm that recovers estimated* $\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*$ *with arbitrary accuracy. Specifically, for all* $\varepsilon > 0$*, and given that the eigenvalues of* $\boldsymbol{\Sigma}^*$ *lie in the interval* $[\lambda_{min}, \lambda_{max}]$*, the algorithm uses* $\tilde{\mathcal{O}}\left(\frac{d^2(\lambda_{max}/\lambda_{min})^2}{\alpha\varepsilon^2}\right)$ *samples and produces estimates that satisfy the following:*

$$\left\|\boldsymbol{\Sigma}^{*-1/2}(\boldsymbol{\mu}^* - \hat{\boldsymbol{\mu}})\right\|_2 \leq \mathcal{O}(\varepsilon);$$
$$\text{and } \left\|\boldsymbol{I} - \boldsymbol{\Sigma}^{*-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{*-1/2}\right\|_F \leq \mathcal{O}(\varepsilon).$$

Note that the sample complexity is proportional to $1/\alpha$*. Furthermore, under the above conditions, we have* $\mathrm{d_{TV}}(\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*), \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})) \leq \mathcal{O}(\varepsilon)$.

With the following lemma we will show a lower bound, which shows that even if the TV distance is large, in which case the distributions are easily distinguishable in the classical sampling model, the censoring model makes the distribution hard to distinguish.

Lemma 9. *Given* $m = o(1/\sqrt{\lambda_{min}})$ *censored samples according to the missingness model* $\mathbb{S}$ *and* $\varepsilon = \Omega(\sqrt{\lambda_{min}})$*. No algorithm can estimate the true mean with accuracy* $O(\varepsilon)$ *and probability larger than* $2/3$*.*

Note that, for $\varepsilon = \Omega(\sqrt{\lambda_{min}})$ the TV distance between the distributions P_λ and Q_λ is $\Omega(1)$, yet the $\Omega(1/\sqrt{\lambda_{min}})$ censored samples are necessary.

7.4 Mean Estimation under Linear Thresholding Missingness

In this section, we present sufficient conditions for mean estimation under linear thresholding missingness. As earlier, we let $\mathbb{S}$ denote the missingness model, and $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})$ denote the ground truth distribution. Our observations are drawn from $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})^{\mathbb{S}}$.

We will make the following two assumptions on the missingness mechanism and the ground truth distribution. Our first assumption ensures that any small subset of coordinates is observed simultaneously with non-negligible probability.

Assumption 7.1.3. *There exist some $\alpha, \beta > 0$ such that for any set $A \subseteq [d]$ of size at most βd,*

$$\Pr_{\mathbf{y} \sim N(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})}[A \subseteq \mathbb{S}(\mathbf{y})] \geq \alpha.$$

Note that we only consider the case where βd is a positive integer without loss of generality.

This is a stronger version of assumption 7.1.1. Our second assumption postulates existence of an "anchoring" subset.

Definition 7.4.1 (Anchored missingness)**.** *A subset $C \subseteq [d]$ is* γ-anchoring *if*

(i) $C \subseteq \mathbb{S}(\mathbf{y})$ for any $\mathbf{y}$, and

(ii) for any $A \subseteq [d]$, $\Pr_{\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})}[\mathbb{S}(\mathbf{y}) = A \mid \mathbf{y}_C]$ is either 0 or at least γ.

Assumption 1.4. *There exists a γ-anchoring subset C for the true distribution $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})$ in combination with the missingness model $\mathbb{S}$.*

Given the assumptions above, we will prove the following result showing that we can accurately and efficiently recover the mean of the distribution using censored samples:

Theorem 7.1.5. *For a known covariance matrix $\boldsymbol{\Sigma}$, suppose we can observe samples from $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})$ censored through a linear thresholding missingness model $\mathbb{S}$. If assumption 7.1.3 and assumption 7.1.4 are satisfied, there exists a polynomial-time algorithm that recovers estimated $\boldsymbol{\mu}^*$ with arbitrary accuracy. Specifically, for*

all $\varepsilon > 0$*, the algorithm uses* $\mathrm{poly}(d, 1/\alpha, 1/\beta, 1/\gamma, \lambda_{\max}(\mathbf{\Sigma})/\lambda_{\min}(\mathbf{\Sigma}),\ 1/\varepsilon, \log(1/\delta))$ *samples and running time, and with probability at least* $1 - \delta$*, produces an estimate* $\hat{\boldsymbol{\mu}}$ *such that* $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*\|_\Sigma \leq \varepsilon$.

General outline In this section, we present and analyze our mean estimation algorithm of `MissingDescent` under anchor missingness models. As a high-level overview, our approach involves running a Projected Stochastic Gradient Descent (PSGD) algorithm on a negative log-likelihood function whose optimal value coincides with the true mean. The steps of proof are as follows:

- We develop an appropriate objective function in Section 7.4.1 and design an efficient mean estimation algorithm `MissingDescent` in Algorithm 10, assuming that any small subset of coordinates is observed with sufficiently high probability (assumption 7.1.3), and the observed missingness pattern is not very rare conditioned on the values of the observed coordinates (assumption 7.1.4).
- We show that our objective function is strongly convex with respect to the correct parameterization and hence the optimum is unique. Furthermore, it is equal to the true mean.
- We analyze our `MissingDescent` algorithm in Section 7.4.3 by showing that Algorithm 10 approximately optimizes ℓ with bounds on the runtime and sample complexity.
- Specifically, we show in Algorithm 9 in Section 7.4.2 that we can use the `Initialize` algorithm to efficiently compute an initial feasible point to start the optimization.
- In the `SampleGradient` algorithm in Algorithm 12, we demonstrate that it is possible to obtain an estimate of $\Delta\ell(\boldsymbol{\mu})$ that is approximately unbiased by sampling from the conditional distribution. Additionally, we use the `ProjectToDomain` algorithm in Algorithm 11 to project a current guess back onto the domain.

7.4.1 Negative Log-likelihood Objective Function with Anchor Missingness

We will approach the mean estimation problem via optimization of the population log-likelihood with respect to a given parameter estimate μ for the true mean μ^*. Define $g_{\boldsymbol{\mu}}$ to be the density function of $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$g_{\boldsymbol{\mu}}(\mathbf{y}) = (2\pi)^{\frac{-d}{2}} |\boldsymbol{\Sigma}|^{\frac{-1}{2}} \exp\left(-\frac{(\mathbf{y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{2}\right).$$

Recall the notation $g_{\boldsymbol{\mu}}^{\mathbb{S}}$ defined in Section 7.2 to be the density function of $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ censored by $\mathbb{S}$. We can then write down the population negative log-likelihood ℓ as:

$$\begin{aligned}\ell(\boldsymbol{\mu}) &= \mathop{\mathbb{E}}_{(A,\mathbf{x})\sim\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma})^{\mathbb{S}}}[-\log g_{\boldsymbol{\mu}}^{\mathbb{S}}(A,\mathbf{x})] \\ &= \mathop{\mathbb{E}}_{(A,\mathbf{x})}\left[-\log \int_{\mathbf{y}} \mathbb{1}_{\{\}}\mathbb{S}(\mathbf{y}) = A] \cdot \delta(\mathbf{y}_A - \mathbf{x}) \cdot g_{\boldsymbol{\mu}}(\mathbf{y}) d\mathbf{y}\right].\end{aligned}$$

In the second equality, and everywhere later, $(A, \mathbf{x})$ is an observation sampled from the censored version of the true distribution: $\mathcal{N}(\boldsymbol{\mu}^*, \Sigma)^{\mathbb{S}}$. The integral above marginalizes over all $\mathbf{y}$ for which the missingness model would yield the observation $(A, \mathbf{x})$.

The gradient with respect to $\boldsymbol{\mu}$ of $\boldsymbol{\nabla}\ell(\boldsymbol{\mu})$ can be expressed as

$$\mathop{\mathbb{E}}_{(A,\mathbf{x})}\left[-\frac{\int_{\mathbf{y}} \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}) \cdot \mathbb{1}_{\{\}}\mathbb{S}(\mathbf{y}) = A] \cdot \delta(\mathbf{y}_A - \mathbf{x}) g_{\boldsymbol{\mu}}(\mathbf{y}) d\mathbf{y}}{\int_{\mathbf{y}} \mathbb{1}_{\{\}}\mathbb{S}(\mathbf{y}) = A] \cdot \delta(\mathbf{y}_A - \mathbf{x}) \cdot g_{\boldsymbol{\mu}}(\mathbf{y}) d\mathbf{y}}\right] \tag{7.2}$$

$$= -\mathop{\mathbb{E}}_{(A,\mathbf{x})}[\mathop{\mathbb{E}}_{\mathbf{y}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)}[\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}) \mid \mathbb{S}(\mathbf{y}) = A, \mathbf{y}_A = \mathbf{x}]] \tag{7.3}$$

Lemma 10. *For any $\boldsymbol{\mu} \in \mathbb{R}^d$, it holds that: $\ell(\boldsymbol{\mu}) \geq \ell(\boldsymbol{\mu}^*)$.*

Lemma 11 (Strong Convexity with Missing Entries)**.** *Given our missingness model and assumption 7.1.3 with $\beta = \frac{c}{d}$ for some integer $c \in \{1, \dots d\}$, we have that the function with general covariance $\ell(\boldsymbol{\mu})$ is λ-strongly convex for $\lambda = \alpha\beta/\lambda_{\max}(\boldsymbol{\Sigma})$.*

Remark. Convexity may not hold if the missignnness pattern is not linear thresholding. For example, even for $d = 1$, if $\mu^* = 0$ and $\mathbb{S}(y) = \{1\}$ if $y \in [2, 4]$ and $\emptyset$ otherwise, the function $\ell(\mu)$ is not convex.

7.4.2 Algorithm

Initialization Our first step for efficiently optimizing the negative log-likelihood function is finding a good initial point for the PSGT. Specifically, we take the empirical mean $\hat{\mu}$. This is a biased estimate, but we show below that this is good enough for initialization: the distance of the empirical estimates and true mean μ^* is a constant that depends only on the constant β, mass α and λ_{max} of the known $\mathbf{\Sigma}$. The pseudocode for `Initialize` appears in Algorithm 9.

Algorithm 9: [`Initialize`] Initialization for the main algorithm.

Input : Access to data generator $\mathcal{O}$, parameter $\beta = \frac{c}{d}$ for some integer $c \in \{1, \ldots, d\}$, number of samples M_{init}

1 $\mathbf{w} \leftarrow$ empty array of length d
2 $\mathbf{X} \leftarrow$ matrix with M_{init} rows, each an independent sample from $\mathcal{O}$
3 **for** $i \leftarrow 0$ **to** $\lceil 1/\beta \rceil - 1$ **do**
4 $s \leftarrow i\beta d + 1$
5 $t \leftarrow \min\{(i+1)\beta d, d\}$
6 $\mathbf{Y}_i \leftarrow$ submatrix of $\mathbf{X}$ consisting of columns $s, s+1, \ldots, t$
7 Remove all rows of $\mathbf{Y}_i$ containing at least one $*$
8 $\hat{\boldsymbol{\mu}}_i \leftarrow$ average of the rows of $\mathbf{Y}_i$
9 $\mathbf{w}[s, s+1, \ldots, t] \leftarrow \hat{\boldsymbol{\mu}}_i$
10 **return** $\mathbf{w}$

By assumption 7.1.3, we have that after line 7 in `Initialize`, each $\mathbf{Y}_i$ is the truncation of a βd-dimensional gaussian where the truncation set has mass at least α. Using Lemma 6 of [Das+18], the mean of such a truncated gaussian is $O(\sqrt{\log(1/\alpha)})$ distance away from the untruncated mean. Hence, we have $\|\mathbb{E}[\mathbf{w}] - \boldsymbol{\mu}^*\|_2^2 = \sum_i \|\mathbb{E}[\hat{\boldsymbol{\mu}}_i] - \boldsymbol{\mu}[i\beta d + 1, \ldots, (i+1)\beta d]\|_2^2 \leq \lambda_{\max} \sum_i \|\mathbb{E}[\hat{\boldsymbol{\mu}}_i] - \boldsymbol{\mu}[i\beta d + 1, \ldots, (i+1)\beta d]\|_{\mathbf{\Sigma}}^2 \leq \mathcal{O}(\frac{\lambda_{\max}}{\beta}\log(1/\alpha))$.[2] Therefore, $\|\mathbb{E}[\mathbf{w}] - \boldsymbol{\mu}^*\|_2 \leq \mathcal{O}(\sqrt{\frac{\lambda_{\max}}{\beta}\log(1/\alpha)})$. Later in Section 7.4.3, we analyze the number of samples M_{init} required for $\|\mathbf{w} - \boldsymbol{\mu}^*\|_2$ to satisfy this bound with high probability.

Note that, in each iteration of SGD in `MissingDescent` (Algorithm 10), we choose a projection set, to make sure that PSGD converges. Specifically, we project

[2] Define $\boldsymbol{x} = \mathbb{E}[\hat{\boldsymbol{\mu}}_i] - \boldsymbol{\mu}[i\beta d + 1, \ldots, (i+1)\beta d]$, and the eigenvalue decomposition of $\mathbf{\Sigma}^{-1}$ as $\mathbf{\Sigma}^{-1} = Q^\top D^{-1} Q$. The first inequality holds because $\|\boldsymbol{x}\|_{\mathbf{\Sigma}}^2 = \|\boldsymbol{x}^\top Q^\top D^{-1} Q\boldsymbol{x}\|_2 = \|D^{-1/2} Q\boldsymbol{x}\|_2^2 \geq \frac{1}{\lambda_{\max}}\|Q\boldsymbol{x}\|_2^2 = \frac{1}{\lambda_{\max}}\|\boldsymbol{x}\|_2^2$. Therefore, we have $\|\boldsymbol{x}\|_2^2 \leq \lambda_{\max}\|\boldsymbol{x}\|_{\mathbf{\Sigma}}^2$

Algorithm 10: [`MissingDescent`] Mean recovery algorithm given access to an oracle that generates samples with incomplete data.

Input : Access to data generator $\mathcal{O}$, parameters $\beta, \lambda_{\text{sgd}}, \eta_{\text{lmc}}, R_{\text{lmc}}, r_{\text{proj}}$, M_{init}, M_{sgd}, M_{grad}

$\boldsymbol{\mu}^{(0)} \leftarrow$ `Initialize`$(\mathcal{O}, \beta, M_{\text{init}})$

for $i \leftarrow 1$ **to** M_{sgd} **do**

  Sample $(A^{(i)}, \mathbf{x}^{(i)})$ from $\mathcal{O}$

  $\eta_i \leftarrow \frac{1}{\lambda_{\text{sgd}} \cdot i}$

  $\mathbf{g}^{(i)} \leftarrow$ `SampleGradient`$((A^{(i)}, \mathbf{x}^{(i)}), \boldsymbol{\mu}^{(i-1)}, \eta_{\text{lmc}}, R_{\text{lmc}}, M_{\text{grad}})$

  $\mathbf{v}^{(i)} \leftarrow \boldsymbol{\mu}^{(i-1)} - \eta_i \mathbf{g}^{(i)}$

  $\boldsymbol{\mu}^{(i)} \leftarrow$ `ProjectToDomain`$(\boldsymbol{\mu}^{(0)}, \mathbf{v}^{(i)}, r_{\text{proj}})$

$\bar{\boldsymbol{\mu}} \leftarrow \frac{1}{M_{\text{sgd}}} \sum_{i=1}^{M_{\text{sgd}}} \boldsymbol{\mu}^{(i)}$

return $\bar{\boldsymbol{\mu}}$

a current guess back to a $\mathcal{B}_{\boldsymbol{\Sigma}}$ ball scaled by r_{proj} and centered at $\boldsymbol{\mu}^{(0)}$ as shown below:

Algorithm 11: [`ProjectToDomain`] The function that projects a current guess back to the domain onto the $\mathcal{B}_{\boldsymbol{\Sigma}}$ ball.

Input : $\boldsymbol{\mu}^{(0)}, \mathbf{v}$, parameter r_{proj}

return $\boldsymbol{\mu}^{(0)} + \min\{r_{proj}, \|(\mathbf{v} - \boldsymbol{\mu}^{(0)})\|_{\Sigma}\} \cdot \frac{(\mathbf{v} - \boldsymbol{\mu}^{(0)})}{(\|\mathbf{v} - \boldsymbol{\mu}^{(0)})\|_{\Sigma}}$

Our goal is to minimize the population negative log-likelihood ℓ via (projected) stochastic gradient descent while maintaining its strong-convexity. Specifically, Algorithm 10 above describes this strategy. In order to apply Algorithm 10 to our log-likelihood objective function, we need to solve the following three algorithmic problems:

- **Initialization**: efficiently compute an initial feasible point from which to start the optimization. The pseudocode for `Initialize` appears in Algorithm 9;
- **Gradient estimation**: design a nearly unbiased sampler for $\nabla \ell(\boldsymbol{\mu})$ using Langevin sampling. The `SampleGradient` pseudocode appears in Algorithm 12;
- **Efficient projection**: perform an efficient projection into a set of feasible points to make sure that PSGD converges. The pseudocode presents in Algorithm 11.

Algorithm 12: [`SampleGradient`] Sampler for $\nabla \ell(\boldsymbol{\mu})$.

Input : $(A, \mathbf{x})$, $\boldsymbol{\mu}$, parameters η, R, M

$a \leftarrow |A|$

Compute $\boldsymbol{\mu}_{\text{cond}}$ and $\boldsymbol{\Sigma}_{\text{cond}}$ as in (E.10) and (E.11)

Let W be such that $\boldsymbol{\Sigma}_{\text{cond}} = WW^\top$

Compute $\mathcal{L} = (W^{-1}\mathcal{K}) \cap \mathcal{B}_\Sigma(W^{-1}\boldsymbol{\mu}_{\text{cond}}, R)$

$\mathbf{z}^{(0)} \leftarrow$ a point in $\mathcal{L}$

for $t = 0$ **to** $M - 1$ **do**

  Sample $\boldsymbol{\zeta}^{(t)}$ from $\mathcal{N}(0, I_{d-a})$

  $\mathbf{z}^{(t+1)} \leftarrow \Pi_{\mathcal{L}}\left(\mathbf{z}^{(t)} - \eta(\mathbf{z}^{(t)} - W^{-1}\boldsymbol{\mu}_{\text{cond}}) + \sqrt{\eta} \cdot \boldsymbol{\zeta}^{(t)}\right)$

return $-\boldsymbol{\Sigma}^{-1}(\mathbf{x} \circ (W\mathbf{z}^{(M)}) - \boldsymbol{\mu})$

7.4.3 Analysis of `MissingDescent`

We show in this section that Algorithm 10 approximately optimizes ℓ with bounds on the runtime and sample complexity. The following lemma describes the ingredients necessary to obtain such bounds:

Lemma 12 (Lemma 6 in [Che+22])**.** *Let $f : \mathbb{R}^k \to \mathbb{R}$ be a convex function, $K \subseteq \mathbb{R}^k$ a convex set, and fix an initial estimate $\mathbf{x}^{(0)} \in K$. Now, let $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}$ be the iterates generated by running T steps of projected SGD using gradient estimates $\mathbf{g}^{(1)}, \ldots, \mathbf{g}^{(T)}$ satisfying $\mathbb{E}[\mathbf{g}^{(i)} \mid \mathbf{x}^{(i-1)}] = \nabla f(\mathbf{x}^{(i-1)}) + \mathbf{b}^{(i)}$. Let $\mathbf{x}_* = \arg\min_{\mathbf{x} \in K} f(\mathbf{x})$ be a minimizer of f. Then, if we assume:*

(i) ***Bounded step variance****: $\mathbb{E}[\|\mathbf{g}^{(i)}\|_2^2] \le \rho^2$,*

(ii) ***Strong convexity****: f is λ-strongly convex, and*

(iii) ***Bounded gradient bias****: $\|\mathbf{b}^{(i)}\|^2 \le \frac{\rho^2}{2\lambda \cdot diam(K) \cdot i}$,*

then the average iterate $\widehat{\mathbf{x}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}^{(T)}$ satisfies $\mathbb{E}[f(\widehat{\mathbf{x}}) - f(\mathbf{x}_*)] \le \frac{\rho^2}{\lambda T}(1 + \log(T))$.

We study each of the three conditions in Lemma 12 above, before wrapping up with the overall analysis.

7.4.3.1 Strong Convexity

To show convergence of stochastic gradient descent on ℓ, we require *strong convexity* such that the optimum of ℓ is unique. Specifically, we need to show: $\boldsymbol{\nabla}^2 \ell(\boldsymbol{\mu}) \succeq \beta I$ for some parameter $\beta > 0$ such that the probability mass is at least a

constant. Once we proved the strong convexity, we can apply projected stochastic gradient descent (PSGD) to recover the parameter μ.

Lemma 13 (Strong Convexity with Missing Entries)**.** *Given our missingness model and assumption 7.1.3 with $\beta = \frac{c}{d}$ for some integer $c \in \{1, \dots d\}$, we have that the function with general covariance $\ell(\boldsymbol{\mu})$ is λ-strongly convex for $\lambda = \alpha\beta/\lambda_{\max}(\boldsymbol{\Sigma})$.*

7.4.3.2 Bounded Step Variance and Gradient Bias

In this section, we analyze Algorithm 12, `SampleGradient` with an illustration of the relationship between convex sets appeared in this section is available in Fig. E.2. We first study the distribution of $\mathbf{z}^{(M)}$, which then will allow us to show an additive approximation guarantee for the output of the algorithm.

Theorem 7.4.2. *Assume $\|\boldsymbol{\mu}^* - \boldsymbol{\mu}\|_{\boldsymbol{\Sigma}} \le S$. For $R = \tilde{O}(\sqrt{d} + S + \log(1/\gamma\varepsilon))$, if $M = \operatorname{poly} d, S, 1/\gamma, 1/\varepsilon$ and $\eta = \tilde{\Theta}(R^2/M)$, then*

$$\mathrm{d_{TV}}(\mathbf{z}^{(M)}, \mathcal{N}(W^{-1}\boldsymbol{\mu}_{cond}, I)) \le \varepsilon.$$

Fix $(A, \mathbf{x})$. Without loss of generality, assume $\boldsymbol{\mu}^* = \mathbf{0}$.

Corollary 7.4.3. *Let $\hat{\mathbf{g}}$ be the output of Algorithm 12 with inputs $\tilde{\mathbf{x}}$ and $\boldsymbol{\mu}$ and parameters R, M, η as in Theorem 7.4.2. Also, let $\mathbf{g} = -\mathbb{E}_{\mathbf{y}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}) \mid \mathbb{S}(\mathbf{y}) = A, \mathbf{y}_A = \mathbf{x}]$. Then, we have that:*

$$\|\mathbb{E}[\hat{\mathbf{g}}] - \mathbf{g}\|_2 \le \varepsilon \cdot \operatorname{poly} S, d, 1/\gamma, 1/\varepsilon, \lambda_{\max}, 1/\lambda_{\min} \tag{7.4}$$

Furthermore, we have the following bound

$$\mathbb{E}[\|\hat{\mathbf{g}}\|_2^2] \le \operatorname{poly} d, 1/\gamma, S, 1/\lambda_{\min} \tag{7.5}$$

7.5 Discussion and Future Work

In the context of linear-thresholding missingness with a known covariance matrix, we can obtain the mean by initially observing that the set $\mathbf{y} : \mathbb{S}(\mathbf{y}) = A \wedge \mathbf{y}_A = \mathbf{x}$ is convex for any set A and any $\mathbf{x} \in \mathbb{R}^{|A|}$. By leveraging the fact (Corollary 2.1 of [KP77]) that the variance of a Gaussian decreases when conditioned on a convex

set, we can establish that the Hessian of our likelihood function is positive definite. This property ensures that our objective function is strongly convex and thus we can learn the distribution from a MNAR model. However, in scenarios where the covariance matrix is unknown, recovering the distribution becomes much more challenging as the Hessian of our likelihood function incorporates a fourth moment. Thus, we leave this as our future work.

Chapter 8

Toward Universal Laws of Outlier Propagation

8.1 Introduction

Anomaly detection plays a crucial role in business, technology, and medicine. Typical use cases range from fraud detection in finance and online trading [Don04], performance drops in manufacturing lines [STB17] and cloud computing applications [Gan+21; Ma+20; Har+23], health monitoring in intensive care units [Mas+16], to explaining extreme weather and climate events [ZSA22]. It has motivated a vast effort towards developing methodologies relevant to outlier analysis. As example, we refer the reader to some of the early works in statistics and computer science [Fre95; RW96; RL03; Agg17]. In complex systems, an anomaly will typically cause a large cascades of related anomalies [Pan+22]. In order to mitigate them, it is not sufficient to merely *detect* the anomalies; we must also identify which of the anomalies was the root cause [Bud+22; Ikr+22; Li+22; Har+23; Wan+23b; Wan+23a]. Thus, we implicitly face the *counterfactual* question of what conditions could have been different to prevent the (usually undesired) anomalous event.

To render a complex system accessible to human understanding, we begin with a causal model of its relevant mechanisms, specifying not only their default behavior, but also their behavior under modifications called *interventions*. Such a model should be modular in two respects. First, we may want to understand the *causal pathway*, along which a perturbation of any part of the system propagates through its components until it generates the event. Second, we want to "blame" some component(s) of the system, while acknowledging that others worked as expected.

Causal Bayesian networks offer a framework that supports both kinds of modular description, specifying causal relations via a directed acyclic graph (DAG) G with random variables $X_1, \ldots, X_n$ as nodes [Pea09; PCR93]. Under the causal Markov condition [Pea09], the joint distribution factorizes according to

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_j \mid \mathrm{Pa}_j), \tag{8.1}$$

where Pa_j denotes the parents of X_j in G, i.e., its direct causes. We will think of each conditional distribution $P(X_j \mid \mathrm{Pa}_j)$ as an *independent mechanism* of the system, which can in principle be changed or replaced without changing the others (see 2.1 and 2.2 in [PJS17] for a historical overview). Following this modular view, our root cause analysis will assume that in the case of an anomalous observation, most of the mechanisms worked as expected; thus, the anomaly can be blamed on a small number of mechanisms that act as root causes, in alignment with [Sch+21]'s "sparse mechanism shift hypothesis".

To our knowledge, [Bud+22] provide the most elaborate formalization of the idea of *attributing anomalies to mechanisms*. We develop our concepts starting from this baseline.

8.1.1 Outlier scores from p-values

To quantitatively attribute an anomalous event to upstream nodes, [Bud+22] first introduce what they call an Information Theoretic (IT) outlier score[1] via

$$\lambda_\tau(x) := -\log P(\tau(X) \geq \tau(x)), \tag{8.2}$$

where x denotes an observation of the random variable X, and $\tau : \mathcal{X} \to \mathbb{R}$ is an appropriate feature statistic, whose choice we discuss later.

λ_τ can be viewed as a statistical test of the null hypothesis that x was sampled from P: setting the base of the logarithm to 2 yields the p-value $2^{-\lambda_\tau(x)}$. A small p-value (or large λ_τ) corresponds to an unusual sample under P, which can be labeled an outlier. Since x is a single observation, anomaly scoring thus reduces to classical hypothesis testing on a sample size of 1 [She20; Vov20][2].

[1] While [Bud+22] use the natural base e, we use base 2 logarithms to align with binary program lengths in algorithmic information theory. In effect, we express the outlier score in units of bits [Fra05].

[2] Note that [TPS12] discuss anomalies as change points over multiple observations, whereas we focus on anomalies confined to an individual observation.

8.1.2 Quantitative root cause analysis in causal Bayesian networks

To quantitatively attribute an anomalous observation $(x_1, \ldots, x_n)$ to different mechanisms, where each x_j represents a value of the corresponding variable X_j, the arXiv version of [Bud+22] define *conditional outlier scores*:

$$\lambda_\tau(x_j \mid \mathrm{pa}_j) := -\log P(\tau(X_j) \geq \tau(x_j) \mid \mathrm{Pa}_j = \mathrm{pa}_j).$$

They demonstrate that this score can be equivalently interpreted as measuring the anomalousness of the noise term[3]. The feature functions τ_j can be node-specific, which is essential when the variables X_j have different characteristics (e.g., different dimensionality or data types). When the specific choice of τ is not crucial for the discussion, we will simplify notation by dropping the subscript. The arXiv version of [Bud+22] extends this framework by introducing a joint outlier score through "convolution":

$$\begin{aligned} \lambda(x_1, \ldots, x_n) \;&:=\; \sum_{j=1}^{n} \lambda(x_j \mid \mathrm{pa}_j) \\ &-\; \log \sum_{i=1}^{n-1} \frac{(\sum_{j=1}^{n} \lambda(x_j \mid \mathrm{pa}_j))^i}{i!}. \end{aligned} \tag{8.3}$$

The second term serves as a correction factor that ensures the score maintains the properties of an IT score, provided that the conditional distributions have densities with respect to the Lebesgue measure.

8.1.3 Monotonicity of scores

[Oka+24] formalize an intuitive principle about anomaly propagation: unless the connecting mechanisms themselves are also anomalous, a moderate outlier in a cause (measured by $\lambda(x_1)$) should not produce an extreme outlier in its effect (measured by $\lambda(x_2)$). This can be demonstrated for the bivariate causal relation $X_1 \to X_2$ as follows:

Lemma 14 (Weak IT outliers rarely cause strong ones)**.** *If λ are IT scores and X_1 is a continuous variable, the following inequality holds:*

$$P(\lambda(X_2) \geq \lambda(x_2) \mid \lambda(X_1) \geq c) \leq 2^{c-\lambda(x_2)}.$$

[3]This aligns in spirit with [VMB23], where events are also backtracked by attributing them to the noise variables.

In other words, whenever we generate an anomaly x_1 randomly from $P(X_1 \mid \lambda(X_1) \geq c)$ and then generate X_2 via its mechanism $P(X_2 \mid X_1)$, the resulting outlier score is unlikely to be much larger than c. The intuitive reason is that the probability occupied by events with high scores is much smaller than that of a given event having a lower score.

Note that there is no statement for any single x_1 – some of them may generate outliers that are higher with large probability – but "most" of them do not. The reason that we cannot make any statement about particular values is crucial for the present paper, because it shows how independence of mechanisms comes into play: whenever one chooses, on purpose, an x_1 such that it results in a large outlier downstream, there is no reason that this particular x_1 needs to have high outlier score. One can easily think, for instance, of the deterministic relation $X_2 = f(X_1)$, where f is a highly non-linear function that results in rather anomalous values x_2 for some tiny region of values x_1 which are not anomalous.

8.1.4 Limitations of current approach

Rigid definition of outliers λ_τ defines outliers in terms of a feature statistic τ that is chosen in advance. Sometimes, we only know what makes a sample abnormal after seeing it, so we would like to be able to choose τ with hindsight. For example, suppose P is univariate Gaussian. It seems natural to define $\tau(x) := |x - \mu|$, so that the anomaly detector flags extremely high or low values. However, we might also like to flag observations such as $x = 0$ or $x = \mu$: while such x may lie in a region of high density, the probability of obtaining such a specific value is zero under P. Although any other value has also probability zero, one may agree that observing $x = 0$ and $x = \mu$ would be surprising because these values are *special* (in a sense specified later).

As another example, suppose each component of a multivariate observation x is the reading of a different sensor. The signal transmission from all d sensors may be broken in such a way that all components equal to the same constant "idle" state c. Such coincidences also indicate an unusual event.

[Agg16] describe a broad variety of different outliers beyond the above toy examples: These can be unusual frequencies of words in text documents [MN20], un-

expected patterns in images [Yak+21], unusually large cliques in graphs [Hoo+16], or points lying in low density regions [Bre+00].

No general decomposition rule While Equation 8.3 nicely decomposes the joint outlier score into mechanism-specific conditional scores, a key limitation is that this decomposition relies on defining the joint score based on a sum of conditional scores. This does not imply that any reasonable outlier score (e.g., using a generic feature function τ) for the joint observation can be decomposed in this manner.

8.1.5 Our contributions

The contribution of this paper is purely conceptual. We do not propose another outlier detection or root cause analysis method, but instead provide a theoretical framework for calibrating and interpreting outlier scores. The framework ensures that outlier scores meet three critical criteria: (i) Comparability across diverse probability spaces and data modalities. (ii) Non-increasingness along causal chains of downstream effects, regardless of variable modalities. (iii) Well-defined attribution of joint system outlier scores to anomalies of mechanisms.

Our ideas were guided by the following general working hypothesis, which we believe applies far beyond the subject of this paper:

Principle 8.1.1 (Information Theory as a Guide)**.** *Good information theoretic concepts enable many nice theorems, but they are often hard to work with in practice. However, together with distributional assumptions (e.g. Gaussianity) they can boil down to simple concepts (e.g. linear algebraic expressions). The resulting formulae may be valid and useful beyond the distributional assumptions (e.g. by virtue of linear algebra).*

While Shannon information is sometimes hard to estimate from small sample sizes, *algorithmic* information is even worse: Kolmogorov complexity is not even computable. Furthermore, its identities hold only up to machine-dependent additive constants. Therefore, we owe it to the reader to show that our algorithmic information theoretic concepts trigger insights that can be applied in practice, as we will try in Section 8.5.

The paper is structured as follows. Section 8.2 describes concepts from statistics, information theory, and causality that we build upon. Section 8.3 derives the decomposition, and Section 8.4 shows that the joint score is non-increasing under marginalization. Section 8.5 discusses simple examples and Section 8.6 a toy experiment. For a cleaner exposition, we defer some formal proofs and definitions to the supplementary material.

8.2 Key ingredients

8.2.1 Statistical testing with e-values instead of p-values

While p-values are the most famous measure of evidence in statistical testing, e-values are recently gaining popularity for their superior ability to aggregate evidence across multiple tests [RW24]. These values are inconsistently scaled in the literature, with e-values being comparable to reciprocals of p-values and exponentials of algorithmic randomness scores. To remove any obfuscation coming from scaling conventions, we introduce the following definitions.

Definition 8.2.1. *A probability-bounded test (**p-test**) in ratio form is a statistic* $\Lambda : \mathcal{X} \to [0, \infty]$ *satisfying* $\forall \varepsilon > 0, P\left(\Lambda(X) \geq 1/\varepsilon\right) \leq \varepsilon$. *An expectation-bounded test (**e-test**) in ratio form is a statistic* $\Lambda : \mathcal{X} \to [0, \infty]$ *satisfying* $\mathbf{E}_{X \sim P}\left(\Lambda(X)\right) \leq 1$.

We say a statistic $\Lambda : \mathcal{X} \to [0, \infty]$ is a p-test (or e-test) in *probability form*, if $1/\Lambda$ is a p-test (or e-test) in ratio form. Note that what is commonly called a "p-value" is a p-test in probability form, satisfying $\Lambda(X) \leq \epsilon$ with probability at most ϵ. In contrast, what the literature calls an "e-value" is an e-test in *ratio* form.

Similarly, a statistic $\lambda : \mathcal{X} \to [-\infty, \infty]$ is a p-test (or e-test) in *log form*, if 2^{λ} is a p-test (or e-test) in ratio form. Our convention is to use lowercase symbols like λ to indicate log form.

By Markov's inequality, every e-test is also a p-test. Conversely, [VW21] describe a number of ways to *calibrate* any given p-test into an e-test. For more details on tests, see Appendix F.1 in the supplementary material.

p-tests provide a straightforward way to control the Type I error rate, i.e., false positives. To ensure that samples from P are labeled as anomalies at a rate no

larger than a desired threshold ϵ, we flag only those samples whose p-test scores are above $1/\epsilon$ (i.e., below ϵ when expressed in probability form).

Since every e-test is also a p-test, e-tests also achieve this false positive rate; however, they are more conservative. In return, e-tests offer many advantages related to composability and optional stopping [GHK20; Ram+23; RW24], and we will find them more convenient for decomposing anomaly scores by mechanism.

8.2.2 Basic notions from algorithmic information theory

The intuition behind universal tests is that an observation is anomalous precisely when it is more compressible than usual for the underlying random process. To formalize description lengths, we fix a universal prefix-free Turing machine. The conditional Kolmogorov complexity $K(x \mid y)$ is the bit length of the shortest program p that outputs x when given access to another string y [LV97]. It satisfies the Kraft inequality $\sum_x 2^{-K(x|y)} < 1$. When y is an empty string, we write $K(x)$. Just as Shannon's entropy measures information content for a probability distribution, $K(x)$ measures it for an individual sample x.

Using a standard prefix-free encoding of n-tuples, [LV97] also define the joint Kolmogorov complexity $K(x_1, \ldots, x_n)$. By analogy with Shannon's mutual information, they define the algorithmic mutual information between x and y, conditional on z^*, by

$$\begin{aligned} I(x : y \mid z^*) &:= K(x \mid z^*) + K(y \mid z^*) - K(x, y \mid z^*) \\ &\stackrel{+}{=} K(x \mid z^*) - K(x \mid (y, z)^*). \end{aligned}$$

Here, z^* denote a shortest program that outputs z, and $\stackrel{+}{=}$ denotes equality up to a constant dependent on the universal machine, but not on x or y[4]. We say x and y are conditionally independent, given z, if $I(x : y \mid z^*) \stackrel{+}{=} 0$.

8.2.3 Universal tests

A Martin-Löf or Levin test (i.e., semicomputable p-test or e-test, respectively) can be thought of as combining all computable anomaly scoring algorithms, in

[4]Note that x^* contains more information than x, because x is easily generated from x^* but not vice versa. Conditioning on x instead of x^* would result in error terms that are often constant, and at worst logarithmic in the length of x [Gác21].

order to detect the broadest possible variety of anomalies. Appendix F.2 in the supplementary material contains full details; here we give a brief overview.

Definition 8.2.2 (Domination property)**.** *For two statistical tests λ_1 and λ_2 expressed in log form, we say* λ_1 dominates λ_2 *if there exists a constant $c \in \mathbb{R}$ such that for all observations x in the sample space,*

$$\lambda_1(x) \geq \lambda_2(x) - c.$$

Intuitively, this means λ_1 can detect at least all the anomalies that λ_2 can detect, up to a constant term.

The domination property provides a natural way to compare the power of different statistical tests. The class of semicomputable tests contains a *universal* test that dominates all the others:

Theorem 8.2.3 (Universality of randomness deficiency)**.** *Let $\mathcal{X}$ be a discrete space that can be interpreted as a subset of $\{0,1\}^*$ in a canonical way, and P be a computable probability distribution on $\mathcal{X}$ (i.e., with finite description length). Then, the randomness deficiency of $x \in \mathcal{X}$, defined by*

$$\delta(x) := -\log P(x) - K(x \mid P^*), \tag{8.4}$$

is a universal e-test, dominating all other semicomputable e-tests.

The intuition behind Eq. (8.4) is that typical samples from a distribution P are optimally compressed by encodings of length $-\log P(x)$. When a sample x can be compressed beyond this theoretical limit, we consider it "anomalous". This includes the case where we observe $x = 0$ from a discretized centered Gaussian distribution. Despite being the mode of the distribution, 0 is considered anomalous because its negative log-likelihood $-\log P(x)$ is substantially larger than its Kolmogorov complexity $K(0) \stackrel{+}{=} 0$.

Note that when a distributional estimate $\hat{P}$ is inferred from data, the "reconstruction loss" $-\log \hat{P}(x)$ can be used as an outlier score [Bis93]. Without further considerations, this score is not calibrated because there may be a huge region with

low density, and thus all low density points together may still be likely. However, if the term $K(x \mid P^*)$ is also small, we know that x is special within this huge set.[5]

Example 1 (Uniform distribution). When P is uniform over all 2^d binary strings of length d, (8.4) becomes $\delta(x) = d - K(x \mid P^*) \stackrel{+}{=} d - K(x \mid d)$, measuring the degree to which x is compressible.

8.2.4 Algorithmic Markov condition and independence of mechanisms

Our goal is to demonstrate how the randomness deficiency decomposes across causal mechanisms in a Bayesian network, subject to postulates that connect algorithmic information theory to causality. To achieve this, we establish a theoretical framework based on fundamental postulates.

[JS10] propose an adaptation of Equation 8.1 that characterizes algorithmic dependencies between *individual objects*, rather than statistical dependencies between *random variables*:

Postulate 8.2.4 (Algorithmic Markov condition). Let $x_1, \ldots, x_n$ be binary words describing objects whose causal relations are given by the DAG G. Then the joint complexity of $(x_1, \ldots, x_n)$ decomposes as

$$K(x_1, \ldots, x_n) \stackrel{+}{=} \sum_{j=1}^{n} K(x_j \mid \mathrm{pa}_j^*).$$

Furthermore, for any three sets R, S, T of nodes, we have

$$I(R : S \mid T^*) \stackrel{+}{=} 0,$$

whenever R and S are d-separated by T.

[JS10] argue that within a causal Bayesian network, where nodes represent random variables, the mechanisms $P_{X_j \mid \mathrm{Pa}_j}$ constitute independent information-bearing objects. Given m observations of n-tuples $(x_1^1, \ldots, x_n^1), \ldots, (x_1^m, \ldots, x_n^m)$, they define a DAG G^m that connects the n mechanisms $P_{X_i \mid \mathrm{Pa}_i}$ with the $n \times m$ observations in

[5] Note that conformal p-values [Bat+21] do not require any parametric assumptions or density estimation, since their coverage guarantees rely on exchangeability alone. However, statements on *conditional* coverage [Bar+19] are too weak for our purpose of calibrating *conditional* outlier scores.

the following way: Each x_i^l has the observations pa_i^l and the *node* $P_{X_i|\,\mathrm{Pa}_i}$ as parents, formalizing the idea that the mechanisms $P_{X_i|\,\mathrm{Pa}_i}$ also determine how the parents in G influence the respective observation. Fig. 8.1 shows G^m for the DAG $X \to Y$. For our purpose, it is sufficient to consider G^1, where we have only one observation x_j from each node X_j in G. Accordingly, we construct G^1 from G as follows: replace each variable X_j with its observation x_j, and make $P_{X_j|\,\mathrm{Pa}_j}$ a parent of X_j. Each $P_{X_j|\,\mathrm{Pa}_j}$ becomes a root node, and these are the only root nodes in G^1. This placement of $P_{X_j|\,\mathrm{Pa}_j}$ as root nodes (and thus algorithmically independent) formalizes the Independence of Mechanisms[6]. Applying the algorithmic Markov condition to G^1 yields:

Lemma 15 (Conditional irrelevance of other mechanisms and predecessors when parents are given)**.** *Let $X_1, \ldots, X_n$ be causally ordered. Given its parents* pa_j *and the mechanism $P_{X_j|\mathrm{Pa}_j}$, none of the other mechanisms $(P_{X_i|\mathrm{Pa}_i})_{i\neq j}$ and none of the causal predecessors $(x_i)_{i<j}$ enable further compression of x_j. That is,*

$$x_j \perp\!\!\!\perp (x_i)_{i<j}, (P_{X_i|\mathrm{Pa}_i})_{i\neq j} \mid (\mathrm{pa}_j, P_{X_j|\mathrm{Pa}_j})^*.$$

8.3 Decomposition of Randomness Deficiency

We start by presentation the decomposition of randomness deficiency in the bivariate case. Specifically, consider the DAG $X \to Y$ between a cause X and its effect Y. Extending Eq. (8.4) with each distribution's shortest representation, define the joint randomness deficiency of outcomes (x, y) with respect to $P_{X,Y}$ by

$$\delta(x, y) := -\log P(x, y) - K(x, y \mid (P_{X,Y})^*), \tag{8.5}$$

and the conditional randomness deficiency by

$$\delta(y \mid x) := -\log P(y \mid x) - K(y \mid (x, P_{Y|X})^*). \tag{8.6}$$

Lemma 16. *(Decomposition of randomness deficiency for a cause-effect pair) For any two random variables $X \to Y$ (i.e., X being the cause of Y), and for specific*

[6] [JS10] call the algorithmic independence of different $P_{X_j|\mathrm{Pa}_j}$ *Independence of Mechanisms*. Here we prefer using this term for the underlying idea of representing them as root nodes.

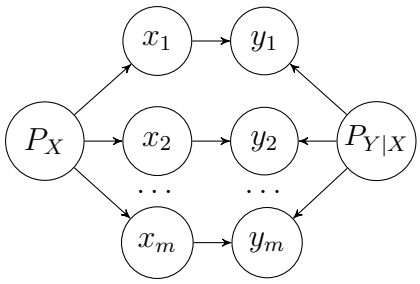


Figure 8.1: $x_1, \ldots, x_m$ sampled from P_X, $y_1, \ldots, y_m$ sampled from $P_{Y|X}$.

observations x and y, the following equality holds under Postulate 8.2.4 for the DAG in Fig. 8.1:

$$\delta(x, y) \stackrel{+}{=} \delta(x) + \delta(y \mid x)$$

The proof relies on Lemma 15, and is provided in Appendix F.3 of the supplementary.

The following example shows why causal assumptions are needed for randomness deficiency to be additive.

Example 2 (No additivity without Independence of Mechanisms). Let $\mathcal{X} = \mathcal{Y} = \{0,1\}^d$ and P_X be the uniform distribution. Furthermore, let $P_{Y|X}$ be the deterministic mechanism

$$P(y \mid x) = 1 \quad \text{iff} \quad y = x \oplus x_0,$$

where $\oplus$ denotes bitwise XOR and x_0 is an algorithmically random string with $K(x_0) = d$. The mechanism $P_{Y|X}$ being deterministic implies that we always have $\delta(y \mid x) \stackrel{+}{=} 0$.

Consider an input $x = x_0$ that violates the conditional independence relation $x \perp\!\!\!\perp P_{Y|X} \mid P_X$. Then, $\delta(x) \stackrel{+}{=} 0$ because x is random with respect to P_X, but $\delta(x, y) \stackrel{+}{=} d$ because $(x, y) = (x_0, 0^d)$ is a simple function of x_0, which is easily deciphered from $P_{X,Y}$. Since x is non-generic relative to $P_{Y|X}$, we find that the randomness deficiency of (x, y) cannot be attributed to x, nor to the mechanism generating y from x.

We now generalize Lemma 16 to the multivariate case:

Theorem 8.3.1 (Decomposition of multivariate joint randomness deficiency)**.** *Let the set of strings $x_1, x_2, \ldots, x_n$ be causally connected by a directed acyclic graph*

G, so that the causal Markov condition holds for G^m. Then the joint randomness deficiency of all strings $x_1, x_2, \ldots, x_n$ decomposes into the conditional randomness deficiencies of the mechanisms:

$$\delta(x_1, \ldots, x_n) \stackrel{+}{=} \sum_{j=1}^{n} \delta(x_j \mid \mathrm{pa}_j),$$

where $\delta(x_j \mid \mathrm{pa}_j)$ denotes the randomness deficiency of x_j given its parents.

The proof is a simple induction over the number n of nodes according to the causal ordering. The induction step from $n-1$ to n uses Lemma 16 where $y := x_n$ and x is the single multivariate parent $x := (x_1, \ldots, x_{n-1})$; see Appendix F.3 in the supplementary for details.

8.4 Monotonicity of Randomness Deficiency

In particular, Theorem 8.3.1 allows us to draw the following straightforward conclusion:

Theorem 8.4.1 (Weak anomalies do not cause stronger ones)**.** *If there is a unique root cause $j \in \{1, \ldots, n\}$, in the sense that*

$$\delta(x_i \mid \mathrm{pa}_i) \stackrel{+}{=} 0 \quad \text{for } i \neq j,$$

and the conditions of Theorem 8.3.1 are met,

$$\text{then,} \qquad \delta(x_i) \stackrel{+}{\leq} \delta(x_j \mid \mathrm{pa}_j) \quad \forall i \in \{1, \ldots, n\}. \tag{8.7}$$

The proof follows easily from Theorem 8.3.1 together with Corollary 4.1.11 from [Gác21] which states non-increasingness of δ under marginalization and thus

$$\delta(x_i) \stackrel{+}{\leq} \delta(x_1, \ldots, x_n).$$

In essence, Equation 8.7 states that none of the nodes displays a randomness deficiency that exceeds the conditional randomness deficiency of the root cause. In a nutshell, "weak outliers cannot cause strong ones". – We have thus found an AIT version of Lemma 14.[7] It is worth noting that [Lev84] (see also [Gác21]) demonstrated

[7] Hence, in the language of modern resource theories in physics [CFS16], Definition 5.1., an anomaly is a "resource" and δ a "monotone" measuring its value.

a similar principle, calling it *randomness conservation*. It says that the output of a mechanism cannot exhibit a substantially larger randomness deficiency than the input - under the condition that the mechanism itself has a constant description length. In physics, randomness deficiency corresponds to a lack of entropy [Zur89; Gác94]. Therefore, the second law of thermodynamics amounts to an instance of randomness conservation [EH25].

In our context, *simplicity* of mechanisms is generalized by the *independence* of mechanisms. Example 2 illustrates why the independence of mechanisms principle is essential. It describes a scenario where a value y with randomness deficiency d emerges, yet this deficiency cannot be attributed to either the input x or the mechanism generating y from x.

8.5 Relation to computable anomaly scores

This section describes simple scenarios in which randomness deficiency boils down to simple and well-known scores. One may ask why one should start with an uncomputable concept in the first place to end up with something simple anyway. The answer is that the properties we have shown for randomness deficiency guide us in calibrating simple scores, with desirable properties such as *decomposition of joint scores into scores of the mechanisms* and *monotonicity of scores*.

Example 3 (z-score). For a Gaussian variable $X \sim \mathcal{N}(\mu, \sigma)$, the squared z-score reads $z^2(x) := (x-\mu)^2/\sigma^2$. At a fixed level of precision, $K(x \mid P^*) \stackrel{+}{\leq} 2\log|x-\mu|$. Substituting the log likelihood

$$-\log P(x) = \log\sqrt{2\pi}\sigma + \frac{\log e}{2} z^2(x),$$

and treating σ as a constant, yields

$$\delta(x) \stackrel{+}{\geq} \frac{\log e}{2} z^2(x) - 2\log|x-\mu|.$$

Note that while the well-known identity $\mathbb{E}[z^2] = 1$ tells us that z^2 is an e-value, the universal e-value $2^\delta \approx e^{z^2/2}$ provides far stronger evidence at the tails. Now we generalize to $n > 1$ dimensions:

Example 4 (Decomposing squared Mahalanobis distance). The randomness deficiency of a random vector $\mathbf{x} \in \mathbb{R}^n$, drawn from a centered Gaussian, satisfies

$$\delta(\mathbf{x}) \stackrel{+}{\geq} \frac{\log e}{2} \mathbf{x}^T \Sigma_{\mathbf{X}}^{-1} \mathbf{x} - O(\log \|\mathbf{x}\|_\infty),$$

where $\Sigma_{\mathbf{X}}$ denotes the covariance matrix and n is treated as a constant. The leading term is $(\log e)/2$ times the squared Mahalanobis distance (M-distance, for short).

Let $\mathbf{X}$ be generated by a causal Bayesian network with linear structural equations $\mathbf{X} = A\mathbf{X} + \mathbf{N}$, where A is strictly lower triangular and N_i are independent noise variables with variance σ_i^2. Using transformation $\mathbf{N} = (I - A)\mathbf{X}$, we obtain a diagonal form

$$\mathbf{x}^T \Sigma_{\mathbf{X}}^{-1} \mathbf{x} = \sum_{i=1}^{n} \frac{n_i^2}{\sigma_i^2}.$$

For large n_i, the expression n_i^2/σ_2 is roughly proportional to the randomness deficiency of the mechanism $P(X_i \mid PA_i)$. Hence, decomposition of randomness deficiencies translates asymptotically into decompositions of the squared M-distance (as used for multivariate anomaly detection [Agg16]) into z^2-scores. Further, M-distance is non-increasing with respect to marginalization to a subset of variables (see Appendix F.5 in the supplementary), resembling the monotonicity of randomness deficiency.

Example 4 supports Principle 8.1.1: a conservative bound on the randomness deficiency provides us with a computable e-value, as a function of the z^2-scores. Moreover, the conclusion that the z^2-score of any variable cannot exceed the sum of all these "noise scores" still holds, because our reasoning above is entirely based on linear algebra.

The following example shows that root causes may have smaller *marginal* randomness deficiency than their downstream effects. In particular, while a root cause may have a low outlier score, its conditional outlier score can be substantially higher. This discrepancy can then result in a downstream outlier whose score is larger than the score of the root cause, but not larger than the conditional score of the root cause.

Example 5 (Root cause with small marginal score). With $N_i \sim \mathcal{N}(0, 1)$, consider the three-node model

$$X_1 := N_1,$$

$$X_2 := 2X_1 + N_2,$$
$$X_3 := X_1 - X_2 + N_3.$$

Then, $X_2 \sim \mathcal{N}(0, \sqrt{5})$. Suppose n_1, n_3 take on typical values, while n_2 is anomalously large. Then, the marginal randomness deficiency $\delta(x_2)$ increases with $n_2^2/5$, while the conditional randomness deficiency $\delta(x_2 \mid x_1)$ increases with $n_2^2/1$ (where we've ignored the factor of $\frac{\log e}{2}$ from **??**). Since $X_3 \sim \mathcal{N}(0, \sqrt{3})$, $\delta(x_3)$ increases with $n_2^2/3$. Thus, for large perturbations of x_2, we have $\delta(x_2) \ll \delta(x_3) \ll \delta(x_2 \mid x_1) \approx \delta(x_1, x_2)$. Looking only at marginal scores, it would seem as though x_2 caused a stronger outlier x_3. Nonetheless, $\delta(x_3) \overset{+}{\leq} \delta(x_2 \mid x_1)$ in agreement with Theorem 8.4.1.

This paradox resembles Example 2.1 in [Li+24], where it is phrased in terms of z^2-scores. Following Principle 8.1.1, we can describe it entirely in terms of M-distances (which are just z^2-scores in the one-dimensional case): the M-distance of x_3 can be larger than the M-distance of its root cause x_2, but not larger than the M-distance of the pair (x_1, x_2), i.e., the full cause of x_3. In other words, to fully quantify the anomaly introduced by perturbing x_2, we must also account for the destroyed coupling between x_1 and x_2, not only for the size of the value x_2 itself.

The insight from algorithmic information theory is that the anomaly scores of the vector (x_1, x_2) and the scalar x_3 are indeed comparable, and also comparable to the conditional score of x_2 given x_1, because their calibration was guided by the randomness deficiency.

In the following example, which could be easily generalized to words over an arbitrary alphabet, randomness deficiency essentially boils down to relative entropy:

Example 6 (m-bit binary word). Let us first consider an m-bit word whose bits are set to 1 independently with probability p. Counting the number of words w with fixed Hamming weight ℓ yields the lower bound

$$\delta(w) \overset{+}{\geq} -\ell \log p - (m - \ell) \log(1 - p) - \log \binom{m}{\ell} - O(\log m).$$

Using Stirling's approximation $\log m! = m \log m - m \log e + O(\log m)$, one can show that

$$\delta(w) \overset{+}{\geq} m \cdot D_{\mathrm{KL}}\left(\frac{\ell}{m} \| p\right) - O(\log m), \tag{8.8}$$

with the binary Kullback-Leibler distance

$$D_{\mathrm{KL}}\left(q\|p\right) := q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}.$$

Here we see that words with unexpectedly low or unexpectedly Hamming weight are both assigned high outlier scores, enabling us to compare both types of outliers. This works despite the probability mass function being monotonic in the Hamming weight ℓ.[8]

Many works [Agg16; Ako+12; NC03] propose detecting anomalies by compression, but consider anomalies that have *higher* compression length than usual, seemingly in conflict with the randomness deficiency which flags *low* compression lengths. Example 6 resolves this paradox: for $p < 1/2$, any word with $K(w) > m \cdot H(p)$ must have Hamming weight larger than mp. Accordingly, its randomness deficiency is bounded from below by Equation 8.8. Hence, paradoxically, an unusually *high* compression length also implies non-zero randomness deficiency, because it entails an increase of the log likelihood term that outweighs the observed increase in compression length.

More sophisticated notions of anomalies can be obtained, for instance, when X is graph-valued and an anomaly is given by large cliques [Agg16] – e.g. when fraudsters work together frequently on illegal activities, their communication is densely connected. To estimate the random deficiency of a graph G with a clique of size k, it suffices to define a probability distribution on the set of graphs with m nodes, and bound a graph's description length by counting the number of graphs with cliques of size k.

8.6 Experiment with Lempel-Ziv Compression

So far we have accounted for the term $K(x \mid P^*)$ only by very rough upper bounds rather than trying to approximate it. Here we describe a toy scenario in which we can detect anomalies using the Lempel-Ziv compression algorithm; more difficult scenarios may demand more powerful compression algorithms that are closer to general AI.

[8] [CT06b], in Section 11.2, call a word "ϵ-typical" when it satisfies $D_{\mathrm{KL}}\left(\frac{\ell}{m}\|p\right) \leq \epsilon$. Their equation (11.67) finds that such a word occurs with probability at least $1 - (m+1)^2 \cdot e^{-\epsilon m}$.

Consider the causal DAG: $X_1 \to X_2 \to \cdots \to X_n$, with the structural equations

$$X_1 = N_1; \quad X_j = X_{j-1} + N_j \quad \text{for } j = 2, \ldots, n. \tag{8.9}$$

Moreover, suppose that every N_j for $j = 1, \ldots, n$ is drawn from the uniform distribution on the set of numbers in $[0, 1]$ discretized to $d \gg 1$ digits of precision. By choosing the uniform distribution, the terms $-\log P(x_j \mid x_{j-1})$ become constant; moreover, we have $K(x_j \mid (x_{j-1}, P)^*) \stackrel{+}{=} K(x_j \mid x_{j-1})$. Since the conditional distribution of each X_j, given its parent, is uniform over the (discretized) interval $[x_{j-1},\, x_{j-1} + 1]$, the conditional randomness deficiency of x_j reads

$$\delta(x_j \mid x_{j-1}) \stackrel{+}{=} d \cdot \log 10 - K(x_j \mid x_{j-1}).$$

Suppose now we inject an anomaly at some node j by setting n_j to some numbers in $\{0, \ldots, 0.9\}$. As a result, x_j and x_{j-1} now coincide in at most 2 digits, so that $K(x_j \mid x_{j-1}) \stackrel{+}{=} 0$ and $\delta(x_j \mid x_{j-1}) \stackrel{+}{=} d \cdot \log 10$.

Following [SJS10], we approximate $K(x_j \mid x^*_{j-1}) \stackrel{+}{=} K(x_j, x_{j-1}) - K(x_{j-1})$ by $R(x_j, x_{j-1}) - R(x_{j-1})$, where R denotes the length of a compressed encoding using the Lempel-Ziv algorithm. Whenever n_j corrupts to a 1-digit number, Lempel-Ziv recognizes that x_j and x_{j-1} coincide with respect to all but 1 or 2 digits. Thus, $R(x_j, x_{j-1}) \approx R(x_{j-1})$, which lets us infer the randomness deficiency of almost $d \log 10$.

We conducted experiments to verify our findings for $n = 4$, with the noise uniformly drawn from numbers in $[0, 1]$ with $d = 10$ digits. We randomly chose one of the 4 nodes as "root cause" and set all but one digit of its noise variables to zero. To detect the root cause, we selected the label j that minimized Lempel-Ziv compression length and found the right one in 100 out of 100 runs.

From a theoretical point of view, the example shows that the joint observation can be anomalous (here in the sense of showing two variables whose digits coincide largely), although none of the variables show unexpected behaviour with respect to their *marginal* distribution. Hence, the root cause can only be found by inspecting which *mechanism* behaves unexpectedly.

8.7 Conclusion and Future Work

Algorithmic randomness deficiency offers a principled and flexible definition of outliers, without prior specification of the feature that exposes the anomaly. On a causal Bayesian network, we saw that the randomness deficiency of a joint observation decomposes along individual causal mechanisms, subject to the Independent Mechanisms Principle. This allows us to trace anomalous observations back to their root causes, identifying specific mechanisms that exhibit atypical behavior. Furthermore, we showed that weak outliers cannot be responsible for producing strong outliers, thus extending Levins law of randomness conservation. This foundational insight can help calibrating anomaly scores in a way that supports root cause analysis in complex systems.

Chapter 9

Conclusion and Future Work

9.1 Summary of Contributions

This thesis developed a unified framework for high-dimensional inference, bridging the gap between foundational theory and the complexities of real-world data. The work's contributions fall into two complementary themes: establishing optimal performance guarantees under idealized settings and deploying robust methods to maintain valid inference under data bias.

Part I: Learning, Testing, and Inference from Structured Models (Chapters 3 – 5) This section established a theoretical bedrock by characterizing the limits of inference on structured models.

- We derived the first general and rigorous identifiability conditions for linear Andersson-Madigan-Perlman (AMP) chain graph models [WB21b] (Chapter 3), a broad class generalizing linear structural equation models.
- We developed algorithms for learning Gaussian trees and polytrees that achieved minimax-optimal sample complexity for both distribution learning (in KL divergence) and exact structure recovery [Wan+24] (Chapter 4). The analysis provided explicit finite-sample guarantees and matching lower bounds, which were previously missing for the faithful learning setting.
- We introduced a mutual information tester for linear models with optimal $O(1/\varepsilon)$ sample complexity (Chapter 5), demonstrating a novel technique to bypass the difficult task of direct mutual information estimation.

- We introduced a principled framework using Algorithmic Information Theory (AIT) in Chapter 8 to quantify outliers in terms of randomness deficiency [**<empty citation>**]. This led to the discovery of universal laws governing outlier propagation in causal Bayesian networks, allowing anomalies to be traced back to their root mechanisms.

Part II: Learning, Testing, and Inference from Biased Data (Chapters 6 – 8) Building on the principles of optimal testing and statistical structure, this section introduced computationally tractable and robust procedures for handling common data imperfections.

- We provided the first sharp analysis for Gaussian mean testing under truncation (Chapter 6), showing a critical phase transition in complexity that depends on the relationship between truncation mass ε and accuracy α. Critically, we demonstrated that knowing the truncation set dramatically reduces the complexity to $\Theta(\sqrt{d})$ across all parameter regimes.
- For learning high-dimensional Gaussians from censored data (Chapter 7), we provided efficient algorithms for the difficult problem of estimation under Missing Not At Random (MNAR) settings, including self-censoring and linear thresholding models.
- We introduced a principled framework using Algorithmic Information Theory (AIT) in Chapter 8 to quantify outliers in terms of randomness deficiency [EWJ25]. This led to the discovery of universal laws governing outlier propagation in causal Bayesian networks, allowing anomalies to be traced back to their root mechanisms.

9.2 Limitations and Future Directions

The following points represent the primary limitations of this work, which also serve as promising avenues for future research.

9.2.1 Generalizing Model Assumptions

- **Non-Parametric Extension**: A core constraint, particularly in Part I, is the reliance on linear and Gaussian assumptions. A major challenge is extending the minimax optimality guarantees from Gaussian/linear models to more complex non-parametric or non-linear model families.
- **Robustness to Violations**: Future work must explore the stability of our optimal algorithms when explicit assumptions, such as tree-faithfulness (Chapter 4) or the log-concavity of noise (Chapter 5), are mildly violated in practice.

9.2.2 The Challenge of Unknown Bias Mechanisms

- **Learning the Bias Model**: The methods developed in Part II, while robust, often assume that the nature of the data bias (the truncation set S or the censoring mechanism) is explicitly known. The next frontier is developing algorithms that can simultaneously learn the model parameters and infer the underlying bias mechanism itself, a critical and difficult open problem.
- **Unknown Covariance**: Specifically in the censored data setting, extending the robust mean estimation (Chapter 7) to the case where the covariance matrix Σ is unknown remains a significant hurdle, as the log-likelihood function incorporates a higher-order moment term.

9.2.3 Practical Scalability and Experimental Validation

- **Beyond Synthetic Data**: The reliance on synthetic data for validation in several chapters (e.g., Chapter 3) indicates a need for real-world datasets that can rigorously test the performance of these algorithms.
- **Engineering for Scale**: While the algorithms are theoretically efficient (polynomial-time), translating them into practical implementations capable of handling web-scale or trillion-parameter datasets requires the adoption of distributed computation strategies.

- **Generalizing Causality Tools**: The AIT framework in Chapter 8 should be extended beyond simple chain structures to include more general graph types and be applied to real-world anomaly detection systems.

9.3 Concluding Remarks

The contributions of this dissertation provide perspective on the trade-offs between statistical reliability and computational efficiency in modern data analysis. By establishing theoretical limits and then building robust methods that respect those limits, we have advanced both the foundational understanding and the practical methodology for reliable inference from imperfect, high-dimensional data.

Bibliography

[ACT20] J. Acharya, C. L. Canonne, and H. Tyagi. "Distributed Signal Detection under Communication Constraints". In: *COLT*. Vol. 125. Proceedings of Machine Learning Research. PMLR, 2020, pp. 41–63.

[Ada+20] M. F. Adak, P. Lieberzeit, P. Jarujamrus, and N. Yumusak. "Classification of alcohols obtained by QCM sensors with different characteristics using ABC based neural network". In: *Engineering Science and Technology, an International Journal* 23.3 (2020), pp. 463–469.

[Ada+19] S. P. Adam, S.-A. N. Alexandropoulos, P. M. Pardalos, and M. N. Vrahatis. "No free lunch theorem: A review". In: *Approximation and optimization: Algorithms, complexity and applications* (2019), pp. 57–82.

[Agg17] C. C. Aggarwal. "An introduction to outlier analysis". Springer, 2017.

[Agg16] C. C. Aggarwal. "Outlier Analysis". 2nd. Springer Publishing Company, Incorporated, 2016. ISBN: 3319475770.

[Ako+12] L. Akoglu, H. Tong, J. Vreeken, and C. Faloutsos. "Fast and reliable anomaly detection in categorical data". In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. CIKM '12. Maui, Hawaii, USA: Association for Computing Machinery, 2012, pp. 415–424. ISBN: 9781450311564.

[All00] P. D. Allison. "Multiple imputation for missing data: A cautionary tale". In: *Sociological methods & research* 28.3 (2000), pp. 301–309.

[AE10] G. Altay and F. Emmert-Streib. "Inferring the conservative causal core of gene regulatory networks". In: *BMC systems biology* 4.1 (2010), pp. 1–13.

[Ame73] T. Amemiya. “Regression analysis when the dependent variable is truncated normal”. In: *Econometrica: Journal of the Econometric Society* (1973), pp. 997–1016.

[Ana+12] A. Anandkumar, D. Hsu, F. Huang, and S. Kakade. “Learning mixtures of tree graphical models”. In: vol. 2. cited By 15. 2012, pp. 1052–1060.

[Ana+11] A. Anandkumar, K. Chaudhuri, D. J. Hsu, S. M. Kakade, L. Song, and T. Zhang. “Spectral methods for learning multivariate latent tree structure”. In: *Advances in neural information processing systems*. 2011, pp. 2025–2033.

[AV13] A. Anandkumar and R. Valluvan. “Learning loopy graphical models with latent variables: Efficient methods and guarantees”. In: *The Annals of Statistics* (2013), pp. 401–435.

[And58] T. W. Anderson. “An introduction to multivariate statistical analysis”. Vol. 2. Wiley New York, 1958.

[AMP97] S. A. Andersson, D. Madigan, and M. D. Perlman. “A characterization of Markov equivalence classes for acyclic digraphs”. In: *The Annals of Statistics* 25.2 (1997), pp. 505–541.

[AMP01] S. A. Andersson, D. Madigan, and M. D. Perlman. “Alternative Markov properties for chain graphs”. In: *Scandinavian journal of statistics* 28.1 (2001), pp. 33–85.

[ABM08] P. Armitage, G. Berry, and J. N. S. Matthews. “Statistical methods in medical research”. John Wiley & Sons, 2008.

[ATB21] M. Azadkia, A. Taeb, and P. Bühlmann. “A Fast Non-parametric Approach for Local Causal Structure Learning”. In: *arXiv preprint arXiv:2111.14969* (2021).

[Bak+22] A. Bakshi, I. Diakonikolas, H. Jia, D. M. Kane, P. K. Kothari, and S. S. Vempala. “Robustly learning mixtures of k arbitrary gaussians”. In: *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*. 2022, pp. 1234–1247.

[BC14] N. Balakrishnan and E. Cramer. “The art of progressive censoring”. In: *Statistics for industry and technology* (2014).

[Bar64] A. J. Baranchik. “Multiple regression and estimation of the mean of a multivariate normal distribution.” Tech. rep. STANFORD UNIV CALIF, 1964.

[Bar+19] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. “The limits of distribution-free conditional predictive inference”. In: *Information and Inference: A Journal of the IMA* (2019).

[BS99] D. R. Barr and E. T. Sherrill. “Mean and variance of truncated normal distributions”. In: *The American Statistician* 53.4 (1999), pp. 357–361.

[Bat+21] S. Bates, E. J. Candès, L. Lei, Y. Romano, and M. Sesia. “Testing for outliers with conformal p-values”. In: *The Annals of Statistics* (2021).

[BS10] M. Belkin and K. Sinha. “Polynomial learning of distribution families”. In: *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE. 2010, pp. 103–112.

[Bel62] C. Bell. “Mutual information and maximal correlation as measures of dependence”. In: *The Annals of Mathematical Statistics* (1962), pp. 587–595.

[BMS20] R. Bhattacharya, D. Malinsky, and I. Shpitser. “Causal inference under interference and network uncertainty”. In: *Uncertainty in Artificial Intelligence*. PMLR. 2020, pp. 1028–1038.

[Bha+20] R. Bhattacharya, R. Nabi, I. Shpitser, and J. M. Robins. “Identification in missing data models represented by directed acyclic graphs”. In: *Uncertainty in Artificial Intelligence*. PMLR. 2020, pp. 1149–1158.

[Bha+25] A. Bhattacharyya, C. Daskalakis, T. Gouleakis, and Y. Wang. “Learning High-dimensional Gaussians from Censored Data”. In: *arXiv preprint arXiv:2504.19446* (2025).

[Bha+21] A. Bhattacharyya, S. Gayen, E. Price, and N. Vinodchandran. “Near-optimal learning of tree-structured distributions by Chow-Liu”. In: *Proceedings of the 53rd annual acm SIGACT symposium on theory of computing*. 2021, pp. 147–160.

[BY22] A. Bhattacharyya and Y. Yoshida. “Property Testing: Problems and Techniques”. Springer Nature, 2022.

[Bis93] C. M. Bishop. “Novelty Detection and Neural Network Validation”. In: *ICANN ’93*. Ed. by S. Gielen and B. Kappen. London: Springer London, 1993, pp. 789–794.

[Bla+22] G. Blanc, J. Lange, A. Malik, and L. Tan. “On the power of adaptivity in statistical adversaries”. In: *COLT*. Vol. 178. Proceedings of Machine Learning Research. PMLR, 2022, pp. 5030–5061.

[BM09] O. Boldea and J. R. Magnus. “Maximum likelihood estimation of the multivariate normal mixture model”. In: *Journal of the American Statistical Association* 104.488 (2009), pp. 1539–1549.

[Bol89] K. A. Bollen. “Measurement models: The relation between latent and observed variables”. In: *Structural equations with latent variables* (1989), pp. 179–225.

[Bor10] C. Borgelt. “A conditional independence algorithm for learning undirected graphical models”. In: *Journal of Computer and System Sciences* 76.1 (2010), pp. 21–33.

[BT14] T. Bouezmarni and A. Taamouti. “Nonparametric tests for conditional independence using conditional distributions”. In: *Journal of Nonparametric Statistics* 26.4 (2014), pp. 697–719.

[Bre+96] R. Breen et al. “Regression models: Censored, sample selected, or truncated data”. Vol. 111. Sage, 1996.

[Bre+00] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. “LOF: identifying density-based local outliers”. In: *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. 2000, pp. 93–104.

[BK96] J. M. Brick and G. Kalton. “Handling missing data in survey research”. In: *Statistical methods in medical research* 5.3 (1996), pp. 215–238.

[BEL18] S. Bubeck, R. Eldan, and J. Lehec. “Sampling from a log-concave distribution with projected langevin monte carlo”. In: *Discrete & Computational Geometry* 59.4 (2018), pp. 757–783.

[Bud+22] K. Budhathoki, L. Minorics, P. Blöbaum, and D. Janzing. "Causal structure-based root cause analysis of outliers". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 2357–2369.

[Bun71] P. Buneman. "The recovery of trees from measures of dissimilarity". In: *Mathematics in the archaeological and historical sciences* (1971).

[CT06a] E. J. Candes and T. Tao. "Near-optimal signal recovery from random projections: Universal encoding strategies?" In: *IEEE transactions on information theory* 52.12 (2006), pp. 5406–5425.

[Can20] C. L. Canonne. "A survey on distribution testing: Your data is big. But is it blue?" In: *Theory of Computing* (2020), pp. 1–100.

[Can+21] C. L. Canonne, X. Chen, G. Kamath, A. Levi, and E. Waingarten. "Random restrictions of high dimensional distributions and uniformity testing with subcube conditioning". In: *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM. 2021, pp. 321–336.

[Can+25] C. L. Canonne, T. Gouleakis, Y. Wang, and J. Q. Yang. "Gaussian Mean Testing under Truncation". In: *arXiv preprint arXiv:2504.04682* (2025).

[Can+23] C. L. Canonne, S. B. Hopkins, J. Li, A. Liu, and S. Narayanan. "The Full Landscape of Robust Mean Testing: Sharp Separations between Oblivious and Adaptive Contamination". In: *arXiv preprint arXiv:2307.10273* (2023).

[Can+18] C. L. Canonne, I. Diakonikolas, D. M. Kane, and A. Stewart. "Testing conditional independence of discrete distributions". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018*. ACM, 2018, pp. 735–748. URL: https://doi.org/10.1145/3188745.3188756.

[Car+21] G. Carreras, G. Miccinesi, A. Wilcock, N. Preston, D. Nieboer, L. Deliens, M. Groenvold, U. Lunder, A. van der Heide, and M. Baccini. "Missing not at random in end of life care studies: multiple imputation

and sensitivity analysis on data from the ACTION study". In: *BMC medical research methodology* 21.1 (2021), pp. 1–12.

[CCS13] J. Cha, B. R. Cho, and J. L. Sharp. "Rethinking the truncated normal distribution". In: *International Journal of Experimental Design and Process Optimisation* 3.4 (2013), pp. 327–363.

[Cha+13] S.-O. Chan, I. Diakonikolas, X. Sun, and R. A. Servedio. "Learning mixtures of structured distributions over discrete domains". In: *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*. SIAM. 2013, pp. 1380–1394.

[Cha96] J. T. Chang. "Full reconstruction of Markov models on evolutionary trees: identifiability and consistency". In: *Mathematical biosciences* 137.1 (1996), pp. 51–73.

[CH91] J. T. Chang and J. A. Hartigan. "Reconstruction of evolutionary trees from pairwise distributions on current species". In: *Computing science and statistics: Proceedings of the 23rd symposium on the interface*. Citeseer. 1991, pp. 254–257.

[CSV17] M. Charikar, J. Steinhardt, and G. Valiant. "Learning from untrusted data". In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. 2017, pp. 47–60.

[CDW19] W. Chen, M. Drton, and Y. S. Wang. "On causal discovery with an equal-variance assumption". In: *Biometrika* 106.4 (2019), pp. 973–980.

[Che+20] Y. Cheng, I. Diakonikolas, R. Ge, and M. Soltanolkotabi. "High-dimensional robust mean estimation via gradient descent". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 1768–1778.

[Che+22] Y. Cherapanamjeri, C. Daskalakis, A. Ilyas, and M. Zampetakis. "What Makes A Good Fisherman? Linear Regression under Self-Selection Bias". In: *arXiv preprint arXiv:2205.03246* (2022).

[CMY20] Y. Cherapanamjeri, S. Mohanty, and M. Yau. "List decodable mean estimation in nearly linear time". In: *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2020, pp. 141–148.

[Che+18] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. "Double/debiased machine learning for treatment and structural parameters". 2018.

[CNS18] V. Chernozhukov, W. Newey, and R. Singh. "De-biased machine learning of global and local parameters using regularized Riesz representers". In: *arXiv preprint arXiv:1802.08667* (2018).

[CNS21] V. Chernozhukov, W. K. Newey, and R. Singh. "A simple and general debiased machine learning theorem with finite sample guarantees". In: *arXiv preprint arXiv:2105.15197* (2021).

[Chi96] D. M. Chickering. "Learning Bayesian networks is NP-complete". In: *Learning from data*. Springer, 1996, pp. 121–130.

[Chi02a] D. M. Chickering. "Learning equivalence classes of Bayesian-network structures". In: *The Journal of Machine Learning Research* 2 (2002), pp. 445–498.

[Chi02b] D. M. Chickering. "Optimal structure identification with greedy search". In: *Journal of machine learning research* 3.Nov (2002), pp. 507–554.

[Chi02c] D. M. Chickering. "Optimal structure identification with greedy search". In: *Journal of machine learning research* 3.Nov (2002), pp. 507–554.

[Chi20] M. Chickering. "Statistically efficient greedy equivalence search". In: *Conference on Uncertainty in Artificial Intelligence*. PMLR. 2020, pp. 241–249.

[CHM04] M. Chickering, D. Heckerman, and C. Meek. "Large-sample learning of Bayesian networks is NP-hard". In: *Journal of Machine Learning Research* 5 (2004), pp. 1287–1330.

[Cho+11] M. J. Choi, V. Y. Tan, A. Anandkumar, and A. S. Willsky. "Learning latent tree graphical models". In: *Journal of Machine Learning Research* 12 (2011), pp. 1771–1812.

[Cho+23] D. Choo, J. Q. Yang, A. Bhattacharyya, and C. L. Canonne. "Learning bounded-degree polytrees with known skeleton". In: *arXiv preprint arXiv:2310.06333* (2023).

[CW73] C. Chow and T. Wagner. "Consistency of an estimate of tree-dependent probability distributions (corresp.)" In: *IEEE Transactions on Information Theory* 19.3 (1973), pp. 369–371.

[CL68] C. K. Chow and C. N. Liu. "Approximating discrete probability distributions with dependence trees". In: *IEEE Trans. Inf. Theory* 14.3 (1968), pp. 462–467. URL: https://doi.org/10.1109/TIT.1968.1054142.

[CFS16] B. Coecke, T. Fritz, and R. W. Spekkens. "A mathematical theory of resources". In: *Information and Computation* 250 (2016). Quantum Physics and Logic, pp. 59–86. ISSN: 0890-5401.

[Coh57] A. C. Cohen. "On the solution of estimating equations for truncated and censored samples from normal populations". In: *Biometrika* 44.1/2 (1957), pp. 225–236.

[Coh91a] A. C. Cohen. "Truncated and censored samples: theory and applications". CRC press, 1991.

[Coh91b] A. C. Cohen. "Truncated and censored samples: theory and applications". CRC press, 1991.

[CM14] D. Colombo and M. H. Maathuis. "Order-independent constraint-based causal structure learning." In: *J. Mach. Learn. Res.* 15.1 (2014), pp. 3741–3782.

[Col+11] D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson. "Learning high-dimensional DAGs with latent and selection variables". In: *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*. AUAI Press. 2011, pp. 850–850.

[CD17] P. Constantinou and A. P. Dawid. "Extended conditional independence and applications in causal inference". In: *The Annals of Statistics* (2017), pp. 2618–2653.

[CT06b] T. M. Cover and J. A. Thomas. "Elements of information theory". 2nd ed. Hoboken, NJ, USA: Wiley-Interscience, 2006.

[CW93] D. R. Cox and N. Wermuth. "Linear dependencies represented by chain graphs". In: *Statistical science* (1993), pp. 204–218.

[DVZ18] D. Dadush, L. A. Végh, and G. Zambelli. "Geometric rescaling algorithms for submodular function minimization". In: *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM. 2018, pp. 832–848.

[DSZ22] H. Dai, P. Spirtes, and K. Zhang. "Independence testing-based approach to causal discovery under measurement error and linear non-gaussian models". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 27524–27536.

[Das99a] S. Dasgupta. "Learning mixtures of Gaussians". In: *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*. IEEE. 1999, pp. 634–644.

[Das99b] S. Dasgupta. "Learning polytrees". In: *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. 1999, pp. 134–141.

[Das+18] C. Daskalakis, T. Gouleakis, C. Tzamos, and M. Zampetakis. "Efficient statistics, in high dimensions, from truncated samples". In: *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2018, pp. 639–649.

[Das+19] C. Daskalakis, T. Gouleakis, C. Tzamos, and M. Zampetakis. "Computationally and statistically efficient truncated regression". In: *Conference on Learning Theory*. PMLR. 2019, pp. 955–960.

[Das+21a] C. Daskalakis, V. Kontonis, C. Tzamos, and E. Zampetakis. "A statistical taylor theorem and extrapolation of truncated densities". In: *Conference on Learning Theory*. PMLR. 2021, pp. 1395–1398.

[DMR11] C. Daskalakis, E. Mossel, and S. Roch. "Evolutionary trees and the Ising model on the Bethe lattice: a proof of Steel's conjecture". In: *Probability Theory and Related Fields* 149.1-2 (2011), pp. 149–189.

[DP20] C. Daskalakis and Q. Pan. "Tree-structured Ising models can be learned efficiently". In: *arXiv preprint arXiv:2010.14864* (2020).

[DP21] C. Daskalakis and Q. Pan. "Sample-optimal and efficient learning of tree Ising models". In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. 2021, pp. 133–146.

[DRZ20] C. Daskalakis, D. Rohatgi, and E. Zampetakis. "Truncated linear regression in high dimensions". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 10338–10347.

[Das+21b] C. Daskalakis, P. Stefanou, R. Yao, and E. Zampetakis. "Efficient Truncated Linear Regression with Unknown Noise Variance". In: *Advances in Neural Information Processing Systems* 34 (2021).

[De+24] A. De, H. Li, S. Nadimpalli, and R. A. Servedio. "Detecting low-degree truncation". In: *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. 2024, pp. 1027–1038.

[DNS23] A. De, S. Nadimpalli, and R. A. Servedio. "Testing Convex Truncation". In: *SODA*. SIAM, 2023, pp. 4050–4082.

[DV55] W. L. Deemer Jr and D. F. Votaw Jr. "Estimation of parameters of truncated or censored exponential distributions". In: *The Annals of Mathematical Statistics* 26.3 (1955), pp. 498–504.

[DM01] M. A. Delgado and W. G. Manteiga. "Significance testing in nonparametric regression based on the bootstrap". In: *The Annals of Statistics* 29.5 (2001), pp. 1469–1507.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22.

[Dia+20] I. Diakonikolas, S. B. Hopkins, D. Kane, and S. Karmalkar. "Robustly learning any clusterable mixture of gaussians". In: *arXiv preprint arXiv:2005.06417* (2020).

[Dia+19] I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart. "Robust estimators in high-dimensions without the computational intractability". In: *SIAM Journal on Computing* 48.2 (2019), pp. 742–864.

[Dia+18] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. "Robustly learning a gaussian: Getting optimal error, efficiently". In: *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM. 2018, pp. 2683–2702.

[DK19] I. Diakonikolas and D. M. Kane. "Recent advances in algorithmic high-dimensional robust statistics". In: *arXiv preprint arXiv:1911.05911* (2019).

[DK23] I. Diakonikolas and D. M. Kane. "Algorithmic high-dimensional robust statistics". Cambridge university press, 2023.

[DKP22] I. Diakonikolas, D. M. Kane, and A. Pensia. "Gaussian Mean Testing Made Simple". In: *arXiv preprint arXiv:2210.13706* (2022).

[DKS17] I. Diakonikolas, D. M. Kane, and A. Stewart. "Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures". In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2017, pp. 73–84.

[DKS16] I. Diakonikolas, D. M. Kane, and A. Stewart. "Statistical Query Lower Bounds for Robust Estimation of High-dimensional Gaussians and Gaussian Mixtures". In: *CoRR* abs/1611.03473 (2016).

[Dix60] W. J. Dixon. "Simplified estimation from censored normal samples". In: *The Annals of Mathematical Statistics* (1960), pp. 385–391.

[Don06] D. L. Donoho. "Compressed sensing". In: *IEEE Transactions on information theory* 52.4 (2006), pp. 1289–1306.

[Don04] S. Donoho. "Early detection of insider trading in option markets". In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2004, pp. 420–429.

[Drt+09] M. Drton et al. "Discrete chain graph models". In: *Bernoulli* 15.3 (2009), pp. 736–753.

[DE06] M. Drton and M. Eichler. "Maximum likelihood estimation in Gaussian chain graph models under the alternative Markov property". In: *Scandinavian journal of statistics* 33.2 (2006), pp. 247–257.

[Drt+17] M. Drton, S. Lin, L. Weihs, and P. Zwiernik. "Marginal likelihood and model selection for Gaussian latent tree and forest models". In: (2017).

[DJW18] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. “Minimax optimal procedures for locally private estimation”. In: *Journal of the American Statistical Association* 113.521 (2018), pp. 182–201.

[Ebe17] F. Eberhardt. “Introduction to the foundations of causal discovery”. In: *International Journal of Data Science and Analytics* 3.2 (2017), pp. 81–91.

[EH25] A. Ebtekar and M. Hutter. “Foundations of algorithmic thermodynamics”. In: *Physical Review E* 111 (1 2025), p. 014118.

[EWJ25] A. Ebtekar, Y. Wang, and D. Janzing. “Toward Universal Laws of Outlier Propagation”. In: *The 41st Conference on Uncertainty in Artificial Intelligence*. 2025.

[Edw12] D. Edwards. “Introduction to graphical modelling”. Springer Science & Business Media, 2012.

[EB01] C. K. Enders and D. L. Bandalos. “The relative performance of full information maximum likelihood estimation for missing data in structural equation models”. In: *Structural equation modeling* 8.3 (2001), pp. 430–457.

[Eve+17] G. Even, O. Fischer, P. Fraigniaud, T. Gonen, R. Levi, M. Medina, P. Montealegre, D. Olivetti, R. Oshman, I. Rapaport, et al. “Three notes on distributed property testing”. In: *31st International Symposium on Distributed Computing (DISC 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 2017.

[Fis31] R. Fisher. “Properties and applications of Hh functions”. In: *Mathematical tables* 1 (1931), pp. 815–852.

[FKT20] D. Fotakis, A. Kalavasis, and C. Tzamos. “Efficient parameter estimation of truncated boolean product distributions”. In: *Conference on Learning Theory*. 2020.

[Fra05] M. P. Frank. “The indefinite logarithm, logarithmic units, and the nature of entropy”. In: (2005). arXiv: physics/0506128 [physics].

[Fre95] J. Freeman. “Outliers in statistical data”. In: *Journal of the Operational Research Society* 46.8 (1995), pp. 1034–1035.

[FNP13] N. Friedman, I. Nachman, and D. Pe'er. "Learning Bayesian network structure from massive datasets: The" sparse candidate" algorithm". In: *arXiv preprint arXiv:1301.6696* (2013).

[FY96] N. Friedman and Z. Yakhini. "On the Sample Complexity of Learning Bayesian Networks". In: *Uncertainty in Artifical Intelligence (UAI)*. Feb. 1996. eprint: 1302.3579. URL: https://arxiv.org/abs/1302.3579.

[Fry90] M. Frydenberg. "The chain graph Markov property". In: *Scandinavian Journal of Statistics* (1990), pp. 333–353.

[Gác94] P. Gács. "The Boltzmann entropy and randomness tests". In: *Proceedings Workshop on Physics and Computation. PhysComp'94*. IEEE. Dallas, TX, USA, Nov. 1994, pp. 209–216.

[Gác21] P. Gács. "Lecture notes on descriptional complexity and randomness". In: *arXiv preprint arXiv:2105.04704* (2021).

[Gal98] F. Galton. "An examination into the registered speeds of american trotting horses, with remarks on their value as hereditary data". In: *Proceedings of the Royal Society of London* 62.379-387 (1898), pp. 310–315.

[Gan+21] Y. Gan, M. Liang, S. Dev, D. Lo, and C. Delimitrou. "Sage: practical and scalable ML-driven performance debugging in microservices". In: *Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*. ASPLOS '21. Virtual, USA: Association for Computing Machinery, 2021, pp. 135–151. ISBN: 9781450383172.

[GA21] M. Gao and B. Aragam. "Efficient Bayesian network structure learning via local Markov boundary search". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 4301–4313.

[GDA20a] M. Gao, Y. Ding, and B. Aragam. "A polynomial-time algorithm for learning nonparametric causal graphs". In: *Advances in Neural Information Processing Systems* 33 (2020).

[GDA20b] M. Gao, Y. Ding, and B. Aragam. "A polynomial-time algorithm for learning nonparametric causal graphs". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 11599–11611.

[GTA22] M. Gao, W. M. Tai, and B. Aragam. "Optimal estimation of Gaussian DAG models". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 8738–8757.

[GTA23] M. Gao, W. M. Tai, and B. Aragam. "Optimal neighbourhood selection in structural equation models". In: *arXiv preprint arXiv:2306.02244* (2023).

[GVG15] S. Gao, G. Ver Steeg, and A. Galstyan. "Efficient estimation of mutual information for strongly dependent variables". In: *Artificial intelligence and statistics*. PMLR. 2015, pp. 277–286.

[Gas+15] S. Gaspers, M. Koivisto, M. Liedloff, S. Ordyniak, and S. Szeider. "On finding optimal polytrees". In: *Theoretical Computer Science* 592 (2015), pp. 49–58.

[Gau12] L. Gautier. "rpy2: A simple and efficient access to R from Python, 2012". In: *URL http://rpy. sourceforge. net/rpy2. html* (2012).

[GHK15] R. Ge, Q. Huang, and S. M. Kakade. "Learning mixtures of gaussians in high dimensions". In: *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. 2015, pp. 761–770.

[GS10] G. Geenens and L. Simar. "Nonparametric tests for conditional independence in two-way contingency tables". In: *Journal of Multivariate Analysis* 101.4 (2010), pp. 765–788.

[GP93] D. Geiger and J. Pearl. "Logical and algorithmic properties of conditional independence and graphical models". In: *The annals of statistics* 21.4 (1993), pp. 2001–2021.

[GG84] S. Geman and D. Geman. "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images". In: *IEEE Transactions on pattern analysis and machine intelligence* 6 (1984), pp. 721–741.

[GRW06] C. R. Genovese, K. Roeder, and L. Wasserman. "False discovery control with p-value weighting". In: *Biometrika* 93.3 (2006), pp. 509–524.

[GH17a] A. Ghoshal and J. Honorio. "Information-theoretic limits of Bayesian network structure learning". In: *Artificial Intelligence and Statistics*. PMLR. 2017, pp. 767–775.

[GH17b] A. Ghoshal and J. Honorio. "Learning identifiable Gaussian Bayesian networks in polynomial time and sample complexity". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, pp. 6460–6469.

[GH17c] A. Ghoshal and J. Honorio. "Learning identifiable gaussian bayesian networks in polynomial time and sample complexity". In: *Advances in Neural Information Processing Systems* 30 (2017).

[GH18a] A. Ghoshal and J. Honorio. "Learning linear structural equation models in polynomial time and sample complexity". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2018, pp. 1466–1475.

[GH18b] A. Ghoshal and J. Honorio. "Learning linear structural equation models in polynomial time and sample complexity". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2018, pp. 1466–1475.

[Gio+14] F. M. Giorgi, G. Lopez, J. H. Woo, B. Bisikirska, A. Califano, and M. Bansal. "Inferring protein modulation from gene expression data using conditional mutual information". In: *PloS one* 9.10 (2014), e109569.

[GZS19] C. Glymour, K. Zhang, and P. Spirtes. "Review of causal discovery methods based on graphical models". In: *Frontiers in genetics* 10 (2019), p. 524.

[GG21] Z. Goldfeld and K. Greenewald. "Sliced mutual information: A scalable measure of statistical dependence". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 17567–17578.

[Gol17] O. Goldreich. "Introduction to property testing". Cambridge University Press, 2017.

[GLS81] M. Grötschel, L. Lovász, and A. Schrijver. "The ellipsoid method and its consequences in combinatorial optimization". In: *Combinatorica* 1.2 (1981), pp. 169–197.

[GLS12] M. Grötschel, L. Lovász, and A. Schrijver. "Geometric algorithms and combinatorial optimization". Vol. 2. Springer Science & Business Media, 2012.

[GHK20] P. Grünwald, R. de Heide, and W. M. Koolen. "Safe testing". In: *2020 Information Theory and Applications Workshop (ITA)*. IEEE. 2020, pp. 1–54.

[GKM21] N. Grüttemeier, C. Komusiewicz, and N. Morawietz. "On the parameterized complexity of polytree learning". In: *arXiv preprint arXiv:2105.09675* (2021).

[GH02] H. Guo and W. Hsu. "A survey of algorithms for real-time Bayesian network inference". In: *Join workshop on real time decision support and diagnosis systems*. 2002, pp. 1–12.

[GK08] S. Gupta and H. W. Kim. "Linking structural equation modeling to Bayesian networks: Decision support for customer retention in virtual communities". In: *European Journal of Operational Research* 190.3 (2008), pp. 818–833.

[GR20] R. Gurjar and R. Rathi. "Linearly Representable Submodular Functions: An Algebraic Algorithm for Minimization". In: *47th International Colloquium on Automata, Languages, and Programming (ICALP 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik. 2020.

[HS90] C. N. Haas and P. A. Scheff. "Estimation of averages in truncated samples". In: *Environmental science & technology* 24.6 (1990), pp. 912–919.

[HSS08] A. Hagberg, P. Swart, and D. S Chult. "Exploring network structure, dynamics, and function using NetworkX". Tech. rep. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[HM98] V. A. Hajivassiliou and D. L. McFadden. "The method of simulated scores for the estimation of LDV models". In: *Econometrica* (1998), pp. 863–896.

[HK07] S. Halevy and E. Kushilevitz. "Distribution-free property-testing". In: *SIAM Journal on Computing* 37.4 (2007), pp. 1107–1138.

[Har+23] M. Hardt, W. Orchard, P. Blöbaum, S. Kasiviswanathan, and E. Kirschbaum. "The PetShop Dataset – Finding Causes of Performance Issues across Microservices". 2023. arXiv: 2311.04806 [cs.DC].

[HD13] N. Harris and M. Drton. "PC algorithm for nonparanormal graphical models." In: *Journal of Machine Learning Research* 14.11 (2013).

[HW77] J. A. Hausman and D. A. Wise. "Social experimentation, truncated distributions, and efficient estimation". In: *Econometrica: Journal of the Econometric Society* (1977), pp. 919–938.

[Hec97] D. Heckerman. "Bayesian networks for data mining". In: *Data mining and knowledge discovery* 1 (1997), pp. 79–119.

[Hec79] J. J. Heckman. "Sample selection bias as a specification error". In: *Econometrica: Journal of the econometric society* (1979), pp. 153–161.

[HK10] J. Honaker and G. King. "What to do about missing values in time-series cross-section data". In: *American journal of political science* 54.2 (2010), pp. 561–581.

[Hoo+16] B. Hooi, H. A. Song, A. Beutel, N. Shah, K. Shin, and C. Faloutsos. "Fraudar: Bounding graph fraud in the face of camouflage". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 895–904.

[HLZ20] S. Hopkins, J. Li, and F. Zhang. "Robust and heavy-tailed mean estimation made simple, via regret minimization". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 11902–11912.

[HL19] S. B. Hopkins and J. Li. "How Hard is Robust Mean Estimation?" In: *COLT*. Vol. 99. Proceedings of Machine Learning Research. PMLR, 2019, pp. 1649–1682.

[Hoy+08a] P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. "Nonlinear causal discovery with additive noise models". In: *Advances in neural information processing systems* 21 (2008), pp. 689–696.

[Hoy+08b] P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. "Nonlinear causal discovery with additive noise models". In: *Advances in neural information processing systems* 21 (2008).

[HLH12] F.-M. Hsu, Y.-T. Lin, and T.-K. Ho. "Design and implementation of an intelligent recommendation system for tourist attractions: The integration of EBM model, Bayesian network and Google Maps". In: *Expert Systems with Applications* 39.3 (2012), pp. 3257–3264.

[HR21] L. Hu and O. Reingold. "Robust Mean Estimation on Highly Incomplete Data with Arbitrary Outliers". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 1558–1566.

[Hua10] T.-M. Huang. "Testing conditional independence using maximal nonlinear conditional correlation". In: (2010).

[Hub92] P. J. Huber. "Robust estimation of a location parameter". In: *Breakthroughs in statistics*. Springer, 1992, pp. 492–518.

[HR11] P. J. Huber and E. M. Ronchetti. "Robust statistics". John Wiley & Sons, 2011.

[Ikr+22] A. Ikram, S. Chakraborty, S. Mitra, S. Saini, S. Bagchi, and M. Kocaoglu. "Root cause analysis of failures in microservices through causal discovery". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 31158–31170.

[IFF01] S. Iwata, L. Fleischer, and S. Fujishige. "A combinatorial strongly polynomial algorithm for minimizing submodular functions". In: *Journal of the ACM (JACM)* 48.4 (2001), pp. 761–777.

[Jak+22] M. E. Jakobsen, R. D. Shah, P. Bühlmann, and J. Peters. "Structure learning for directed trees". In: *The Journal of Machine Learning Research* 23.1 (2022), pp. 7076–7172.

[JGK23] F. Jamshidi, L. Ganassali, and N. Kiyavash. "On sample complexity of conditional independence testing with Von Mises estimator with application to causal discovery". In: *arXiv preprint arXiv:2310.13553* (2023).

[JS10] D. Janzing and B. Schölkopf. "Causal inference using the algorithmic Markov condition". In: *IEEE Transactions on Information Theory* 56.10 (2010), pp. 5168–5194.

[JV18] M. A. Javidian and M. Valtorta. "On the properties of MVR chain graphs". In: *arXiv preprint arXiv:1803.04262* (2018).

[JVJ20] M. A. Javidian, M. Valtorta, and P. Jamshidi. "AMP Chain Graphs: Minimal Separators and Structure Learning Algorithms". In: *Journal of Artificial Intelligence Research* 69 (2020), pp. 419–470.

[Jia21] H. Jiang. "Minimizing convex functions with integral minimizers". In: *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM. 2021, pp. 976–985.

[KB07] M. Kalisch and P. Bühlman. "Estimating high-dimensional directed acyclic graphs with the PC-algorithm." In: *Journal of Machine Learning Research* 8.3 (2007).

[Kal+12] M. Kalisch, M. Mächler, D. Colombo, M. H. Maathuis, P. Bühlmann, et al. "Causal inference using graphical models with the R package pcalg". In: *Journal of Statistical Software* 47.11 (2012), pp. 1–26.

[Kan21] D. M. Kane. "Robust learning of mixtures of gaussians". In: *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM. 2021, pp. 1246–1258.

[KP77] M. Kanter and H. Proppe. "Reduction of variance for Gaussian densities via restriction to convex sets". In: *Journal of Multivariate Analysis* 7.1 (1977), pp. 74–81.

[KS01] D. Karger and N. Srebro. "Learning Markov networks: Maximum bounded tree-width graphs". In: *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics. 2001, pp. 392–401.

[Kea+94] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. E. Schapire, and L. Sellie. "On the learnability of discrete distributions". In: *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*. 1994, pp. 273–282.

[Kho+19] P. Khosravi, Y. Liang, Y. Choi, and G. V. d. Broeck. "What to expect of classifiers? reasoning about logistic regression with missing features". In: *arXiv preprint arXiv:1903.01620* (2019).

[KF09a] D. Koller and N. Friedman. "Probabilistic graphical models: principles and techniques". MIT press, 2009.

[KF09b] D. Koller and N. Friedman. "Probabilistic graphical models: principles and techniques". MIT press, 2009.

[KTZ19] V. Kontonis, C. Tzamos, and M. Zampetakis. "Efficient truncated statistics with unknown truncation". In: *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2019, pp. 1578–1595.

[KSG04] A. Kraskov, H. Stögbauer, and P. Grassberger. "Estimating mutual information". In: *Physical review E* 69.6 (2004), p. 066138.

[Kra10] A. Krause. "SFO: A toolbox for submodular function optimization". In: *Journal of Machine Learning Research* 11 (2010), pp. 1141–1144.

[KSG08] A. Krause, A. Singh, and C. Guestrin. "Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies." In: *Journal of Machine Learning Research* 9.2 (2008).

[KFL01] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. "Factor graphs and the sum-product algorithm". In: *IEEE Transactions on information theory* 47.2 (2001), pp. 498–519.

[Lac+20] S. Lachapelle, P. Brouillard, T. Deleu, and S. Lacoste-Julien. "Gradient-Based Neural DAG Learning". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[LRV16] K. A. Lai, A. B. Rao, and S. Vempala. “Agnostic estimation of mean and covariance”. In: *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2016, pp. 665–674.

[LAR22] W.-Y. Lam, B. Andrews, and J. Ramsey. “Greedy relaxations of the sparsest permutation algorithm”. In: *Uncertainty in Artificial Intelligence*. PMLR. 2022, pp. 1052–1062.

[Lau96] S. L. Lauritzen. “Graphical models”. Vol. 17. Clarendon Press, 1996.

[LW89] S. L. Lauritzen and N. Wermuth. “Graphical models for associations between variables, some of which are qualitative and some quantitative”. In: *The annals of Statistics* (1989), pp. 31–57.

[Lee14] A. Lee. “Table of the Gaussian" tail" functions; when the" tail" is larger than the body”. In: *Biometrika* 10.2/3 (1914), pp. 208–214.

[LSW15] Y. T. Lee, A. Sidford, and S. C.-w. Wong. “A faster cutting plane method and its implications for combinatorial and convex optimization”. In: *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*. IEEE. 2015, pp. 1049–1065.

[Lei+20] Z. Lei, K. Luh, P. Venkat, and F. Zhang. “A fast spectral algorithm for mean estimation with sub-gaussian rates”. In: *Conference on Learning Theory*. PMLR. 2020, pp. 2598–2612.

[Lev84] L. A. Levin. “Randomness conservation inequalities; information and independence in mathematical theories”. In: *Information and Control* 61.1 (1984), pp. 15–37.

[Lev76] L. A. Levin. “Uniform tests of randomness”. In: *Doklady Akademii Nauk*. Vol. 227. 1. Russian Academy of Sciences. 1976, pp. 33–35.

[LPM01] M. Levitz, M. D. Perlman, and D. Madigan. “Separation and completeness properties for AMP chain graph Markov models”. In: *Annals of statistics* (2001), pp. 1751–1784.

[Li+24] J. Li, B. B. Chu, I. F. Scheller, J. Gagneur, and M. H. Maathuis. “Root cause discovery via permutations and Cholesky decomposition”. arxiv:2410.12151. 2024.

[LV97] M. Li and P. Vitányi. “An Introduction to Kolmogorov Complexity and its Applications”. New York: Springer, 1997.

[Li+22] M. Li, Z. Li, K. Yin, X. Nie, W. Zhang, K. Sui, and D. Pei. “Causal inference-based root cause analysis for online service systems with intervention recognition”. In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2022, pp. 3230–3240.

[LW08] K.-C. Liang and X. Wang. “Gene regulatory network reconstruction using conditional mutual information”. In: *EURASIP Journal on Bioinformatics and Systems Biology* 2008 (2008), pp. 1–14.

[Lin+14] S. Lin, C. Uhler, B. Sturmfels, and P. Bühlmann. “Hypersurfaces and their singularities in partial correlation testing”. In: *Foundations of Computational Mathematics* 14 (2014), pp. 1079–1116.

[LG96] O. B. Linton and P. Gozalo. “Conditional independence restrictions: testing and estimation”. In: (1996).

[Lit+12] R. J. Little, R. D’Agostino, M. L. Cohen, K. Dickersin, S. S. Emerson, J. T. Farrar, C. Frangakis, J. W. Hogan, G. Molenberghs, S. A. Murphy, et al. “The prevention and treatment of missing data in clinical trials”. In: *New England Journal of Medicine* 367.14 (2012), pp. 1355–1360.

[LR19] R. J. Little and D. B. Rubin. “Statistical analysis with missing data”. Vol. 793. John Wiley & Sons, 2019.

[Liu+11] H. Liu, M. Xu, H. Gu, A. Gupta, J. Lafferty, and L. Wasserman. “Forest density estimation”. In: *The Journal of Machine Learning Research* 12 (2011), pp. 907–951.

[Liu+21] Z. Liu, J. H. Park, T. Rekatsinas, and C. Tzamos. “On Robust Mean Estimation under Coordinate-level Corruption”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 6914–6924.

[LB14] P.-L. Loh and P. Bühlmann. “High-dimensional learning of linear causal networks via inverse covariance estimation”. In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 3065–3105.

[Ma+20] M. Ma, J. Xu, Y. Wang, P. Chen, Z. Zhang, and P. Wang. "Automap: Diagnose your microservice-based web applications automatically". In: *Proceedings of The Web Conference 2020*. 2020, pp. 246–258.

[MXG08] Z. Ma, X. Xie, and Z. Geng. "Structural learning of chain graphs via decomposition". In: *Journal of Machine Learning Research* 9 (2008), p. 2847.

[Maa+18] M. Maathuis, M. Drton, S. Lauritzen, and M. Wainwright. "Handbook of graphical models". CRC Press, 2018.

[Mad86] G. S. Maddala. "Limited-dependent and qualitative variables in econometrics". 3. Cambridge university press, 1986.

[MST21] D. Malinsky, I. Shpitser, and E. J. Tchetgen Tchetgen. "Semiparametric inference for nonmonotone missing-not-at-random data: the no self-censoring model". In: *Journal of the American Statistical Association* (2021), pp. 1–9.

[MH59] N. Mantel and W. Haenszel. "Statistical aspects of the analysis of data from retrospective studies of disease". In: *Journal of the national cancer institute* 22.4 (1959), pp. 719–748.

[Mar+06] G. M. Marchetti et al. "Independencies induced from a graphical Markov model after marginalization and conditioning: the R package ggm". In: *Journal of Statistical Software* 15.6 (2006), pp. 1–15.

[MM92] M. Marcus and H. Minc. "A survey of matrix theory and matrix inequalities". Vol. 14. Courier Corporation, 1992.

[MS07] F. Markowetz and R. Spang. "Inferring cellular networks–a review". In: *BMC bioinformatics* 8.6 (2007), pp. 1–17.

[Mar66] P. Martin-Löf. "The definition of random sequences". In: *Information and control* 9.6 (1966), pp. 602–619.

[MGM21] A. Marx, A. Gretton, and J. M. Mooij. "A weaker faithfulness assumption based on triple interactions". In: *Uncertainty in Artificial Intelligence*. PMLR. 2021, pp. 451–460.

[Mas+16] D. M. Maslove, J. A. Dubin, A. Shrivats, and J. Lee. "Errors, omissions, and outliers in hourly vital signs measurements in intensive care". In: *Critical care medicine* 44.11 (2016), e1021–e1030.

[Mee95] C. Meek. "Complete orientation rules for patterns". Carnegie Mellon [Department of Philosophy], 1995.

[Mer97] R. Merris. "Multilinear algebra". Crc Press, 1997.

[MVL20] S. Misra, M. Vuffray, and A. Y. Lokhov. "Information theoretic optimal learning of gaussian graphical models". In: *Conference on Learning Theory*. PMLR. 2020, pp. 2888–2909.

[MPT13] K. Mohan, J. Pearl, and J. Tian. "Graphical models for inference with missing data". In: *Advances in neural information processing systems* 26 (2013).

[MN20] W. A. Mohotti and R. Nayak. "Efficient outlier detection in text corpus using rare frequency and ranking". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 14.6 (2020), pp. 1–30.

[Moo+16] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. "Distinguishing cause from effect using observational data: methods and benchmarks". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 1103–1204.

[Mor+09] A. Moradi, N. Mousavi, C. Paar, and M. Salmasizadeh. "A comparative study of mutual information analysis under a Gaussian assumption". In: *Information Security Applications: 10th International Workshop, WISA 2009, Busan, Korea, August 25-27, 2009, Revised Selected Papers 10*. Springer. 2009, pp. 193–205.

[MRS13] E. Mossel, S. Roch, and A. Sly. "Robust estimation of latent tree graphical models: Inferring hidden states with inexact parameters". In: *IEEE transactions on information theory* 59.7 (2013), pp. 4357–4373.

[Mou+13] R. Mourad, C. Sinoquet, N. L. Zhang, T. Liu, and P. Leray. "A survey on latent tree models and applications". In: *Journal of Artificial Intelligence Research* 47 (2013), pp. 157–203.

[Mur12] K. P. Murphy. “Machine learning: a probabilistic perspective”. MIT press, 2012.

[NBS20] R. Nabi, R. Bhattacharya, and I. Shpitser. “Full law identification in graphical models of missing data: Completeness results”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 7153–7163.

[NHM+18] P. Nandy, A. Hauser, M. H. Maathuis, et al. “High-dimensional consistency in score-based and hybrid structure learning”. In: *The Annals of Statistics* 46.6A (2018), pp. 3151–3183.

[NP33] J. Neyman and E. S. Pearson. “IX. On the problem of the most efficient tests of statistical hypotheses”. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231.694-706 (1933), pp. 289–337.

[Nie+14] S. Nie, D. D. Mauá, C. P. De Campos, and Q. Ji. “Advances in learning Bayesian networks of bounded treewidth”. In: *Advances in Neural Information Processing Systems*. 2014, pp. 2285–2293.

[NC03] C. C. Noble and D. J. Cook. “Graph-based anomaly detection”. In: *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’03. Washington, D.C.: Association for Computing Machinery, 2003, pp. 631–636. ISBN: 1581137370.

[Oka+24] N. Okati, S. H. G. Mejia, W. R. Orchard, P. Blöbaum, and D. Janzing. “Root Cause Analysis of Outliers with Missing Structural Knowledge”. arxiv:2406.05014. 2024. URL: https://arxiv.org/abs/2406.05014.

[Oue81] D. V. Ouellette. “Schur complements and statistics”. In: *Linear Algebra and its Applications* 36 (1981), pp. 187–295.

[PCR93] P. Spirtes, C. Glymour, and R. Scheines. “Causation, Prediction, and Search”. New York, NY: Springer-Verlag, 1993.

[Pan+22] E. Panjei, L. Gruenwald, E. Leal, C. Nguyen, and S. Silvia. “A survey on outlier explanations”. In: *The VLDB Journal* 31.5 (2022), pp. 977–1008.

[PSP11] A. Parikh, L. Song, and E. P Xing. “A spectral algorithm for latent tree graphical models”. In: (2011).

[Par20a] G. Park. “Identifiability of additive noise models using conditional variances”. In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 2896–2929.

[Par20b] G. Park. “Identifiability of Additive Noise Models Using Conditional Variances.” In: *Journal of Machine Learning Research* 21.75 (2020), pp. 1–34.

[PK20] G. Park and Y. Kim. “Identifiability of Gaussian linear structural equation models with homogeneous and heterogeneous error variances”. In: *Journal of the Korean Statistical Society* 49.1 (2020), pp. 276–292.

[Pas+13] F. Pascal, L. Bombrun, J.-Y. Tourneret, and Y. Berthoumieu. “Parameter estimation for multivariate generalized Gaussian distributions”. In: *IEEE Transactions on Signal Processing* 61.23 (2013), pp. 5960–5971.

[Pea88] J. Pearl. “Probabilistic reasoning in intelligent systems: networks of plausible inference”. Morgan kaufmann, 1988.

[Pea95] J. Pearl. “Causal diagrams for empirical research”. In: *Biometrika* 82.4 (1995), pp. 669–688.

[Pea+00] J. Pearl et al. “Models, reasoning and inference”. In: *Cambridge, UK: CambridgeUniversityPress* 19.2 (2000), p. 3.

[Pea09] J. Pearl. “Causality”. 2nd ed. Cambridge, UK: Cambridge university press, 2009.

[Pea10] J. Pearl. “An introduction to causal inference”. In: *The international journal of biostatistics* 6.2 (2010).

[Pea02] K. Pearson. “On the systematic fitting of curves to observations and measurements”. In: *Biometrika* 1.3 (1902), pp. 265–303.

[Pea20] K. Pearson. “Notes on the history of correlation”. In: *Biometrika* 13.1 (1920), pp. 25–45.

[PL08] K. Pearson and A. Lee. “On the generalised probable error in multiple normal correlation”. In: *Biometrika* 6.1 (1908), pp. 59–68.

[Peñ12] J. M. Peña. "Learning AMP chain graphs under faithfulness". In: *arXiv preprint arXiv:1204.5357* (2012).

[Peñ14a] J. M. Peña. "Learning marginal AMP chain graphs under faithfulness". In: *European Workshop on Probabilistic Graphical Models*. Springer. 2014, pp. 382–395.

[Peñ14b] J. M. Peña. "Marginal AMP chain graphs". In: *International Journal of Approximate Reasoning* 55.5 (2014), pp. 1185–1206.

[Peñ15] J. M. Peña. "Every LWF and AMP chain graph originates from a set of causal models". In: *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*. Springer. 2015, pp. 325–334.

[Peñ16] J. M. Peña. "Alternative Markov and Causal Properties for Acyclic Directed Mixed Graphs". In: *The 32nd Conference on Uncertainty in Artificial Intelligence (UAI 2016), New York City, NY, USA, June 25-29, 2016*. 2016.

[Peñ17a] J. M. Peña. "Identification of strong edges in AMP chain graphs". In: *arXiv preprint arXiv:1711.09990* (2017).

[Peñ17b] J. M. Peña. "Learning Causal AMP Chain Graphs". In: *Advanced Methodologies for Bayesian Networks*. PMLR. 2017, pp. 33–44.

[Peñ18] J. M. Peña. "Reasoning with alternative acyclic directed mixed graphs". In: *Behaviormetrika* 45.2 (2018), pp. 389–422.

[PG16] J. M. Peña and M. Gomez-Olmedo. "Learning marginal AMP chain graphs under faithfulness revisited". In: *International Journal of Approximate Reasoning* 68 (2016), pp. 108–126.

[PB14a] J. Peters and P. Bühlmann. "Identifiability of Gaussian structural equation models with equal error variances". In: *Biometrika* 101.1 (2014), pp. 219–228.

[PB14b] J. Peters and P. Bühlmann. "Identifiability of Gaussian structural equation models with equal error variances". In: *Biometrika* 101.1 (2014), pp. 219–228.

[PJS17] J. Peters, D. Janzing, and B. Schölkopf. "Elements of causal inference: foundations and learning algorithms". The MIT Press, 2017.

[Pet+14] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. "Causal discovery with continuous additive noise models". In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 2009–2053.

[Ple21] O. Plevrakis. "Learning from Censored and Dependent Data: The case of Linear Dynamics". In: *Conference on Learning Theory*. PMLR. 2021, pp. 3771–3787.

[PS11] B. Póczos and J. Schneider. "On the Estimation of *alpha*-Divergences". In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings. 2011, pp. 609–617.

[Qui22] V. Quintas-Martinez. "Finite-Sample Guarantees for High-Dimensional DML". In: *arXiv preprint arXiv:2206.07386* (2022).

[RH20] J. S. Racine and T. Hayfield. "Package np". In: (2020).

[Ram+23] A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer. "Game-theoretic statistics and safe anytime-valid inference". In: *Statistical Science* 38.4 (2023), pp. 576–601.

[RW24] A. Ramdas and R. Wang. "Hypothesis Testing with E-values". 2024. arXiv: 2410.23614 [math.ST].

[RS01] M. Ramoni and P. Sebastiani. "Robust learning with missing data". In: *Machine Learning* 45 (2001), pp. 147–170.

[Ram+17] J. Ramsey, M. Glymour, R. Sanchez-Romero, and C. Glymour. "A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images". In: *International journal of data science and analytics* 3.2 (2017), pp. 121–129.

[RU18] G. Raskutti and C. Uhler. "Learning directed acyclic graph models based on sparsest permutations". In: *Stat* 7.1 (2018), e183.

[RH11] S. Rathmanner and M. Hutter. “A Philosophical Treatise of Universal Induction”. In: *Entropy* 13.6 (2011), pp. 1076–1136.

[REB87] G. REBANE. “The recovery of causal poly-trees from statistical data”. In: *Uncertainty in Artificial Intelligence'87* (1987), pp. 222–228.

[RP13] G. Rebane and J. Pearl. “The recovery of causal poly-trees from statistical data”. In: *arXiv preprint arXiv:1304.2736* (2013).

[Rek+17] T. Rekatsinas, X. Chu, I. F. Ilyas, and C. Ré. “Holoclean: Holistic data repairs with probabilistic inference”. In: *arXiv preprint arXiv:1702.00820* (2017).

[Ric13] T. S. Richardson. “A discovery algorithm for directed cyclic graphs”. In: *arXiv preprint arXiv:1302.3599* (2013).

[RG97] J. M. Robins and R. D. Gill. “Non-response models for the analysis of non-monotone ignorable missing data”. In: *Statistics in medicine* 16.1 (1997), pp. 39–56.

[RRS00] J. M. Robins, A. Rotnitzky, and D. O. Scharfstein. “Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models”. In: *IMA VOLUMES IN MATHEMATICS AND ITS APPLICATIONS* 116 (2000), pp. 1–94.

[RW96] D. M. Rocke and D. L. Woodruff. “Identification of outliers in multivariate data”. In: *Journal of the American Statistical Association* 91.435 (1996), pp. 1047–1061.

[REB+18] D. Rothenhäusler, J. Ernest, P. Bühlmann, et al. “Causal inference in partially linear structural equation models”. In: *The Annals of Statistics* 46.6A (2018), pp. 2904–2938.

[RR97] A. Rotnitzky and J. Robins. “Analysis of semi-parametric regression models with non-ignorable non-response”. In: *Statistics in medicine* 16.1 (1997), pp. 81–102.

[RR95] A. Rotnitzky and J. M. Robins. “Semiparametric regression estimation in the presence of dependent censoring”. In: *Biometrika* 82.4 (1995), pp. 805–820.

[RRS98] A. Rotnitzky, J. M. Robins, and D. O. Scharfstein. "Semiparametric regression for repeated outcomes with nonignorable nonresponse". In: *Journal of the american statistical association* 93.444 (1998), pp. 1321–1339.

[RL03] P. J. Rousseeuw and A. M. Leroy. "Robust regression and outlier detection". John wiley & sons, 2003.

[Rov05] A. Roverato. "A unified approach to the characterization of equivalence classes of DAGs, chain graphs with no flags and chain graphs". In: *Scandinavian Journal of Statistics* 32.2 (2005), pp. 295–312.

[Roy51] A. D. Roy. "Some thoughts on the distribution of earnings". In: *Oxford economic papers* 3.2 (1951), pp. 135–146.

[Rub76] D. B. Rubin. "Inference and missing data". In: *Biometrika* 63.3 (1976), pp. 581–592.

[Rub12] R. Rubinfeld. "Taming big probability distributions". In: *XRDS: Crossroads, The ACM Magazine for Students* 19.1 (2012), pp. 24–28.

[SMS13] J. Safaei, J. Mauch, and L. Stacho. "Learning polytrees with constant number of roots from data". In: *AI 2013: Advances in Artificial Intelligence: 26th Australasian Joint Conference, Dunedin, New Zealand, December 1-6, 2013. Proceedings 26*. Springer. 2013, pp. 447–452.

[SK01] A. Sanjeev and R. Kannan. "Learning mixtures of arbitrary gaussians". In: *Proceedings of the thirty-third annual ACM symposium on Theory of computing*. 2001, pp. 247–257.

[SW12] N. P. Santhanam and M. J. Wainwright. "Information-theoretic limits of selecting binary graphical models in high dimensions". In: *IEEE Transactions on Information Theory* 58.7 (2012), pp. 4117–4134.

[SS05] J. Schafer and K. Strimmer. "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics". In: *Statistical applications in genetics and molecular biology* 4.1 (2005).

[SRR99] D. O. Scharfstein, A. Rotnitzky, and J. M. Robins. "Adjusting for nonignorable drop-out using semiparametric nonresponse models". In: *Journal of the American Statistical Association* 94.448 (1999), pp. 1096–1120.

[ST03] M. Schlather and J. A. Tawn. "A dependence measure for multivariate and spatial extreme values: Properties and inference". In: *Biometrika* 90.1 (2003), pp. 139–156.

[Sch+21] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. "Toward Causal Representation Learning". In: *Proceedings of the IEEE* 109.5 (2021), pp. 612–634.

[Sch00] A. Schrijver. "A combinatorial algorithm minimizing submodular functions in strongly polynomial time". In: *Journal of Combinatorial Theory, Series B* 80.2 (2000), pp. 346–355.

[Sch18] I. Schur. "Über endliche Gruppen und hermitesche Formen". In: *Mathematische Zeitschrift* 1.2 (1918), pp. 184–207.

[SW81] B. Schweizer and E. F. Wolff. "On nonparametric measures of dependence for random variables". In: *The annals of statistics* 9.4 (1981), pp. 879–885.

[Scu09] M. Scutari. "Learning Bayesian networks with the bnlearn R package". In: *arXiv preprint arXiv:0908.3817* (2009).

[Scu+14] M. Scutari, P. Howell, D. J. Balding, and I. Mackay. "Multiple quantitative trait analysis using Bayesian networks". In: *Genetics* 198.1 (2014), pp. 129–137.

[SW13] S. R. Seaman and I. R. White. "Review of inverse probability weighting for dealing with missing data". In: *Statistical methods in medical research* 22.3 (2013), pp. 278–295.

[Sen+17] R. Sen, A. T. Suresh, K. Shanmugam, A. G. Dimakis, and S. Shakkottai. "Model-powered conditional independence test". In: *Advances in neural information processing systems* 30 (2017).

[SP20] R. D. Shah and J. Peters. "The hardness of conditional independence testing and the generalised covariance measure". In: (2020).

[She20] A. Shen. "Randomness tests: theory and practice". In: *Fields of Logic and Computation III: Essays Dedicated to Yuri Gurevich on the Occasion of His 80th Birthday*. Springer, 2020, pp. 258–290.

[Shi+21] C. Shi, T. Xu, W. Bergsma, and L. Li. "Double generative adversarial networks for conditional independence testing". In: *Journal of Machine Learning Research* 22.285 (2021), pp. 1–32.

[Shi14] S. Shimizu. "LiNGAM: Non-Gaussian methods for estimating causal structures". In: *Behaviormetrika* 41 (2014), pp. 65–98.

[Shi+06a] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. "A linear non-Gaussian acyclic model for causal discovery". In: *Journal of Machine Learning Research* 7.Oct (2006), pp. 2003–2030.

[Shi+06b] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan. "A linear non-Gaussian acyclic model for causal discovery." In: *Journal of Machine Learning Research* 7.10 (2006).

[SMP15] I. Shpitser, K. Mohan, and J. Pearl. "Missing data as a causal and probabilistic problem". Tech. rep. CALIFORNIA UNIV LOS ANGELES DEPT OF COMPUTER SCIENCE, 2015.

[STA17] I. Shpitser, E. T. Tchetgen, and R. Andrews. "Modeling interference via symmetric treatment decomposition". In: *arXiv preprint arXiv:1709.01050* (2017).

[Son09] K. Song. "Testing conditional independence via Rosenblatt transforms". In: (2009).

[Son+14] L. Song, H. Liu, A. Parikh, and E. Xing. "Nonparametric latent tree graphical models: Inference, estimation, and structure learning". In: *arXiv preprint arXiv:1401.3940* (2014).

[SXP11] L. Song, E. P. Xing, and A. P. Parikh. "Kernel embeddings of latent tree graphical models". In: *Advances in Neural Information Processing Systems*. 2011, pp. 2708–2716.

[SP15] D. Sonntag and J. M. Peña. "Chain graphs and gene networks". In: *Foundations of Biomedical Knowledge Representation*. Springer, 2015, pp. 159–178.

[Spi10] P. Spirtes. “Introduction to causal inference.” In: *Journal of Machine Learning Research* 11.5 (2010).

[SG91] P. Spirtes and C. Glymour. “An algorithm for fast recovery of sparse causal graphs”. In: *Social science computer review* 9.1 (1991), pp. 62–72.

[SGS01] P. Spirtes, C. Glymour, and R. Scheines. “Causation, prediction, and search”. MIT press, 2001.

[SGS00] P. Spirtes, C. N. Glymour, and R. Scheines. “Causation, prediction, and search”. MIT press, 2000.

[Spi+00] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. “Causation, prediction, and search”. MIT press, 2000.

[SM95] P. Spirtes and C. Meek. “Learning Bayesian networks with discrete variables from data.” In: *KDD*. Vol. 1. 1995, pp. 294–299.

[SU22] C. Squires and C. Uhler. “Causal structure learning: A combinatorial perspective”. In: *Foundations of Computational Mathematics* (2022), pp. 1–35.

[Sre03] N. Srebro. “Maximum likelihood bounded tree-width Markov networks”. In: *Artificial intelligence* 143.1 (2003), pp. 123–138.

[Ste81] C. M. Stein. “Estimation of the mean of a multivariate normal distribution”. In: *The annals of Statistics* (1981), pp. 1135–1151.

[SJS10] B. Steudel, D. Janzing, and B. Schölkopf. “Causal Markov condition for submodular information measures”. In: *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)* (2010), pp. 464–476.

[Ste+02] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig. “The mutual information: detecting and evaluating dependencies between variables”. In: *Bioinformatics* 18.suppl_2 (2002), S231–S240.

[Sti89] S. M. Stigler. “Francis Galton’s account of the invention of correlation”. In: *Statistical Science* (1989), pp. 73–79.

[SR09] M. Studen, A. Roverato, and. tpánová. "Two operations of merging and splitting components in a chain graph". In: *Kybernetika* 45.2 (2009), pp. 208–248.

[SW07] L. Su and H. White. "A consistent characteristic function-based test for conditional independence". In: *Journal of Econometrics* 141.2 (2007), pp. 807–834.

[SW08] L. Su and H. White. "A nonparametric Hellinger metric test for conditional independence". In: *Econometric Theory* 24.4 (2008), pp. 829–864.

[STB17] G. A. Susto, M. Terzi, and A. Beghi. "Anomaly detection approaches for semiconductor manufacturing". In: *Procedia Manufacturing* 11 (2017), pp. 2018–2024.

[SVZ23] B. Szabó, L. Vuursteen, and H. van Zanten. "Optimal high-dimensional and nonparametric distributed testing under communication constraints". In: *Ann. Statist.* 51.3 (2023), pp. 909–934. ISSN: 0090-5364. URL: https://doi.org/10.1214/23-aos2269.

[Sza80] T. H. Szatrowski. "Necessary and sufficient conditions for explicit solutions in the multivariate normal estimation problem for patterned means and covariances". In: *The Annals of Statistics* (1980), pp. 802–810.

[TAW10] V. Y. Tan, A. Anandkumar, and A. S. Willsky. "Learning Gaussian tree models: Analysis of error exponents and extremal structures". In: *IEEE Transactions on Signal Processing* 58.5 (2010), pp. 2701–2714.

[TAW11] V. Y. Tan, A. Anandkumar, and A. S. Willsky. "Learning high-dimensional Markov forest distributions: Analysis of error rates". In: *Journal of Machine Learning Research* 12 (2011), pp. 1617–1653.

[TPS12] A. G. Tartakovsky, A. S. Polunchenko, and G. Sokolov. "Efficient computer network anomaly detection by changepoint detection methods". In: *IEEE Journal of Selected Topics in Signal Processing* 7.1 (2012), pp. 4–11.

[TWS18] E. J. T. Tchetgen, L. Wang, and B. Sun. "Discrete choice models for nonmonotone nonignorable missing data: Identification and inference". In: *Statistica Sinica* 28.4 (2018), p. 2069.

[Tch06] E. J. T. Tchetgen. "Statistical methods for robust inference in causal and missing data models". Harvard University, 2006.

[Tha04] P. Thagard. "Causal inference in legal decision making: Explanatory coherence vs. Bayesian networks". In: *Applied Artificial Intelligence* 18.3-4 (2004), pp. 231–249.

[Tob58] J. Tobin. "Estimation of relationships for limited dependent variables". In: *Econometrica: journal of the Econometric Society* (1958), pp. 24–36.

[TS13] T. H. Tom Claassen Joris M. Mooij and P. Smyth. "Learning Sparse Causal Models is not NP-hard". In: *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, UAI 2013, Bellevue, WA, USA, August 11-15, 2013*. AUAI Press, 2013.

[TMD22] D. Tramontano, A. Monod, and M. Drton. "Learning linear non-Gaussian polytree models". In: *Uncertainty in Artificial Intelligence*. PMLR. 2022, pp. 1960–1969.

[Tro+01] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. "Missing value estimation methods for DNA microarrays". In: *Bioinformatics* 17.6 (2001), pp. 520–525.

[Tsi06] A. A. Tsiatis. "Semiparametric theory and missing data". In: (2006).

[Tuk60] J. W. Tukey. "A survey of sampling from contaminated distributions". In: *Contributions to probability and statistics* (1960), pp. 448–485.

[Uhl17] C. Uhler. "Gaussian Graphical Models: An Algebraic and Geometric Perspective". 2017. arXiv: 1707.04345 [math.ST].

[Uhl+13] C. Uhler, G. Raskutti, P. Bühlmann, and B. Yu. "Geometry of the faithfulness assumption in causal inference". In: *Annals of Statistics* 41.2 (2013), pp. 436–463.

[Val84] L. G. Valiant. "A Theory of the Learnable". In: *Commun. ACM* 27.11 (Nov. 1984), pp. 1134–1142. ISSN: 0001-0782. URL: https://doi.org/10.1145/1968.1972.

[Van18] S. Van Buuren. "Flexible imputation of missing data". CRC press, 2018.

[VLP08] F. Van Harmelen, V. Lifschitz, and B. Porter. "Handbook of knowledge representation". Elsevier, 2008.

[Vap99] V. N. Vapnik. "An overview of statistical learning theory". In: *IEEE transactions on neural networks* 10.5 (1999), pp. 988–999.

[VP90] T. Verma and J. Pearl. "Equivalence and Synthesis of Causal Models In Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence (P. Bonissone, M. Henrion, L. Kanal and J. Lemmer, eds.) 220–227". 1990.

[VMB23] J. Von Kügelgen, A. Mohamed, and S. Beckers. "Backtracking Counterfactuals". In: *Proceedings of the Second Conference on Causal Learning and Reasoning*. Ed. by M. van der Schaar, C. Zhang, and D. Janzing. Vol. 213. Proceedings of Machine Learning Research. PMLR, Nov. 2023, pp. 177–196.

[Vov20] V. Vovk. "Non-algorithmic theory of randomness". In: *Fields of Logic and Computation III: Essays Dedicated to Yuri Gurevich on the Occasion of His 80th Birthday*. Springer, 2020, pp. 323–340.

[VW21] V. Vovk and R. Wang. "E-values: Calibration, combination and applications". In: *The Annals of Statistics* 49.3 (2021), pp. 1736–1754.

[Wai19] M. J. Wainwright. "High-dimensional statistics: A non-asymptotic viewpoint". Vol. 48. Cambridge University Press, 2019.

[WJ08a] M. J. Wainwright and M. I. Jordan. "Graphical models, exponential families, and variational inference". Now Publishers Inc, 2008.

[WJ08b] M. J. Wainwright and M. I. Jordan. "Graphical models, exponential families, and variational inference". Now Publishers Inc, 2008.

[Wan+23a] D. Wang, Z. Chen, Y. Fu, Y. Liu, and H. Chen. "Incremental causal graph learning for online root cause analysis". In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2023, pp. 2269–2278.

[Wan+23b] D. Wang, Z. Chen, J. Ni, L. Tong, Z. Wang, Y. Fu, and H. Chen. "Interdependent causal networks for root cause localization". In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2023, pp. 5051–5060.

[WWR10] W. Wang, M. J. Wainwright, and K. Ramchandran. "Information-theoretic bounds on model selection for Gaussian Markov random fields". In: *2010 IEEE International Symposium on Information Theory*. IEEE. 2010, pp. 1373–1377.

[WH18] X. Wang and Y. Hong. "Characteristic function based testing for conditional independence: A nonparametric regression approach". In: *Econometric Theory* 34.4 (2018), pp. 815–849.

[WD20] Y. S. Wang and M. Drton. "High-dimensional causal discovery under non-Gaussianity". In: *Biometrika* 107.1 (2020), pp. 41–59.

[WB21a] Y. Wang and A. Bhattacharyya. "Identifiability of AMP chain graph models". In: *arXiv preprint arXiv:2106.09350* (2021).

[WB21b] Y. Wang and A. Bhattacharyya. "Identifiability of AMP chain graph models". In: *arXiv preprint arXiv:2106.09350* (2021).

[Wan+24] Y. Wang, M. Gao, W. M. Tai, B. Aragam, and A. Bhattacharyya. "Optimal estimation of Gaussian (poly) trees". In: *arXiv preprint arXiv:2402.06380* (2024).

[Wan+20] Y. Wang, V. Menkovski, H. Wang, X. Du, and M. Pechenizkiy. "Causal discovery from incomplete data: a deep learning approach". In: *arXiv preprint arXiv:2001.05343* (2020).

[War92] A. Warga. "Bond returns, liquidity, and missing data". In: *Journal of Financial and Quantitative Analysis* 27.4 (1992), pp. 605–617.

[WR06] L. Wasserman and K. Roeder. "Weighted hypothesis testing". In: (2006). arXiv: math/0604172 [math].

[WGY20] D. Wei, T. Gao, and Y. Yu. "DAGs with No Fears: A Closer Look at Continuous Optimization for Learning Bayesian Networks". In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. 2020.

[Wol23] D. H. Wolpert. "The implications of the no-free-lunch theorems for meta-induction". In: *Journal for General Philosophy of Science* 54.3 (2023), pp. 421–432.

[WW15] S. Wood and M. S. Wood. "Package mgcv". In: *R package version* 1 (2015), p. 29.

[Woo07] J. M. Wooldridge. "Inverse probability weighted estimation for general missing data problems". In: *Journal of econometrics* 141.2 (2007), pp. 1281–1301.

[WDS19] S. Wu, A. G. Dimakis, and S. Sanghavi. "Learning distributions generated by one-layer ReLU networks". In: *Advances in neural information processing systems* 32 (2019).

[Yak+21] K. Yakovlev, I. E. I. Bekkouch, A. M. Khan, and A. M. Khattak. "Abstraction-based outlier detection for image data". In: *Intelligent Systems and Applications: Proceedings of the 2020 Intelligent Systems Conference (IntelliSys) Volume 1*. Springer. 2021, pp. 540–552.

[Yu97] B. Yu. "Assouad, fano, and le cam". In: *Festschrift for Lucien Le Cam: research papers in probability and statistics*. Springer, 1997, pp. 423–435.

[Yu15] Y.-L. Yu. "Submodular Analysis, Duality and Optimization". http://www.cs.cmu.edu/~yaoliang/mynotes/submodular.pdf. Accessed: 2021-02-18. 2015.

[Yu+19] Y. Yu, J. Chen, T. Gao, and M. Yu. "DAG-GNN: DAG Structure Learning with Graph Neural Networks". In: *International Conference on Machine Learning*. 2019, pp. 7154–7163.

[Zha+13] B. Zhang, C. Gaiteri, L.-G. Bodea, Z. Wang, J. McElwee, A. A. Podtelezhnikov, C. Zhang, T. Xie, L. Tran, R. Dobrin, et al. "Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimers disease". In: *Cell* 153.3 (2013), pp. 707–720.

[Zha+17] H. Zhang, S. Zhou, K. Zhang, and J. Guan. "Causal discovery using regression-based conditional independence tests". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1. 2017.

[ZS16] J. Zhang and P. Spirtes. "The three faces of faithfulness". In: *Synthese* 193.4 (2016), pp. 1011–1027.

[ZH09] K. Zhang and A. Hyvärinen. "On the identifiability of the post-nonlinear causal model". In: *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press. 2009, pp. 647–655.

[Zha+12] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. "Kernel-based conditional independence test and application in causal discovery". In: *arXiv preprint arXiv:1202.3775* (2012).

[Zha+16] K. Zhang, Z. Wang, J. Zhang, and B. Schölkopf. "On estimation of functional causal models: general results and application to the post-nonlinear causal model". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 7.2 (2016), p. 13.

[Zhe+18] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing. "DAGs with no tears: Continuous optimization for structure learning". In: *Advances in Neural Information Processing Systems*. 2018, pp. 9472–9483.

[Zhe+20] X. Zheng, C. Dan, B. Aragam, P. Ravikumar, and E. Xing. "Learning sparse nonparametric DAGs". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 3414–3425.

[ZSA22] J. Zscheischler, J. Sillmann, and L. Alexander. "Introduction to the special issue: Compound weather and climate events". In: *Weather and Climate Extremes* 35 (2022), p. 100381. ISSN: 2212-0947. URL: https://www.sciencedirect.com/science/article/pii/S2212094721000712.

[ZMD12] O. Zuk, S. Margel, and E. Domany. "On the Number of Samples Needed to Learn the Correct Structure of a Bayesian Network". In: (June 2012). eprint: 1206.6862. URL: https://arxiv.org/abs/1206.6862.

[Zur89] W. H. Zurek. "Algorithmic randomness and physical entropy". In: *Physical Review A* 40.8 (1989), p. 4731.

Appendix A

Supplementary Material - Chapter 3

A.1 Proof of Fact 2.2.9

Proof. It is well-known that [Uhl17] if $\mathrm{Cov}(X) = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix} : \mathrm{Cov}(X_A) = \Sigma_{AA}, \mathrm{Cov}(X_B \mid X_A) = \Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB}$. On the other hand, it follows from the properties of Schur complement [Oue81] that: $\det(\Sigma) = \det(\Sigma_{AA}) \cdot \det(\Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB})$. The result follows. □

A.2 Proof of Theorem 3.4.1

Proof. We show that Algorithm 1 recovers the chain graph under the assumptions of the Theorem 3.4.1. This follows immediately from the following lemma, as it shows that at every step i, the algorithm chooses as τ_i a chain component whose parents are contained in the current $\mathcal{A}$.

Lemma 17. *Let $\mathcal{A}$ be an ancestral set of chain components, and let $P = \{v : v \in \tau \in \mathcal{A}\}$. Assume the condition (3.3) above. Suppose τ_1 and τ_2 are chain components in $\mathcal{C} \setminus \mathcal{A}$ such that $\mathrm{Pa}(\tau_1) \subseteq \mathcal{A}$ but $\mathrm{Pa}(\tau_2) \not\subseteq \mathcal{A}$. Then:*

$$d_{|\tau_1|}\left(\mathop{\mathbb{E}}_{X_P}\mathop{\mathrm{Cov}}_{X_{\tau_1}}(X_{\tau_1} \mid X_P)\right) < d_{|\tau_2|}\left(\mathop{\mathbb{E}}_{X_P}\mathop{\mathrm{Cov}}_{X_{\tau_2}}(X_{\tau_2} \mid X_P)\right)$$

Proof. Note that τ_1 must precede τ_2 in the topological ordering π, and hence (3.3) holds with $\tau = \tau_1$ and $\tau' = \tau_2$. We invoke the law of conditional covariance (Fact

2.2.8).

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{X_P} \mathop{\mathrm{Cov}}_{X_{\tau'}}(X_{\tau'} \mid X_P) \\
&= \mathop{\mathbb{E}}_{X_P} \mathop{\mathbb{E}}_{X_{\mathrm{Pa}(\tau')}} \left[\mathop{\mathrm{Cov}}_{X_{\tau'}}(X_{\tau'} \mid X_P, X_{\mathrm{Pa}(\tau')}) \mid X_P \right] + \\
&\quad \mathop{\mathbb{E}}_{X_P} \mathop{\mathrm{Cov}}_{X_{\mathrm{Pa}(\tau')}} \left(\mathop{\mathbb{E}}_{X_{\tau'}} (X_{\tau'} \mid X_P, X_{\mathrm{Pa}(\tau')}) \mid X_P \right) \\
&= \mathop{\mathbb{E}}_{X_P} \mathop{\mathbb{E}}_{X_{\mathrm{Pa}(\tau')}} \left[\mathop{\mathrm{Cov}}_{X_{\tau'}}(X_{\tau'} \mid X_{\mathrm{Pa}(\tau')}) \mid X_P \right] + \\
&\quad \mathop{\mathbb{E}}_{X_P} \mathop{\mathrm{Cov}}_{X_{\mathrm{Pa}(\tau')}} \left(\mathop{\mathbb{E}}_{X_{\tau'}} (X_{\tau'} \mid X_{\mathrm{Pa}(\tau')}) \mid X_P \right) \\
&= \mathop{\mathbb{E}}_{X_{\mathrm{Pa}(\tau')}} \left[\mathop{\mathrm{Cov}}_{X_{\tau'}}(X_{\tau'} \mid X_{\mathrm{Pa}(\tau')}) \right] + \\
&\quad \mathop{\mathbb{E}}_{X_P} \mathop{\mathrm{Cov}}_{X_{\mathrm{Pa}(\tau')}} \left(\mathop{\mathbb{E}}_{X_{\tau'}} (X_{\tau'} \mid X_{\mathrm{Pa}(\tau')}) \mid X_P \right)
\end{aligned}
$$

The second equality follows from the fact that $X_{\tau'}$ is independent of X_P, conditioned on $X_{\mathrm{Pa}(\tau')}$. Now, note that the second term in the last line above is positive definite if P does not contain $\mathrm{Pa}(\tau')$. Therefore, using the fact that $d_{|\tau'|}$ is positive and super-additive:

$$
\begin{aligned}
&d_{|\tau'|} \left(\mathop{\mathbb{E}}_{X_P} \mathop{\mathrm{Cov}}_{X_{\tau'}}(X_{\tau'} \mid X_P) \right) \\
&\geq d_{|\tau'|} \left(\mathop{\mathbb{E}}_{X_{\mathrm{Pa}(\tau')}} \left[\mathop{\mathrm{Cov}}_{X_{\tau'}}(X_{\tau'} \mid X_{\mathrm{Pa}(\tau')}) \right] \right) + \\
&d_{|\tau'|} \left(\mathop{\mathbb{E}}_{X_P} \mathop{\mathrm{Cov}}_{X_{\mathrm{Pa}(\tau')}} \left(\mathop{\mathbb{E}}_{X_{\tau'}} (X_{\tau'} \mid X_{\mathrm{Pa}(\tau')}) \mid X_P \right) \right) \\
&> d_{|\tau'|} \left(\mathop{\mathbb{E}}_{X_{\mathrm{Pa}(\tau')}} \left[\mathop{\mathrm{Cov}}_{X_{\tau'}}(X_{\tau'} \mid X_{\mathrm{Pa}(\tau')}) \right] \right) \\
&\geq d_{|\tau|} \left(\mathop{\mathbb{E}}_{X_{\mathrm{Pa}(\tau)}} \left[\mathop{\mathrm{Cov}}_{X_\tau}(X_\tau \mid X_{\mathrm{Pa}(\tau)}) \right] \right) \\
&= d_{|\tau|} \left(\mathop{\mathbb{E}}_{X_P} \left[\mathop{\mathrm{Cov}}_{X_\tau}(X_\tau \mid X_P) \right] \right)
\end{aligned}
$$

The third inequality is due to (3.3). The last equality holds since $\mathrm{Pa}(\tau) \subseteq P$, and hence, X_τ is independent of $X_{P \setminus \mathrm{Pa}(\tau)}$ conditioned on $X_{\mathrm{Pa}(\tau)}$. □

□

A.3 Non-parametric algorithm

Algorithm 1 can be converted into a finite-sample algorithm, as shown in Algorithm 13. Statistical guarantees can be provided, assuming the following conditions, analogous to those in [GDA20a].

Algorithm 13: Finite-sample algorithm for learning the topological order of a chain graph with chain component decomposition $\mathcal{C}$ of size t.

Input : $X^{(1)}, \ldots, X^{(n)}$, $\eta > 0$

$\mathcal{A}, P \leftarrow \emptyset$;
$i \leftarrow 0$;
while $|\mathcal{A}| \neq t$ **do**
  Randomly split the samples in half;
  For each $\tau \in \mathcal{C} \setminus \mathcal{A}$, use the first half of samples to estimate $\mathbb{E}[X_\tau \mid X_P]$ via a nonparametric estimator $\widehat{F}_{\tau,P}(X_P)$;
  For each $\tau \in \mathcal{C} \setminus \mathcal{A}$, use the second half of samples to get:

$$\widehat{\sigma}^2_{\tau,P} = \det\left(\frac{1}{n/2}\sum_{i=1}^{n/2}\left((X_\tau^{(i)})^{\otimes 2} - \widehat{F}^{\otimes 2}_{\tau,P}(X_P^{(i)})\right)\right)$$

  ;
  Set $\tau^* = \arg\min_{\tau \in \mathcal{C}\setminus\mathcal{A}} \widehat{\sigma}^2_{\tau,P}$, and let $\widehat{T}_i = \{\tau \in \mathcal{C} \setminus \mathcal{A} : |\widehat{\sigma}^2_{\tau,P} - \widehat{\sigma}^2_{\tau^*,P}| < \eta\}$;
  $\mathcal{A} \leftarrow \mathcal{A} \cup \{\widehat{T}_i\}$;
  $P \leftarrow P \cup \widehat{T}_i$;
  $i \leftarrow i + 1$;
return the ordering $(\widehat{T}_1, \ldots, \widehat{T}_t)$;

Suppose the DAG on the chain components has a layer decomposition $L_1, \ldots, L_r$, and let $P_j = L_1 \cup \cdots \cup L_{j-1}$ with $d_j = |P_j|$.

Condition A.3.1 (Regularity). For all f and all $\tau \notin P_j$, (a) $\|X_\tau\|_2 \in [0,1]$, (b) $F_{\tau,P_j} : [0,1]^{d_j} \to [0,1]^{|\tau|}$ where $F_{\tau,P_j}(X_{P_j}) = \mathbb{E}[X_\tau \mid P_j]$, (c) $F_{\tau,P_j} \in L^\infty([0,1]^{d_j})$, and (d) $\|\operatorname{Cov}(X_\tau \mid P_j)\|_F \leq \zeta_0 < \infty$.

Condition A.3.2 (Identifiability). $\det(\mathbb{E}\operatorname{Cov}(X_\tau \mid X_{Pa(\tau)})) = \sigma^2$ independent of τ.

Condition A.3.3 (Estimator). The estimator $\widehat{F}$ for estimating conditional expectation $\mathbb{E}[Y \mid Z]$ for joint distributions (Y, Z) satisfies: (i) $\mathbb{E}[Y \mid Z] \in L^\infty \implies \widehat{F} \in L^\infty$, and (b) $\mathbb{E}_{\widehat{F}} \|\widehat{F}(Z) - \mathbb{E}[Y \mid Z]\|_2^2 \to 0$.

Condition A.3.4 (Bounded chain component). The size of each chain component is bounded.

Under these conditions, the proof of Theorem 4.1 in [GDA20a] extends straightforwardedly to show that the probability of recovering the topological order on the chain components goes to 1 with increasing samples if η is chosen appropriately. The only difference is that one needs to use a bound on the Lipschitz constant of the determinant function, which follows because of the bounded chain component condition.

A.4 Proof of Theorem 3.5.1

Proof. For simplicity, suppose that there is a unique topological order $\tau_1 \preceq \tau_2 \preceq \cdots \preceq \tau_m$ for the components outside P. The proof easily extends to the general case. Let $\tau_{<i} = \tau_1 \cup \cdots \cup \tau_{i-1}$.

Consider any *non-empty*[1] subset S that is disjoint from P. We claim:

$$\begin{aligned}
&\det(\mathrm{Cov}(X_S \mid X_P)) \\
&= \prod_{i=1}^{m} \det(\mathrm{Cov}(X_{S\cap\tau_i} \mid X_{S\cap\tau_{<i}}, X_P)) && \text{(A.1)} \\
&\geq \prod_{i=1}^{m} \det(\mathrm{Cov}(X_{S\cap\tau_i} \mid X_{\mathrm{Pa}(\tau_i)})) && \text{(A.2)} \\
&\geq \prod_{i:S\cap\tau_i\neq\emptyset} \det(\mathrm{Cov}(X_{\tau_i} \mid X_{\mathrm{Pa}(\tau_i)})) && \text{(A.3)} \\
&\geq \det(\mathrm{Cov}(X_{\tau_1} \mid X_{\mathrm{Pa}(\tau_1)})) && \text{(A.4)}
\end{aligned}$$

(A.1) is a consequence of Fact 2.2.9. To prove (A.2), we invoke the law of conditional covariance (Fact 2.2.8):

$$\begin{aligned}
&\det(\mathrm{Cov}(X_{S\cap\tau_i} \mid X_{S\cap\tau_{<i}}, X_P)) \\
&= \det(\mathrm{Cov}(X_{S\cap\tau_i} \mid X_{\mathrm{Pa}(\tau_i)}, X_{S\cap\tau_{<i}}, X_P) \\
&\qquad + \mathrm{Cov}(\mathbb{E}[X_{S\cap\tau_i} \mid X_{\mathrm{Pa}(\tau_i)}] \mid X_{S\cap\tau_{<i}}, X_P)) \\
&\geq \det(\mathrm{Cov}(X_{s\cap\tau_i} \mid X_{\mathrm{Pa}(\tau_i)}))
\end{aligned}$$

[1]The condition that S is non-empty was omitted by error in Algorithm 2. Note that the minimization over non-empty sets is still a submodular optimization problem.

The last inequality uses the positive semi-definiteness of covariance matrices and super-additivity of the determinant. The proof of (A.3) uses Fact 2.2.9 as follows:

$$\begin{aligned}
&\det(\mathrm{Cov}(X_{\tau_i} \mid X_{\mathrm{Pa}(\tau_i)})) \\
&= \det(\mathrm{Cov}(X_{S\cap\tau_i} \mid X_{\mathrm{Pa}(\tau_i)}))\cdot \\
&\qquad\qquad \det(\mathrm{Cov}(X_{\tau_i\setminus S} \mid X_{S\cap\tau_i}, X_{\mathrm{Pa}(\tau_i)})) \\
&\leq \det(\mathrm{Cov}(X_{S\cap\tau_i} \mid X_{\mathrm{Pa}(\tau_i)}))
\end{aligned}$$

using condition (iii) of Theorem 3.5.1. (The last inequality is non-strict because $\tau_i \setminus S$ may be empty.) The inequality (A.4) follows from conditions (i) and (ii) of Theorem 3.5.1.

For (A.4) to be an equality, S must be contained in exactly one component τ. For (A.3) to be an equality, S must equal τ. For (A.2) to be an equality, the parents of $S \cap \tau = S$ must be contained in P, and hence $S = \tau = \tau_1$. □

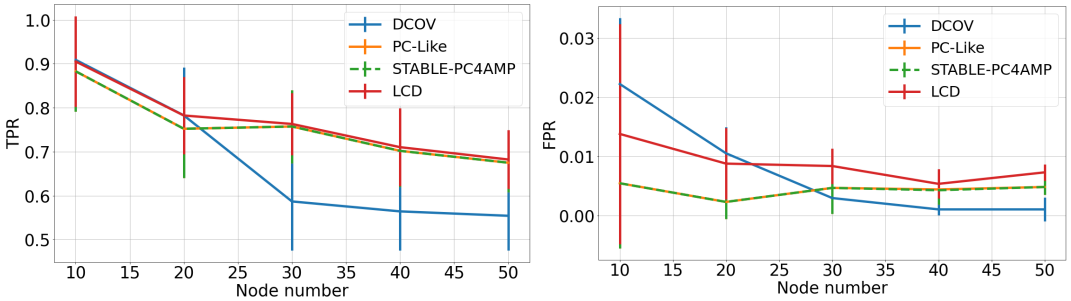
A.5 Performance Evaluation Metrics

We evaluate the performance of each algorithm using the following four metrics:

- The true positive rate (TPR): the ratio of the number of correctly identified edges in estimated graphs (TP) over total number of edges in true graph (Pos).
- The false positive rate (FPR): the ratio of the number of incorrectly identified edges (FP) over total number of gaps (Neg).
- Accuracy (Acc): Acc is defined as the ratio of (TP + TN) over (Pos + Neg).
- Structure hamming distance (SHD): SHD counts the total number operations, including edge additions, deletions, and reversals, that are needed to convert the current resulting graph into the true CG.

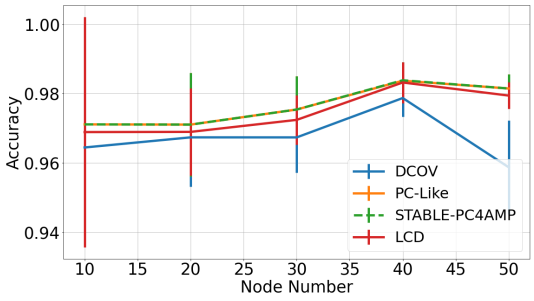
A.6 Agnostic Learning Experiments

In this section, we explore the learning of non-realizable inputs, where our assumptions in Theorem 3.5.1 do not holds. To do this, we conduct chain graph

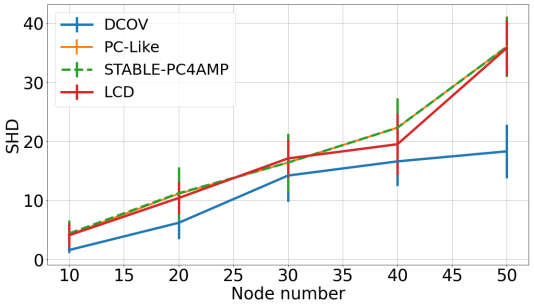


(a) TPR

(b) FPR



(c) ACC



(d) SHD

Figure A.1: Agnostic learning experiments on chain graph using Algorithm 2. The plots show the impact on: (a) True Positive Rate (TPR), (b) False Positive Rate (FPR), (c) Accuracy (ACC), and (d) Structural Hamming Distance (SHD).

experiments by fitting the linear AMP chain graph data generated from the opposite condition of Theorem 3.5.1:

$$\det(\mathrm{Cov}(X_\tau \mid X_{\mathrm{Pa}(\tau)})) \leq 1.$$

The experiment results are available in Figure A.1. Even the SHD performance is the best among all other baseline algorithms. The FPR is higher with less nodes but lower when nodes are more than 30. However, we have the worst ACC. The TPR of our algorithm also decreases sharply when node number beyond 30.

A.7 DAG synthetic data generation

For our experiments on DAG identifiability, we explored Erdős-Rényi graphs using the Python package `networkx` [HSS08]. Our non-zero edge weights are uniformly drawn from the range $(-2, -1] \cup [1, 2)$. Once the graph is generated, the linear i.i.d. data $X \in \mathbb{R}^{n \times d}$ (with d variables and sample size $n = 1000$) is generated

by sampling the model $X = M^T X + z$, where M is a strictly upper triangular matrix and $z \sim \mathcal{N}(0, \sigma^2)$. We conduct two sets of experiments where σ follows one of the following two conditions:

1. The variance for each node is equal and > 1;

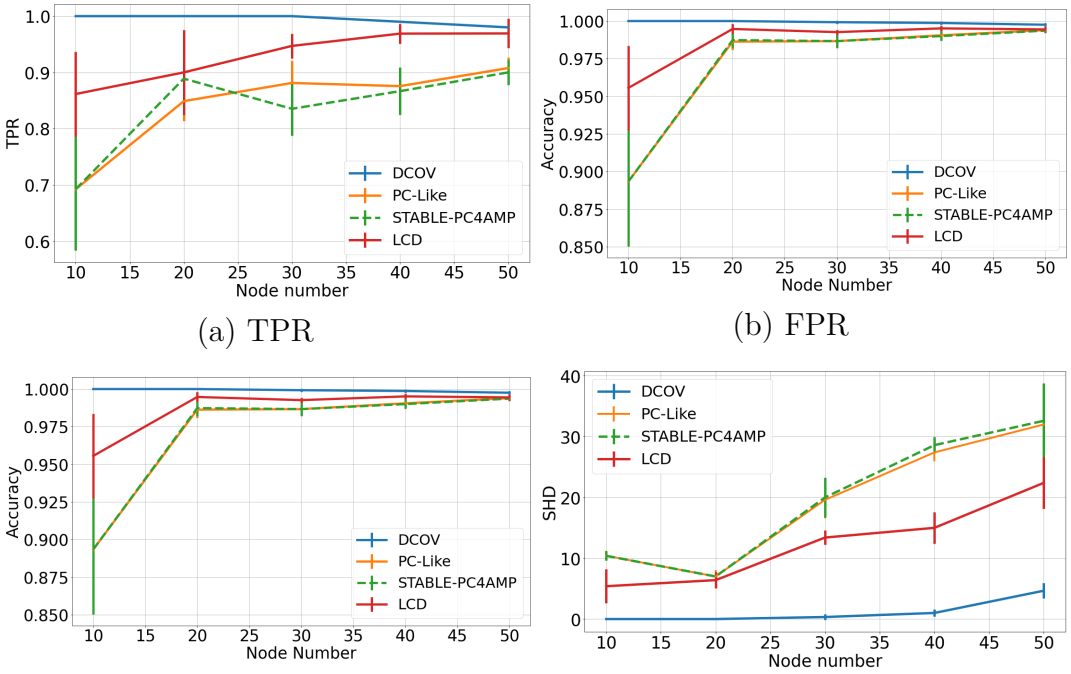
2. The variance for each node is uniformly in $[0.5, 1.5]$;

A.8 Identifiability of DAG structures

In previous section, we conduct identifiability algorithms with theoretical guarantees for learning linear AMP chain graph models from observational data. Since DAG is a special case of AMP chain graph models, it is of interest to see when the variance of each node is equal and > 1, whether Algorithm 2 can successfully identify the special one-node chain component structure and recover the DAG graph from observational data. As shown in Figure A.2, our algorithm achieves the highest TPR, ACC and the lowest FPR and SHD. Thus, we can identify not only the general case of AMP chain graph models but also the special case of DAG structures. This is because when our assumption in Theorem 3.5.1 holds, we can correctly identify both the partitioning into chain components and the topological order on the chain components using a polynomial time algorithm based on submodular function minimization.

A.9 Agnostic learning experiments on synthetic Bayesian Network

We also evaluate the performance of agnostic learning experiments on DAGs when the node variance is uniformly in between $[0.5, 1.5]$. The experiment results are in Figure A.3. Our algorithm on TPR is relatively better than other baseline methods. However, we have the highest FPR, SHD, and the lowest ACC among all other methods. This is because when our condition in Theorem 3.5.1 do not hold, in the worst case, Algorithm 2 will treat all DAG nodes as one chain component and thus return an undirected graph.



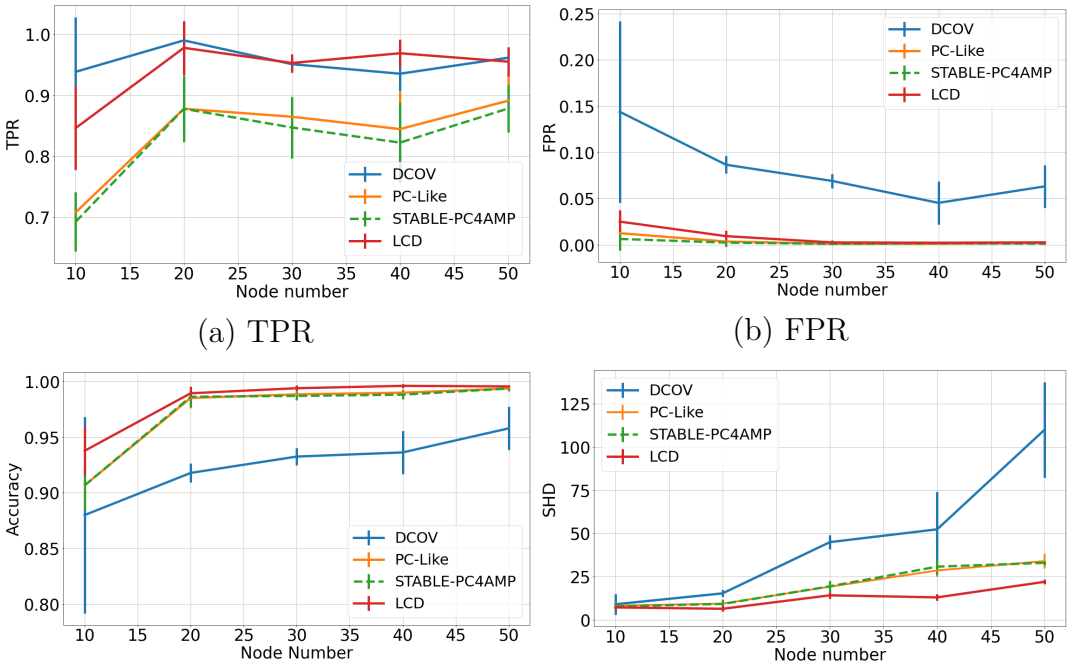
(a) TPR (b) FPR

(c) ACC (d) SHD

Figure A.2: Identifiability of DAG structures using Algorithm 2, variance greater than 1.

A.10 Experiments on Real Bayesian Networks

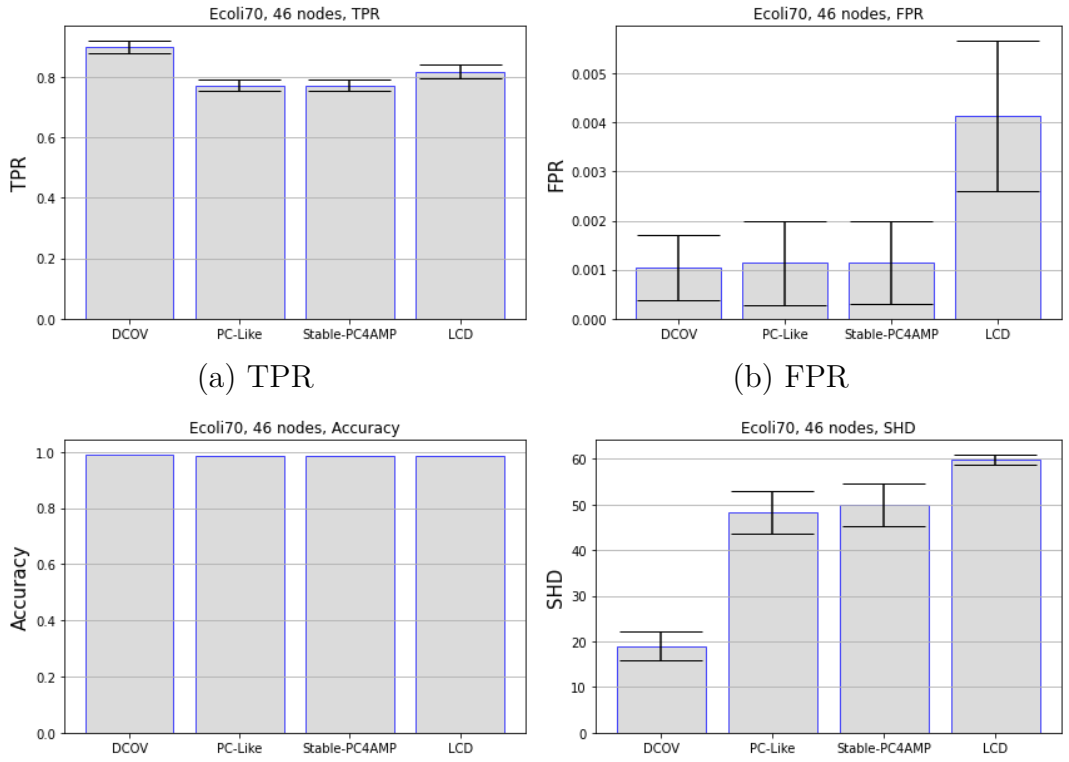
Besides, we are also interested in evaluating the structure learning performance of Algorithm 1 over real Bayesian graphs. Experiment results in Figure A.4 report the performance of the **ECOLI70** graph. Figure A.5 report the experiment results over the **MAGIC-NIAB** graph. We also compared the performance over **MAGIC-IRRI** graph in Figure A.6. Our algorithm performs best among all these three real Bayesian graphs.



(a) TPR (b) FPR

(c) ACC (d) SHD

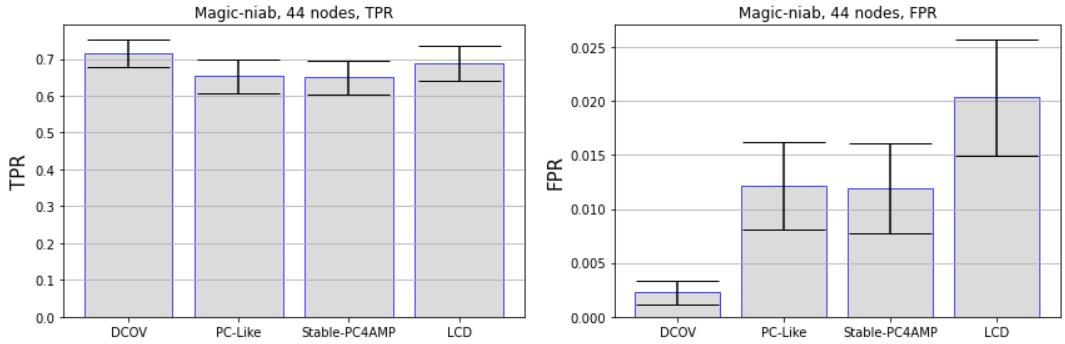
Figure A.3: Agnostic learning experiments on DAGs using Algorithm 2, uniform variance, variance in [0.5-1.5]



(a) TPR (b) FPR

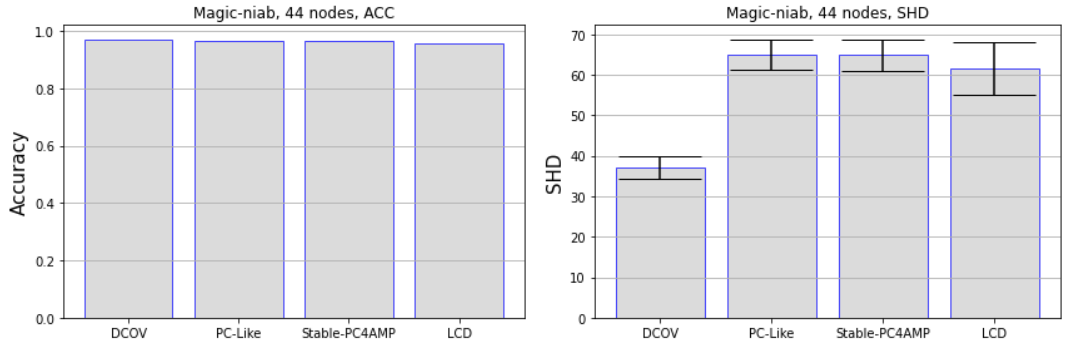
(c) ACC (d) SHD

Figure A.4: Ecoli70, 46 node Bayesian network



(a) TPR

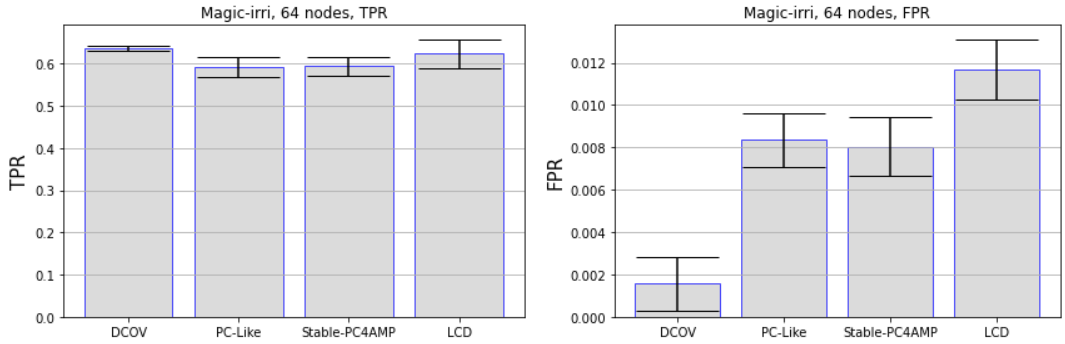
(b) FPR



(c) ACC

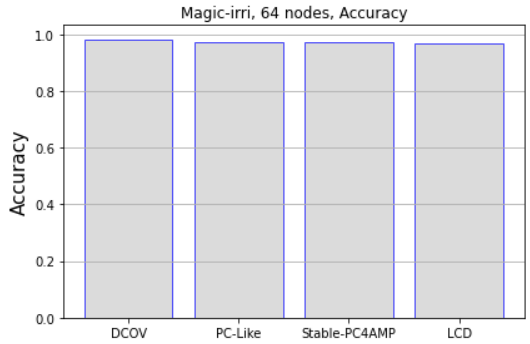
(d) SHD

Figure A.5: Magic-niab, 44 node Bayesian network

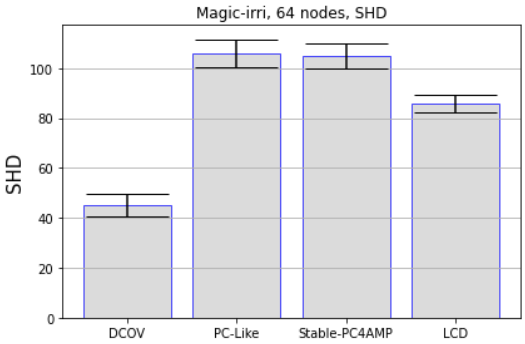


(a) TPR

(b) FPR



(c) ACC



(d) SHD

Figure A.6: Magic-irri, 64 node Bayesian network

Appendix B

Supplementary Material - Chapter 4

B.1 Comparing structure learning and distribution learning

Lemma 18. *Suppose $T \in \mathcal{T}$ and P is parameterized using $\{\beta_k, \sigma_k^2\}_{k=1}^d$ as Equation 8.9 according to T. If there exists a constant $M > 1$ such that for any $k \in [d]$,*

$$|\beta_{kj}| \in [M^{-1}, M], \qquad \forall \beta_{kj} \neq 0$$
$$\sigma_k^2 \in [M^{-1}, M]\,,$$

then P is c-strong Tree-faithful to T for some $c \asymp 1$.

Proof of Lemma 18. Since a directed tree T does not have any v-structures, we only need to verify adjacency faithfulness in Definition 4.3.2. For any two nodes connected as $j \to k$, we want to check whether $\rho(X_j, X_k \,|\, X_\ell)$ is lower bounded by some constant for $\ell \in V \cup \{\emptyset\} \setminus \{j, k\}$. There are four cases of ℓ to consider, see Fig. B.1:

- $\ell = \emptyset$: To simplify the notation, we write

  $$X_k = \beta_k \times X_j + \eta_k$$

  with $\beta_k \in \mathbb{R}$ and $|\beta_k| \in [M^{-1}, M]$, $\mathrm{Var}(\eta_k) = \sigma_k^2$. We also write $V_j^2 := \mathrm{Var}(X_j) \geq \sigma_j^2$. Hence,

  $$\rho(X_j, X_k) = \frac{\beta_k V_j^2}{\sqrt{V_j^2}\sqrt{\beta_k^2 V_j^2 + \sigma_k^2}} = \frac{1}{\sqrt{1 + \sigma_k^2/\beta_k^2 V_j^2}} \geq \frac{1}{\sqrt{1 + \sigma_k^2/\beta_k^2 \sigma_j^2}} \gtrsim 1\,.$$

- $\ell \in \mathrm{an}(j)$: Write $V^2_{j\,|\,\ell} = \mathrm{Var}(X_j \,|\, X_\ell) \geq \sigma_j^2$, hence

$$\rho(X_j, X_k \,|\, X_\ell) = \frac{\beta_k V^2_{j\,|\,\ell}}{\sqrt{V^2_{j\,|\,\ell}}\sqrt{\beta_k^2 V^2_{j\,|\,\ell} + \sigma_k^2}} \geq \frac{1}{\sqrt{1 + \sigma_k^2/\beta_k^2\sigma_j^2}} \gtrsim 1\,.$$

- $\ell \in \mathrm{de}(j)$: Suppose the directed path from j to ℓ is $j \to h_1 \to h_2 \to \ldots \to h_q \to \ell$, q can be 0, then we can write

$$X_\ell = b_1 X_j + u_1\,,$$

with

$$b_1 = \beta_\ell \prod_{i=1}^{q} \beta_{h_i}, \qquad u_1 = \eta_\ell + \beta_\ell \sum_{i=1}^{q} \eta_{h_i} \prod_{t=i+1}^{q} \beta_{h_t}\,,$$

and

$$\nu_1^2 := \mathrm{Var}(u_1) = \sigma_\ell^2 + \beta_\ell^2 \sum_{i=1}^{q} \sigma_{h_i}^2 \prod_{t=i+1}^{q} \beta_{h_t}^2 \geq \beta_\ell^2 \sigma_{h_1}^2 \prod_{t=2}^{q} \beta_{h_t}^2\,.$$

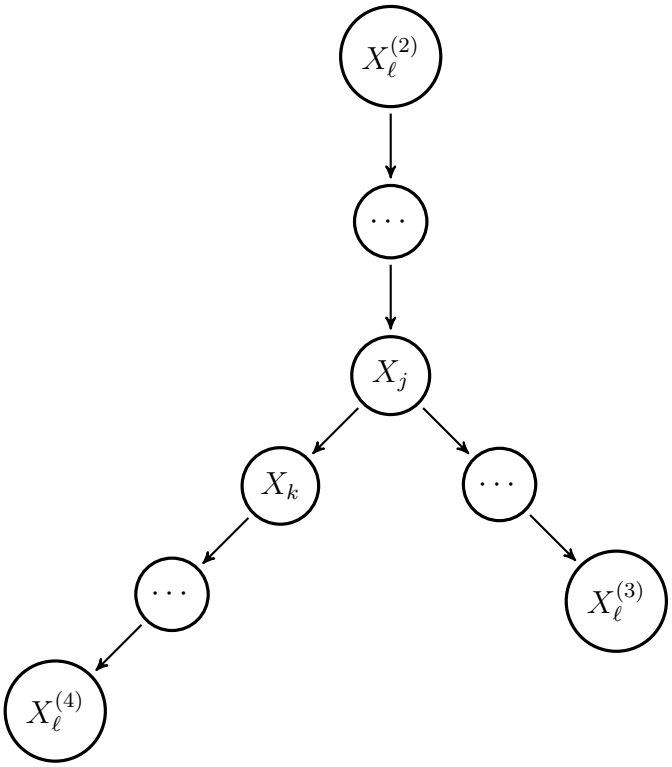


Figure B.1: Four cases of ℓ to verify for c-strong Tree-faithfulness, indicated by the superscript of X_ℓ. The first case is when $\ell = \emptyset$. The second, third and fourth are when ℓ is the ancestor of j, descendant of j and descendant of k.

So we have $b_1^2/\nu_1^2 \le \beta_{h_1}^2/\sigma_{h_1}^2 \asymp 1$. The covariance among X_j, X_k, X_ℓ is

$$\mathrm{Cov}(X_j, X_k, X_\ell) = \begin{pmatrix} V_j^2 & \beta_k V_j^2 & b_1 V_j^2 \\ * & \beta_k^2 V_j^2 + \sigma_k^2 & b_1 \beta_k V_j^2 \\ * & * & b_1^2 V_j^2 + \nu_1^2 \end{pmatrix}.$$

Then the conditional covariance is

$$\mathrm{Cov}(X_j, X_k \,|\, X_\ell) \propto \begin{pmatrix} \nu_1^2 & \beta_k \nu_1^2 \\ * & \beta_k^2 \nu_1^2 + \sigma_k^2 b_1^2 + \sigma_k^2 \nu_1^2 / V_j^2 \end{pmatrix}.$$

Therefore,

$$\rho(X_j, X_k \,|\, X_\ell) = \frac{1}{\sqrt{1 + \frac{\sigma_k^2}{\beta_k^2} \times \frac{b_1^2}{\nu_1^2} + \frac{\sigma_k^2}{V_j^2 \beta_k^2}}} \gtrsim 1\,.$$

- $\ell \in \mathrm{de}(k)$: Similarly, we can write

$$X_\ell = b_2 X_k + u_2, \qquad \mathrm{Var}(u_2) = \nu_2^2\,,$$

with $b_2^2/\nu_2^2 \lesssim 1$. The covariance among X_j, X_k, X_ℓ is

$$\mathrm{Cov}(X_j, X_k, X_\ell) = \begin{pmatrix} V_j^2 & \beta_k V_j^2 & b_2 \beta_k V_j^2 \\ * & V_k^2 & b_2 V_k^2 \\ * & * & b_2^2 V_k^2 + \nu_2^2 \end{pmatrix}.$$

Then the conditional covariance is

$$\mathrm{Cov}(X_j, X_k \,|\, X_\ell) \propto \begin{pmatrix} b_2^2 \sigma_k^2 V_j^2 + \nu_2^2 V_j^2 & \beta_k V_j^2 \nu_2^2 \\ * & \nu_2^2 V_k^2 \end{pmatrix}.$$

Therefore,

$$\rho(X_j, X_k \,|\, X_\ell) = \frac{1}{\sqrt{(1 + \frac{\sigma_k^2}{\beta_k^2 V_j^2})(1 + \frac{b_2^2}{\nu_2^2}\sigma_k^2)}} \gtrsim 1\,.$$

In all four cases, $\rho(X_j, X_k \,|\, X_\ell) \gtrsim 1$, thus c-strong Tree-faithfulness is satisfied with some $c \asymp 1$. □

Lemma 19. *Let $\mathcal{A}$ denote some distribution learning algorithm such that given a tree-structured distribution P, $\mathcal{A}$ takes data from P and outputs $\widehat{P}$ with $D_{\mathrm{KL}}(P \| \widehat{P}) \le \varepsilon$. If $\varepsilon \gtrsim c^2$, then for any estimator $\widehat{T}(\widehat{P})$ for $\overline{T}$ using solely $\widehat{P}$,*

$$\inf_{\widehat{T}(\widehat{P})} \sup_{\substack{T \in \mathcal{T} \\ P \text{ is } c\text{-strong} \\ \text{Tree-faithful to } T}} \sup_{\mathcal{A}} \Pr(\widehat{T}(\widehat{P}) \neq \overline{T}) = 1\,.$$

Proof. We construct $T, T' \in \mathcal{T}$ with different skeletons, and P, P' Markov and strongly faithful to T, T' respectively such that $D_{\mathrm{KL}}(P\|P') \asymp c^2$. In this way, consider the ground truth to be T and P, and supppose $\mathcal{A}$ outputs $\widehat{P} = P'$. Then we have $D_{\mathrm{KL}}(P\|\widehat{P}) \leq \varepsilon$ with $\varepsilon \asymp c^2$. While P and $\widehat{P} = P'$ correspond to different structures, thus any estimator using solely $\widehat{P}$ cannot uniformly find the true structure.

It remains to show the construction: Consider T and T' as follows:

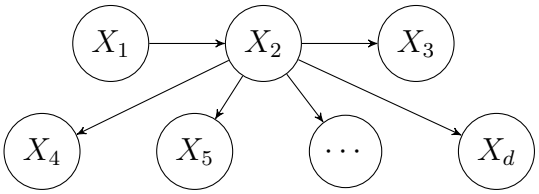


Figure B.2: Tree T

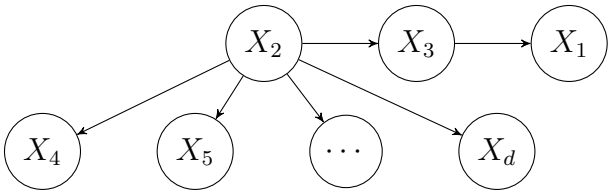


Figure B.3: Tree T'

Figure B.4: Construction for Lemma 19.

We parameterize P, P' as the lower bound construction in Appendix B.3.4:

$$X_k = \beta X_{\mathrm{pa}(k)} + \eta_k \,,$$

where $\beta = \sqrt{2}c$, $\eta_k \sim \mathcal{N}(0,1)$ and Lemma 29 makes sure they are c-strong tree faithful. Now we only need to compute the KL divergence:

$$\begin{aligned}
D_{\mathrm{KL}}(P\|P') &= \mathop{\mathbf{E}}_{P} \log \frac{\prod_k P(X_k \,|\, \mathrm{pa}(k))}{\prod_j P'(X_j \,|\, \mathrm{pa}(j))} \\
&= \mathop{\mathbf{E}}_{P} \log \frac{P(X_3 \,|\, X_2) P(P_2 \,|\, X_1) P(X_1)}{P(X_1 \,|\, X_3) P(X_3 \,|\, X_2) P(X_2)} \\
&= \mathop{\mathbf{E}}_{P} \frac{1}{2} \Big(X_2^2 + (X_1 - \beta X_3)^2 - X_1^2 - (X_3 - \beta X_2)^2 \Big) \\
&= \frac{1}{2} \Big(-\beta^4 + \beta^6 + 2(\beta^2 + \beta^4 - \beta^3) \Big) \\
&\leq 2\beta^2 = 4c^2 \,,
\end{aligned}$$

which completes the proof. □

B.2 Proofs of Section 4.2

B.2.1 Preliminaries

We first state some useful lemmas. They are well-known results for the concentration bound on variances and covariances. For completeness, we provide the proof below.

Lemma 20 (Guarantees of variance recovery)**.** *Suppose X is the random variable of $\mathbb{N}(0, \sigma^2)$ for some $\sigma > 0$. Let $X^{(1)}, \ldots, X^{(n)}$ be the i.i.d. samples of X and $\widehat{\sigma}^2$ be $\frac{1}{n}\sum_{i=1}^{n}(X^{(i)})^2$. Then, for any $t \in (0,1)$, we have*

$$\left|\widehat{\sigma}^2 - \sigma^2\right| < t\sigma^2$$

with probability $1 - O(e^{-\Omega(nt^2)})$.

Proof. We first show that the probability of $\widehat{\sigma}^2 > (1+t)\sigma^2$ is bounded by $e^{-\Omega(nt^2)}$ and the other inequality $\widehat{\sigma}^2 < (1-t)\sigma^2$ follows similarly.

Note that

$$\widehat{\sigma}^2 > (1+t)\sigma^2 \Leftrightarrow e^{\lambda\frac{1}{n}\sum_{i=1}^{n}(X^{(i)})^2} > e^{\lambda(1+t)\sigma^2} \qquad \text{for any } \lambda > 0.$$

By Markov inequality, the probability of $\widehat{\sigma}^2 > (1+t)\sigma^2$ is bounded by

$$\mathbf{E}(e^{\lambda\frac{1}{n}\sum_{i=1}^{n}(X^{(i)})^2})/e^{\lambda(1+t)\sigma^2} = \underbrace{\mathbf{E}(e^{\lambda\frac{1}{n}X^2})^n}_{\text{by i.i.d. assumption}}/e^{\lambda(1+t)\sigma^2}. \tag{B.1}$$

Hence, we need to bound the term $\mathbf{E}(e^{\lambda\frac{1}{n}X^2})$.

$$\mathbf{E}(e^{\lambda\frac{1}{n}X^2}) = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}e^{\lambda\frac{1}{n}x^2}e^{-\frac{1}{2\sigma^2}x^2}dx = \frac{1}{\sqrt{1-\frac{2\sigma^2\lambda}{n}}} \quad \text{as long as } \frac{1}{2\sigma^2} - \frac{\lambda}{n} > 0$$

Moreover, using the inequality $\frac{1}{\sqrt{1-x}} \le e^{\frac{1}{2}x+x^2}$ for $x < \frac{1}{2}$, we have

$$\mathbf{E}(e^{\lambda\frac{1}{n}X^2}) \le e^{\frac{\sigma^2\lambda}{n}+\frac{4\sigma^4\lambda^2}{n^2}} \qquad \text{as long as } \frac{2\sigma^2\lambda}{n} < \frac{1}{2} \tag{B.2}$$

Plugging Equation B.2 into Equation B.1, the probability of $\widehat{\sigma}^2 > (1+t)\sigma^2$ is bounded by

$$(e^{\frac{\sigma^2\lambda}{n}+\frac{4\sigma^4\lambda^2}{n^2}})^n/e^{\lambda(1+t)\sigma^2} = e^{-\frac{4\sigma^4\lambda^2}{n}+\lambda t\sigma^2} = e^{-\frac{4\sigma^4}{n}(\lambda-\frac{nt}{8\sigma^2})^2+\frac{nt^2}{16}}$$

and, by taking $\lambda = \frac{nt}{8\sigma^2}$, it becomes $e^{-\frac{nt^2}{16}}$.

□

Lemma 21 (Guarantees of correlation coefficient recovery)**.** *Suppose* (X, Y) *is the random variable of* $\mathbb{N}(0, \Sigma)$ *for some positive definite* $\Sigma = \begin{bmatrix} \sigma_x^2 & \rho_{xy} \\ \rho_{xy} & \sigma_y^2 \end{bmatrix}$*. Let* $(X^{(1)}, Y^{(1)}), \ldots, (X^{(n)}, Y^{(n)})$ *be the i.i.d. samples of* (X, Y) *and* $\widehat{\rho}_{xy}$ *be* $\frac{1}{n}\sum_{i=1}^{n} X^{(i)}Y^{(i)}$*. Then, for any* $t \in (0, 1)$*, we have*

$$|\widehat{\rho}_{xy} - \rho_{xy}| < t\sigma_x\sigma_y$$

with probability $1 - O(e^{-\Omega(nt^2)})$*.*

Proof. We first show that the probability of $\widehat{\rho}_{xy} > \rho_{xy} + t\sigma_x\sigma_y$ is bounded by $e^{-\Omega(nt^2)}$ and the other inequality $\widehat{\rho}_{xy} < \rho - t\sigma_x\sigma_y$ follows similarly.

Note that

$$\widehat{\rho}_{xy} > \rho_{xy} + t\sigma_x\sigma_y \Leftrightarrow e^{\lambda\frac{1}{n}\sum_{i=1}^{n} X^{(i)}Y^{(i)}} > e^{\lambda(\rho_{xy}+t\sigma_x\sigma_y)} \qquad \text{for any } \lambda > 0.$$

By Markov inequality, the probability of $\widehat{\rho}_{xy} > \rho_{xy} + t\sigma_x\sigma_y$ is bounded by

$$\mathbf{E}(e^{\lambda\frac{1}{n}\sum_{i=1}^{n} X^{(i)}Y^{(i)}})/e^{\lambda(\rho_{xy}+t\sigma_x\sigma_y)} = \underbrace{\mathbf{E}(e^{\lambda\frac{1}{n}XY})^n}_{\text{by i.i.d. assumption}} /e^{\lambda(\rho_{xy}+t\sigma_x\sigma_y)}. \tag{B.3}$$

Hence, we need to bound the term $\mathbf{E}(e^{\lambda\frac{1}{n}XY})$.

$$\begin{aligned}
\mathbf{E}(e^{\lambda\frac{1}{n}XY}) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{\sqrt{(2\pi)^2(\sigma_x^2\sigma_y^2 - \rho_{xy}^2)}} e^{\lambda\frac{1}{n}xy} e^{-\frac{1}{2(\sigma_x^2\sigma_y^2-\rho_{xy}^2)}(\sigma_y^2x^2 - 2\rho_{xy}xy + \sigma_x^2y^2)} dxdy \\
&= \frac{1}{\sqrt{1 - \frac{2\rho_{xy}\lambda}{n} - \frac{\lambda^2\Delta}{n^2}}} \qquad \text{as long as } \sigma_x^2\sigma_y^2 > (\rho_{xy} + \frac{\lambda\Delta}{n})^2 \text{ where } \Delta = \sigma_x^2\sigma_y^2 - \rho_{xy}^2
\end{aligned}$$

Moreover, using the inequality $\frac{1}{\sqrt{1-x}} \leq e^{\frac{1}{2}x+x^2}$ for $x < \frac{1}{2}$, we have

$$\begin{aligned}
\mathbf{E}(e^{\lambda\frac{1}{n}XY}) &\leq e^{\frac{1}{2}(\frac{2\rho_{xy}\lambda}{n} + \frac{\lambda^2\Delta}{n^2}) + (\frac{2\rho_{xy}\lambda}{n} + \frac{\lambda^2\Delta}{n^2})^2} \qquad \text{as long as } \frac{2\rho_{xy}\lambda}{n} + \frac{\lambda^2\Delta}{n^2} < \frac{1}{2} \\
&\leq e^{\frac{\rho_{xy}\lambda}{n} + \frac{\lambda^2\sigma_x^2\sigma_y^2}{2n^2} + (\frac{2\sigma_x\sigma_y\lambda}{n} + \frac{\lambda^2\sigma_x^2\sigma_y^2}{n^2})^2} \qquad \text{using } \rho_{xy} \leq \sigma_x\sigma_y \text{ and } \Delta \leq \sigma_x^2\sigma_y^2 \\
&\leq e^{\frac{\rho_{xy}\lambda}{n} + \frac{19\lambda^2\sigma_x^2\sigma_y^2}{2n^2}} \qquad \text{as long as } \frac{\lambda\sigma_x\sigma_y}{n} < 1
\end{aligned} \tag{B.4}$$

Plugging Equation B.4 into Equation B.3, the probability of $\widehat{\rho}_{xy} > \rho_{xy} + t\sigma_x\sigma_y$ is bounded by

$$(e^{\frac{\rho_{xy}\lambda}{n} + \frac{19\lambda^2\sigma_x^2\sigma_y^2}{2n^2}})^n / e^{\lambda(\rho_{xy}+t\sigma_x\sigma_y)} = e^{-\frac{19\sigma_x^2\sigma_y^2}{2n}\lambda^2 + t\sigma_x\sigma_y\lambda} = e^{-\frac{19\sigma_x^2\sigma_y^2}{2n}(\lambda - \frac{tn}{19\sigma_x\sigma_y})^2 + \frac{t^2n}{38}}$$

and, by taking $\lambda = \frac{tn}{19\sigma_x\sigma_y}$, it becomes $e^{-\frac{t^2n}{38}}$.

□

Corollary B.2.1. *Suppose* $(X_1, \ldots, X_d)$ *is the random variable of* $\mathbb{N}(0, \Sigma)$ *for some positive definite* Σ *where* $\rho_{ij} := \Sigma_{ij}$ *and* $\sigma_i^2 := \Sigma_{ii}$ *for* $i, j = 1, \ldots, d$. *Let* $(X_1^{(1)}, \ldots, X_d^{(1)}), \ldots, (X_1^{(n)}, \ldots, X_d^{(n)})$ *be the i.i.d. samples of* $(X_1, \ldots, X_d)$ *and*

$$\widehat{\rho}_{jk} = \frac{1}{n}\sum_{i=1}^{n} X_j^{(i)} X_k^{(i)} \qquad \text{and} \qquad \widehat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n} (X_j^{(i)})^2.$$

Then, when $n = \Theta(\frac{1}{t^2}\log\frac{d}{\delta})$, *we have, for all* $j, k = 1, \ldots, d$,

$$|\widehat{\rho}_{jk} - \rho_{jk}| \le t\sigma_j\sigma_k \qquad \text{and} \qquad \left|\widehat{\sigma}_j^2 - \sigma_j^2\right| \le t\sigma_j^2$$

with probability $1 - \delta$.

B.2.2 Conditional Mutual Information Tester

In this subsection, we define the conditional mutual information tester used in our main algorithm.

Suppose (X, Y, Z) is the random variable of $\mathbb{N}(0, \Sigma)$ for some positive definite $\Sigma = \begin{bmatrix} \sigma_x^2 & \rho_{xy} & \rho_{xz} \\ \rho_{xy} & \sigma_y^2 & \rho_{yz} \\ \rho_{xz} & \rho_{xy} & \sigma_z^2 \end{bmatrix}$. WLOG, we can express (X, Y, Z) as

$$Y = \beta_{xy} X + \eta_y$$
$$Z = \gamma_{xz} X + \gamma_{yz} Y + \eta_z$$

for some random variables η_y, η_z where

$$\beta_{xy} = \frac{\rho_{xy}}{\sigma_x^2} \qquad \text{and} \qquad \begin{bmatrix} \gamma_{xz} \\ \gamma_{yz} \end{bmatrix} = \begin{bmatrix} \sigma_x^2 & \rho_{xy} \\ \rho_{xy} & \sigma_y^2 \end{bmatrix}^{-1} \begin{bmatrix} \rho_{xz} \\ \rho_{yz} \end{bmatrix}.$$

Let $\sigma_{y|x}^2$ be $\mathbf{E}(\eta_y^2)$ and $\sigma_{z|x,y}^2$ be $\mathbf{E}(\eta_z^2)$. Recall that the mutual information $I(X;Y)$ and the conditional mutual information $I(Y;Z \mid X)$ are defined (equivalently) as

$$I(X;Y) := \frac{1}{2}\log(1 + \frac{\beta_{xy}^2\sigma_x^2}{\sigma_{y|x}^2}) \qquad \text{and} \qquad I(Y;Z \mid X) := \frac{1}{2}\log(1 + \frac{\gamma_{yz}^2\sigma_{y|x}^2}{\sigma_{z|x,y}^2})$$

Let $(X^{(1)}, Y^{(1)}, Z^{(1)}), \ldots, (X^{(n)}, Y^{(n)}, Z^{(n)})$ be the i.i.d. samples of (X, Y, Z). Then we define the empirical mutual information $\widehat{I}(X;Y)$ and the empirical mutual information $\widehat{I}(Y;Z \mid X)$ to be

$$\widehat{I}(X;Y) := \frac{1}{2}\log(1 + \frac{\widehat{\beta}_{xy}^2\widehat{\sigma}_x^2}{\widehat{\sigma}_{y|x}^2}) \qquad \text{and} \qquad \widehat{I}(Y;Z \mid X) := \frac{1}{2}\log(1 + \frac{\widehat{\gamma}_{yz}^2\widehat{\sigma}_{y|x}^2}{\widehat{\sigma}_{z|x,y}^2}) \quad \text{(B.5)}$$

where the $\widehat{\cdot}$ mark indicates the empirical version of the quantity. Namely,

$$\begin{cases} \widehat{\sigma}_x^2 & := \frac{1}{n}\sum_{i=1}^n (X^{(i)})^2, \qquad \widehat{\sigma}_y^2 := \frac{1}{n}\sum_{i=1}^n (Y^{(i)})^2, \qquad \widehat{\sigma}_z^2 := \frac{1}{n}\sum_{i=1}^n (Z^{(i)})^2, \\ \widehat{\rho}_{xy} & := \frac{1}{n}\sum_{i=1}^n X^{(i)}Y^{(i)}, \qquad \widehat{\rho}_{xz} := \frac{1}{n}\sum_{i=1}^n X^{(i)}Z^{(i)}, \qquad \widehat{\rho}_{yz} := \frac{1}{n}\sum_{i=1}^n Y^{(i)}Z^{(i)}, \\ \widehat{\beta}_{xy} & := \frac{\widehat{\rho}_{xy}}{\widehat{\sigma}_x^2}, \qquad \begin{bmatrix} \widehat{\gamma}_{xz} \\ \widehat{\gamma}_{yz} \end{bmatrix} := \begin{bmatrix} \widehat{\sigma}_x^2 & \widehat{\rho}_{xy} \\ \widehat{\rho}_{xy} & \widehat{\sigma}_y^2 \end{bmatrix}^{-1} \begin{bmatrix} \widehat{\rho}_{xz} \\ \widehat{\rho}_{yz} \end{bmatrix}, \\ \widehat{\sigma}_{y|x}^2 & := \widehat{\sigma}_y^2 - \widehat{\beta}_{xy}^2\widehat{\sigma}_x^2 \qquad \text{and} \qquad \widehat{\sigma}_{z|x,y}^2 := \widehat{\sigma}_z^2 - \widehat{\gamma}_{xz}^2\widehat{\sigma}_x^2 - \widehat{\gamma}_{yz}^2\widehat{\sigma}_{y|x}^2. \end{cases} \tag{B.6}$$

Note that the above quantities depend on the samples but we will not emphasize it if the set of samples is clear in the context. Also, it is known that, by the chain rule of mutual information,

$$I(X;Y) - I(X;Z) = I(X;Y \mid Z) - I(X;Z \mid Y) \tag{B.7}$$

$$\widehat{I}(X;Y) - \widehat{I}(X;Z) = \widehat{I}(X;Y \mid Z) - \widehat{I}(X;Z \mid Y). \tag{B.8}$$

From now on, when we have a d-dimensional random variable $(X_1, \ldots, X_d)$, we abuse the notations defined in Equation B.6 by replacing x, y, z with i, j, k for $i, j, k = 1, \ldots, d$.

Lemma 22. *Suppose $(X_1, \ldots, X_d)$ is the random variable of $\mathbb{N}(0, \Sigma)$ for some positive definite Σ where $\rho_{ij} := \Sigma_{ij}$ and $\sigma_i^2 := \Sigma_{ii}$ for $i, j = 1, \ldots, d$. Let*

$$(X_1^{(1)}, \ldots, X_d^{(1)}), \ldots, (X_1^{(n)}, \ldots, X_d^{(n)})$$

be the i.i.d. samples of $(X_1, \ldots, X_d)$ and $\widehat{\gamma}_{ij}, \widehat{\sigma}_{i|j}, \widehat{\sigma}_{i|j,k}$ be the quantities defined in Equation B.6 for $i, j, k = 1, \ldots, d$. Then, when $n = \Theta(\frac{1}{t^2}\log\frac{d}{\delta})$, we have, for all $i, j, k = 1, \ldots, d$,

$$|\widehat{\gamma}_{ij} - \gamma_{ij}| < t\frac{\sigma_{j|i,k}}{\sigma_{i|k}}, \qquad \left|\widehat{\sigma}_{i|j}^2 - \sigma_{i|j}^2\right| < t\sigma_{i|j}^2 \qquad \text{and} \qquad \left|\widehat{\sigma}_{i|j,k}^2 - \sigma_{i|j,k}^2\right| < t\sigma_{i|j,k}^2$$

with probability $1 - \delta$.

Proof. By using Corollary B.2.1 and the definition in Equation B.6, it can be done by a straightforward calculation.

□

Theorem B.2.2 (Conditional Mutual Information Tester)**.** *Suppose* $(X_1, \ldots, X_d)$ *is the random variable of* $\mathbb{N}(0, \Sigma)$ *for some positive definite* Σ*. Let* $(X_1^{(1)}, \ldots, X_d^{(1)}), \ldots, (X_1^{(n)}, \ldots, X_d^{(n)})$ *be the i.i.d. samples of* $(X_1, \ldots, X_d)$ *For any sufficiently small* $\varepsilon, \delta > 0$*, if*

$$n = \Theta(\frac{1}{\varepsilon} \log \frac{d}{\delta}),$$

the following results hold for all $i, j, k = 1, \ldots, d$ *with probability* $1 - \delta$*:*

1. *If* $I(X_i; X_j \mid X_k) = 0$*, then* $\widehat{I}(X_i; X_j \mid X_k) \leq \frac{\varepsilon}{100}$.
2. *If* $I(X_i; X_j \mid X_k) \geq \varepsilon$*, then* $\widehat{I}(X_i; X_j \mid X_k) > \frac{1}{20} I(X_i; X_j \mid X_k) - \frac{\varepsilon}{40}$.

Combining these two cases, we have

$$\widehat{I}(X_i; X_j \mid X_k) > \frac{1}{20} I(X_i; X_j \mid X_k) - \frac{\varepsilon}{40}.$$

Proof. By Lemma 22, with $\Theta(\frac{1}{\varepsilon} \log \frac{d}{\delta})$, we have the following properties for all $i, j, k = 1, \ldots, d$ with probability $1 - \delta$:

$$|\widehat{\gamma}_{ij} - \gamma_{ij}| < \frac{\sqrt{\varepsilon}}{100} \frac{\sigma_{j|i,k}}{\sigma_{i|k}}, \qquad \left|\widehat{\sigma}^2_{i|j} - \sigma^2_{i|j}\right| < \frac{\sqrt{\varepsilon}}{100} \sigma^2_{i|j} \qquad \text{and} \qquad \left|\widehat{\sigma}^2_{i|j,k} - \sigma^2_{i|j,k}\right| < \frac{\sqrt{\varepsilon}}{100} \sigma^2_{i|j,k} \tag{B.9}$$

We express

$$\widehat{I}(X_i; X_j \mid X_k) = \frac{1}{2} \log \left(1 + \widehat{\gamma}^2_{ij} \frac{\widehat{\sigma}^2_{i|k}}{\widehat{\sigma}^2_{j|i,k}}\right) = \frac{1}{2} \log \left(1 + \widehat{\gamma}^2_{ij} \frac{\sigma^2_{i|k}}{\sigma^2_{j|i,k}} \cdot \frac{\widehat{\sigma}^2_{i|k}}{\sigma^2_{i|k}} \cdot \frac{\sigma^2_{j|i,k}}{\widehat{\sigma}^2_{j|i,k}}\right) \tag{B.10}$$

We bound each term $\widehat{\gamma}^2_{ij} \frac{\sigma^2_{i|k}}{\sigma^2_{j|i,k}}$, $\frac{\widehat{\sigma}^2_{i|k}}{\sigma^2_{i|k}}$ and $\frac{\sigma^2_{j|i,k}}{\widehat{\sigma}^2_{j|i,k}}$ for the cases of $I(X_i; X_j \mid X_k) = 0$ and $I(X_i; X_j \mid X_k) \geq \varepsilon$.

We first prove if $I(X_i; X_j \mid X_k) = 0$ then $\widehat{I}(X_i; X_j \mid X_k) \leq \frac{\varepsilon}{100}$. Since $I(X_i; X_j \mid X_k) = 0$, it means that X_i and X_j are independent conditioned on X_k and hence $\gamma_{ij} = 0$. We have $\widehat{\gamma}^2_{ij} \frac{\sigma^2_{i|k}}{\sigma^2_{j|i,k}} \leq \frac{\varepsilon}{100}$. For the term $\frac{\widehat{\sigma}^2_{i|k}}{\sigma^2_{i|k}}$, we have $\frac{\widehat{\sigma}^2_{i|k}}{\sigma^2_{i|k}} \leq 1 + \frac{\sqrt{\varepsilon}}{100}$ by Equation B.9. For the term $\frac{\sigma^2_{j|i,k}}{\widehat{\sigma}^2_{j|i,k}}$, we have $\frac{\sigma^2_{j|i,k}}{\widehat{\sigma}^2_{j|i,k}} \leq \frac{1}{1 - \frac{\sqrt{\varepsilon}}{100}}$ by Equation B.9. Plugging these three inequalities into Equation B.10, we have

$$\widehat{I}(X_i; X_j \mid X_k) = \frac{1}{2} \log \left(1 + \widehat{\gamma}^2_{ij} \frac{\sigma^2_{i|k}}{\sigma^2_{j|i,k}} \cdot \frac{\widehat{\sigma}^2_{i|k}}{\sigma^2_{i|k}} \cdot \frac{\sigma^2_{j|i,k}}{\widehat{\sigma}^2_{j|i,k}}\right) \leq \frac{1}{2} \log \left(1 + \frac{\varepsilon}{100} \cdot \frac{1 + \frac{\sqrt{\varepsilon}}{100}}{1 - \frac{\sqrt{\varepsilon}}{100}}\right) \leq \frac{\varepsilon}{100}$$

for any sufficiently small $\varepsilon > 0$.

We now prove if $I(X_i; X_j \mid X_k) \geq \varepsilon$, then $\widehat{I}(X_i; X_j \mid X_k) > \frac{1}{20} I(X_i; X_j \mid X_k) - \frac{\varepsilon}{40}$. Since $I(X_i; X_j \mid X_k) \geq \varepsilon$, it means that $I(X_i; X_j \mid X_k) = \frac{1}{2}\log(1 + \gamma_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2}) \geq \varepsilon$ and hence $\gamma_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2} \geq e^{2\varepsilon} - 1 \geq 2\varepsilon$. We have $\widehat{\gamma}_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2} \geq \gamma_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2} - \sqrt{\frac{\varepsilon}{100}} \geq 0$. For the term $\frac{\widehat{\sigma}_{i|k}^2}{\sigma_{i|k}^2}$, we have $\frac{\widehat{\sigma}_{i|k}^2}{\sigma_{i|k}^2} \geq 1 - \frac{\sqrt{\varepsilon}}{100}$ by Equation B.9. For the term $\frac{\sigma_{j|i,k}^2}{\widehat{\sigma}_{j|i,k}^2}$, we have $\frac{\sigma_{j|i,k}^2}{\widehat{\sigma}_{j|i,k}^2} \geq \frac{1}{1+\frac{\sqrt{\varepsilon}}{100}}$ by Equation B.9. Plugging these three inequalities into Equation B.10, we have

$$\widehat{I}(X_i; X_j \mid X_k) = \frac{1}{2}\log\left(1 + \widehat{\gamma}_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2} \cdot \frac{\widehat{\sigma}_{i|k}^2}{\sigma_{i|k}^2} \cdot \frac{\sigma_{j|i,k}^2}{\widehat{\sigma}_{j|i,k}^2}\right) \geq \frac{1}{2}\log\left(1 + \left(\gamma_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2} - \sqrt{\frac{\varepsilon}{100}}\right)^2 \cdot \frac{1 - \frac{\sqrt{\varepsilon}}{100}}{1 + \frac{\sqrt{\varepsilon}}{100}}\right).$$

Note that, for any a, b, we have $(a-b)^2 \geq \frac{1}{2}a^2 - b^2$ which implies the term $\left(\gamma_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2} - \sqrt{\frac{\varepsilon}{100}}\right)^2$ is larger than $\frac{1}{2}\gamma_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2} - \frac{\varepsilon}{100}$. Namely, we have

$$\begin{aligned}
\widehat{I}(X_i; X_j \mid X_k) &\geq \frac{1}{2}\log\left(1 + \left(\frac{1}{2}\gamma_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2} - \frac{\varepsilon}{100}\right) \cdot \frac{1 - \frac{\sqrt{\varepsilon}}{100}}{1 + \frac{\sqrt{\varepsilon}}{100}}\right) \\
&\geq \frac{1}{2}\log\left(1 + \frac{1}{3}\gamma_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2} - \frac{\varepsilon}{100}\right) \\
&\geq \frac{1}{2}\log\left(1 + \frac{1}{3}\gamma_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2}\right) - \frac{\varepsilon}{40}
\end{aligned}$$

for any sufficiently small $\varepsilon > 0$. Note that, for any $a > 0$, $\log(1 + \frac{1}{3}a) \geq \frac{1}{10}\log(1+a)$. Namely, we have

$$\begin{aligned}
\widehat{I}(X_i; X_j \mid X_k) &\geq \frac{1}{2}\log\left(1 + \frac{1}{3}\gamma_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2}\right) - \frac{\varepsilon}{40} \geq \frac{1}{20}\log\left(1 + \gamma_{ij}^2 \frac{\sigma_{i|k}^2}{\sigma_{j|i,k}^2}\right) - \frac{\varepsilon}{40} \\
&= \frac{1}{20} I(X_i; X_j \mid X_k) - \frac{\varepsilon}{40}.
\end{aligned}$$

□

B.2.3 Distribution Learning Upper Bounds

In this subsection, we give the formal proof of the upper bounds on the sample complexity for distribution learning in the non-realizable setting Theorem 4.2.1 and realizable setting Theorem 5.3.3:

B.2.3.1 Non-realizable Case

Theorem 4.2.1. *Let P be a Gaussian distribution. Given n i.i.d. samples from P, for any $\varepsilon, \delta > 0$, if $n \gtrsim \frac{d^2}{\varepsilon^2} \log \frac{d}{\delta}$, then $\widehat{T}$ returned by Algorithm 5 satisfies*

$$D_{\mathrm{KL}}(P \parallel P_{\widehat{T}}) \leq \min_{T \in \mathcal{T}} D_{\mathrm{KL}}(P \parallel P_T) + \varepsilon,$$

with probability at least $1 - \delta$.

Proof. Let T^* be $\arg\min_{T \in \mathcal{T}} D_{\mathrm{KL}}(P \parallel P_T)$. By Equation 4.4, we express $D_{\mathrm{KL}}(P \parallel P_{\widehat{T}}) - D_{\mathrm{KL}}(P \parallel P_{T^*})$ as

$$D_{\mathrm{KL}}(P \parallel P_{\widehat{T}}) - D_{\mathrm{KL}}(P \parallel P_{T^*}) = -\sum_{(W,Z) \in \widehat{T}} I(W;Z) + \sum_{(X,Y) \in T^*} I(X;Y)$$

Since $\widehat{T}$ is the output of Algorithm 5, we have

$$\sum_{(X,Y) \in T^*} \widehat{I}(X;Y) - \sum_{(W,Z) \in \widehat{T}} \widehat{I}(W;Z) \leq 0.$$

Hence, we have

$$\begin{aligned} &D_{\mathrm{KL}}(P \parallel P_{\widehat{T}}) - D_{\mathrm{KL}}(P \parallel P_{T^*}) \\ &\quad \leq \sum_{(W,Z) \in \widehat{T}} \widehat{I}(W;Z) - \sum_{(W,Z) \in \widehat{T}} I(W;Z) + \sum_{(X,Y) \in T^*} I(X;Y) - \sum_{(X,Y) \in T^*} \widehat{I}(X;Y) \end{aligned}$$

By the definition in Equation B.5 and Corollary B.2.1, we can show that each $\left|\widehat{I}(X,Y) - I(X,Y)\right| < \frac{\varepsilon}{d}$ for all (X,Y) using $O(\frac{d^2}{\varepsilon^2} \log \frac{d}{\delta})$ samples. Therefore, we have

$$D_{\mathrm{KL}}(P \parallel P_{\widehat{T}}) - D_{\mathrm{KL}}(P \parallel P_{T^*}) < \varepsilon.$$

□

B.2.3.2 Realizable Case

Fact B.2.3 ([Bha+21])**.** *Let T_1 and T_2 be two spanning trees on d vertices such that their symmetric difference consists of the edges $E = \{e_1, e_2, \ldots, e_l\} \in T_1 \setminus T_2$ and $F = \{f_1, f_2, \ldots, f_l\} \in T_2 \setminus T_1$. Then E and F can be paired up, say $\langle e_i, f_i \rangle$, such that for all i, $T_1 \cup \{f_i\} \setminus \{e_i\}$ is a spanning tree.*

Theorem 4.2.2. *Let T^* be a directed tree and P_{T^*} be a T^*-structured Gaussian. Given n i.i.d. samples from P_{T^*}, for any $\varepsilon, \delta > 0$, if $n \gtrsim \frac{d}{\varepsilon} \log \frac{d}{\delta}$, then $\widehat{T}$ returned by Algorithm 5 satisfies*

$$D_{\mathrm{KL}}(P_{T^*} \parallel P_{\widehat{T}}) \leq \varepsilon,$$

with probability at least $1 - \delta$.

Proof. We first consider the edge difference between $\widehat{T}$ and T^*. By fact C.1.1, we can pair up the edges in $\widehat{T} \backslash T^*$ with the edges in $T^* \backslash \widehat{T}$ such that $T^* \cup \{(W, Z)\} \backslash \{(X, Y)\}$ is also a spanning tree for any $(W, Z) \in \widehat{T} \backslash T^*$ and $(X, Y) \in T^* \backslash \widehat{T}$. Let $\widehat{T} \backslash T^*$ be $\{(W_1, Z_1), \ldots, (W_k, Z_k)\}$ and $T^* \backslash \widehat{T}$ be $\{(X_1, Y_1), \ldots, (X_k, Y_k)\}$ such that (W_i, Z_i) pairs up with (X_i, Y_i) for $i = 1, \ldots, k$. Because of that, there exists a path in T^* from W_i to Z_i containing X_i and Y_i. Without loss of generality, we assume that the order of them is $W_i \rightsquigarrow X_i - Y_i \rightsquigarrow Z_i$ in T^*.

Since $\widehat{T}$ is the output of Algorithm 5, we have

$$\sum_{i=1}^{k} \widehat{I}(X_i; Y_i) - \sum_{i=1}^{k} \widehat{I}(W_i; Z_i) \leq 0$$

by the definition of the maximal spanning tree. We first expand the LHS as

$$\begin{aligned}
\sum_{i=1}^{k} \widehat{I}(X_i, Y_i) - \sum_{i=1}^{k} \widehat{I}(W_i, Z_i) &= \sum_{i=1}^{k} \Big(\widehat{I}(X_i, Y_i) - \widehat{I}(X_i; Z_i) + \widehat{I}(X_i; Z_i) - \widehat{I}(W_i; Z_i) \Big) \\
&= \sum_{i=1}^{k} \Big(\widehat{I}(X_i; Y_i \mid Z_i) - \widehat{I}(X_i; Z_i \mid Y_i) + \widehat{I}(X_i; Z_i \mid W_i) - \widehat{I}(W_i; Z_i \mid X_i) \Big) \qquad \text{by Equation B.8} \\
&= \underbrace{\sum_{i=1}^{k} \Big(\widehat{I}(X_i; Y_i \mid Z_i) + \widehat{I}(X_i; Z_i \mid W_i) \Big)}_{:=A} - \underbrace{\sum_{i=1}^{k} \Big(\widehat{I}(X_i; Z_i \mid Y_i) + \widehat{I}(W_i; Z_i \mid X_i) \Big)}_{:=B}.
\end{aligned}$$

In other words, we have $A \leq B$.

Recall that there exists a path $W_i \rightsquigarrow X_i - Y_i \rightsquigarrow Z_i$ in T^* and hence $(X_i, Z_i) \notin T^*$ which further implies $I(X_i; Z_i \mid Y_i) = 0$. Similarly, we have $I(W_i; Z_i \mid X_i) = 0$. By Theorem B.2.2 with $\Theta(\frac{1}{\varepsilon'} \log \frac{d}{\delta})$ samples, we have

$$\widehat{I}(X_i; Z_i \mid Y_i) \leq \varepsilon'/100 \qquad \text{and} \qquad \widehat{I}(W_i; Z_i \mid X_i) \leq \varepsilon'/100 \qquad \text{for all } i = 1, \ldots, k.$$

Plugging them into each term in B, we can bound B by $2k \cdot \varepsilon'/100 \leq d\varepsilon'/50$. Namely, we have

$$A = \sum_{i=1}^{k} \Big(\widehat{I}(X_i; Y_i \mid Z_i) + \widehat{I}(X_i; Z_i \mid W_i) \Big) \leq d\varepsilon'/50.$$

By Theorem B.2.2 with $\Theta(\frac{1}{\varepsilon'}\log\frac{d}{\delta})$ samples, we have

$$\frac{1}{20}I(X_i;Y_i \mid Z_i) - \frac{\varepsilon'}{40} \le \widehat{I}(X_i;Y_i \mid Z_i) \qquad \text{and} \qquad \frac{1}{20}I(X_i;Z_i \mid W_i) - \frac{\varepsilon'}{40} \le \widehat{I}(X_i;Z_i \mid W_i)$$

for all $i = 1, \ldots, k$. In other words,

$$A \ge \frac{1}{20}\sum_{i=1}^{k}\left(I(X_i;Y_i \mid Z_i) + I(X_i;Z_i \mid W_i)\right) - \frac{d\varepsilon'}{40}$$

or

$$\sum_{i=1}^{k}\left(I(X_i;Y_i \mid Z_i) + I(X_i;Z_i \mid W_i)\right) \le \frac{9d\varepsilon'}{10} \tag{B.11}$$

Now, we can bound $D_{\mathrm{KL}}(P_{T^*} \parallel P_{\widehat{T}})$. We express it as

$$\begin{aligned} D_{\mathrm{KL}}(P_{T^*} \parallel P_{\widehat{T}}) &= \sum_{i=1}^{k} I(X_i;Y_i) - \sum_{i=1}^{k} I(W_i;Z_i) = \sum_{i=1}^{k}\left(I(X_i;Y_i) - I(X_i;Z_i) + I(X_i;Z_i) - I(W_i;Z_i)\right) \\ &= \sum_{i=1}^{k}\left(I(X_i;Y_i \mid Z_i) - I(X_i;Z_i \mid Y_i) + I(X_i;Z_i \mid W_i) - I(W_i;Z_i \mid X_i)\right) \end{aligned}$$

Recall that we have $I(X_i;Z_i \mid Y_i) = 0$ and $I(W_i;Z_i \mid X_i) = 0$. Combining with Equation B.11, we have

$$D_{\mathrm{KL}}(P_{T^*} \parallel P_{\widehat{T}}) \le \frac{9d\varepsilon'}{10}$$

with probability at least $1-\delta$. By picking $\varepsilon' = \frac{10\varepsilon}{9d}$, we conclude our result. □

B.2.4 Distribution Learning Lower Bounds

To show the lower bounds, our main idea is to reduce problem B.2.4 defined below to our problem.

Problem B.2.4. Suppose $R^{(1)}$ and $R^{(2)}$ are two distributions such that $D_{\mathrm{KL}}(R^{(1)} \parallel R^{(2)}) \le \delta$. Let P be a distribution on m variables where each variable is distributed as either $R^{(1)}$ or $R^{(2)}$ uniformly and independently. We are given n i.i.d. samples drawn from a distribution P. Our task is to determine which distribution the samples are drawn from correctly for at least $51m/100$ variables. Formally, we define

$$\mathcal{R} := \{(R_1, \ldots, R_m) \mid R_i \in \{R^{(1)}, R^{(2)}\}\}.$$

We pick a distribution uniformly from $\mathcal{R}$ and let $P = (R_1^*, \ldots, R_m^*)$ be this distribution. Then, our goal is to design an algorithm that takes n i.i.d. samples drawn from P as input and returns $(\widehat{R}_1, \ldots, \widehat{R}_m)$ such that $\widehat{R}_i = R_i^*$ for at least $51m/100$ of $\{1, \ldots, m\}$.

Fact B.2.5. *By the standard information-theoretic lower bounds, if $n = o(\frac{1}{\delta})$, then no algorithm can solve problem B.2.4 with probability $2/3$.*

B.2.4.1 Non-realizable Case

We define two distributions $Q^{(1)}, Q^{(2)}$ as follows.

$$Q^{(1)} = \begin{cases} H \sim \mathbb{N}(0,1) \\ X \sim (1+\varepsilon)H + \mathbb{N}(0,1) \\ Y \sim H + \mathbb{N}(0,1) \\ Z \sim H + \mathbb{N}(0,1) \end{cases} \quad \text{and} \quad Q^{(2)} = \begin{cases} H \sim \mathbb{N}(0,1) \\ X \sim H + \mathbb{N}(0,1) \\ Y \sim (1+\varepsilon)H + \mathbb{N}(0,1) \\ Z \sim H + \mathbb{N}(0,1) \end{cases} \tag{B.12}$$

Also, we define $R^{(1)}, R^{(2)}$ to be the corresponding marginal distributions on (X, Y, Z).

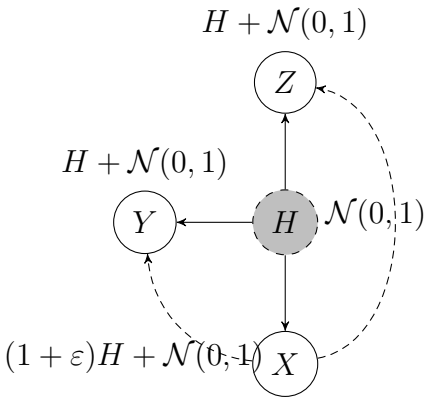


Figure B.5: $R^{(1)}$

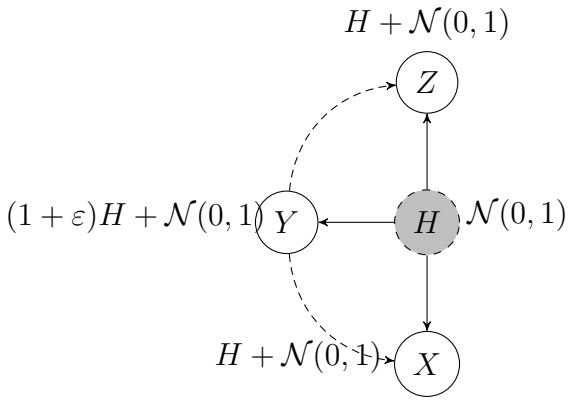


Figure B.6: $R^{(2)}$

Figure B.7: The $\Omega(1/\varepsilon^2)$ bound in the non-realizable setting. The underlying graph is represented with solid lines, while the best estimated tree structure is depicted with dashed lines.

Lemma 23. *Suppose R^* is one of $R^{(1)}$ and $R^{(2)}$ defined in Equation B.12. For any small $\varepsilon > 0$, if a direct tree $\widehat{T}$ satisfies*

$$D_{\mathrm{KL}}(R^* \parallel R^*_{\widehat{T}}) \leq \min_T D_{\mathrm{KL}}(R^* \parallel R^*_T) + \frac{\varepsilon}{100} \tag{B.13}$$

and $\widehat{R} = \arg\min_{R \in \{R^{(1)}, R^{(2)}\}} D_{\mathrm{KL}}(R \parallel R_{\widehat{T}})$*, then* $\widehat{R} = R^*$.

Proof. Since there are three variables, there are only three possible tree structures: $T_1 = Y - X - Z$, $T_2 = X - Y - Z$ and $T_3 = X - Z - Y$. Recall that, by Equation 4.4, we have

$$D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_2}) - D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_1}) = I(X;Z) - I(Y;Z) \geq \frac{\varepsilon}{50} \tag{B.14}$$

and, similarly, we also have

$$D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_3}) - D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_1}) \geq \frac{\varepsilon}{50} \tag{B.15}$$

$$D_{\mathrm{KL}}(R^{(2)} \parallel R^{(2)}_{T_1}) - D_{\mathrm{KL}}(R^{(2)} \parallel R^{(2)}_{T_2}) \geq \frac{\varepsilon}{50} \tag{B.16}$$

$$D_{\mathrm{KL}}(R^{(2)} \parallel R^{(2)}_{T_3}) - D_{\mathrm{KL}}(R^{(2)} \parallel R^{(2)}_{T_2}) \geq \frac{\varepsilon}{50} \tag{B.17}$$

By Equation B.13, Equation B.15 and Equation B.17, we have $\widehat{T} \neq T_3$. Namely, $\widehat{T}$ is either T_1 or T_2 (WLOG, say T_1). By Equation B.13 and Equation B.16, we have $R^* = R^{(1)}$. By Equation B.14, we have

$$D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_1}) \leq D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_2}) - \frac{\varepsilon}{50} < \underbrace{D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_2}) = D_{\mathrm{KL}}(R^{(2)} \parallel R^{(2)}_{T_1})}_{\text{by symmetry}}.$$

Hence, $\widehat{R} = R^{(1)} = R^*$ by the definition of $\widehat{R}$.

□

Theorem 4.2.3. *Suppose P is an unknown Gaussian distribution. Given n i.i.d. samples drawn from P. For any small $\varepsilon > 0$, if $n = o(d^2/\varepsilon^2)$, then for any estimator $\widehat{T}$, the maximum probability of achieving the required accuracy over a hard family of distribution $\mathcal{P}$ is bounded, such that:*

$$\inf_{\widehat{T}} \sup_{P \in \mathcal{P}} \Pr\left(D_{\mathrm{KL}}(P \parallel P_{\widehat{T}}) \leq \min_{T \in \mathcal{T}} D_{\mathrm{KL}}(P \parallel P_T) + \varepsilon \right) \leq 2/3$$

Proof. We will prove the statement by reducing problem B.2.4 to our problem. We first split the d variables into $m = d/3$ groups of 3 variables and for each group we select $R^{(1)}$ or $R^{(2)}$ defined in Equation B.12 (replacing ε with ε/d) uniformly and independently and notice that $D_{\mathrm{KL}}(R^{(1)} \parallel R^{(2)}) = O(\varepsilon^2/d^2)$ by a straightforward calculation. By fact B.2.5, it implies that if $n = o(\frac{d^2}{\varepsilon^2})$ then no algorithm can determine which distribution the samples are drawn from correctly for at least $51m/100$ groups with probability $\frac{2}{3}$.

Suppose there is an algorithm that takes these n i.i.d. samples as input and returns a directed tree $\widehat{T}$ such that

$$D_{\mathrm{KL}}(P \parallel P_{\widehat{T}}) \leq \min_{T \in \mathcal{T}} D_{\mathrm{KL}}(P \parallel P_T) + \varepsilon \tag{B.18}$$

with probability $\frac{2}{3}$. If we manage to show that we can use $\widehat{T}$ to determine which distribution the samples are drawn from correctly for $51m/100$ groups then it implies $n = \Omega(\frac{d^2}{\varepsilon^2})$.

We construct the reduction as follows. For the i-th group of variables, we consider its subtree $\widehat{T}_i$ of $\widehat{T}$ and declare $\widehat{R}_i$ to be the distribution for this group where $\widehat{R}_i$ is defined to be $\arg\min_{R \in \{R^{(1)}, R^{(2)}\}} D_{\mathrm{KL}}(R \parallel R_{\widehat{T}_i})$. To see the correctness, we have the following. Since each group is independent, Equation B.18 can be decomposed into

$$\sum_{i=1}^{m} D_{\mathrm{KL}}(P_i \parallel (P_i)_{\widehat{T}_i}) \leq \sum_{i=1}^{m} \min_{T_i} D_{\mathrm{KL}}(P_i \parallel (P_i)_{T_i}) + \varepsilon$$

where P_i is the random pick of $R^{(1)}$ or $R^{(2)}$ for the i-th group. Therefore, at least $51m/100$ of the terms $D_{\mathrm{KL}}(P_i \parallel (P_i)_{\widehat{T}_i}) - \min_{T_i} D_{\mathrm{KL}}(P_i \parallel (P_i)_{T_i}) \leq \frac{10\varepsilon}{m}$. By Lemma 23, for these $51m/100$ groups, $\widehat{R}_i$ is correctly determined, i.e. $\widehat{R}_i = P_i$ and hence the reduction is completed. □

B.2.4.2 Realizable Case

We define two distributions $R^{(1)}, R^{(2)}$ as follows.

$$R^{(1)} = \begin{cases} X \sim \mathbb{N}(0,1) \\ Y \sim (1-\sqrt{\varepsilon})X + \sqrt{\varepsilon}\mathbb{N}(0,1) \\ Z \sim \frac{1}{2}X + \frac{1}{2}\mathbb{N}(0,1) \end{cases} \quad \text{and} \quad R^{(2)} = \begin{cases} X \sim \mathbb{N}(0,1) \\ Y \sim (1-\sqrt{\varepsilon})X + \sqrt{\varepsilon}\mathbb{N}(0,1) \\ Z \sim \frac{1}{2}Y + \frac{1}{2}\mathbb{N}(0,1) \end{cases} \tag{B.19}$$

Namely, the underlying graph for $R^{(1)}$ is $Y < -X- > Z$ and the underlying graph for $R^{(2)}$ is $X- > Y- > Z$. Both have $X- > Y$ and the only difference is Z.

Lemma 24. *Suppose R^* is one of $R^{(1)}$ and $R^{(2)}$ defined in Equation B.19. For any small $\varepsilon > 0$, if a direct tree $\widehat{T}$ satisfies*

$$D_{\mathrm{KL}}(R^* \parallel R^*_{\widehat{T}}) \leq \frac{\varepsilon}{100} \tag{B.20}$$

and $\widehat{R} = \arg\min_{R \in \{R^{(1)}, R^{(2)}\}} D_{\mathrm{KL}}(R \parallel R_{\widehat{T}})$, then $\widehat{R} = R^*$.

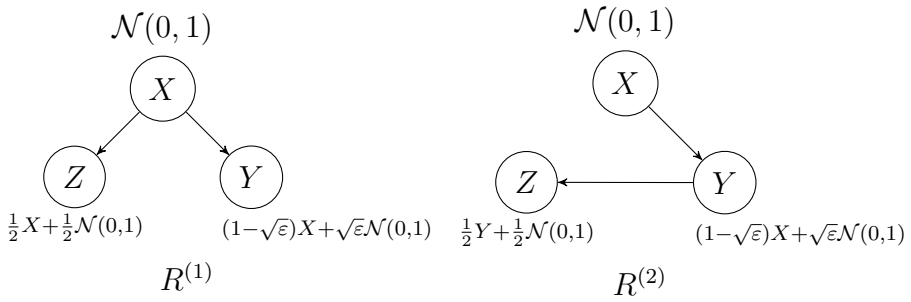


Figure B.8: Realizable setting. The two graphs represent the hard family of distributions $\mathcal{P}$ used for the lower bound proof in Theorem 4.2.4.

Proof. Since there are three variables, there are only three possible tree structures: $T_1 = Y - X - Z$, $T_2 = X - Y - Z$ and $T_3 = X - Z - Y$. Recall that, by Equation 4.4, we have

$$D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_2}) - D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_1}) = I(X;Z) - I(Y;Z) \geq \frac{\varepsilon}{50}$$

Note that $D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_1}) = 0$ and hence

$$D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_2}) \geq \frac{\varepsilon}{50} \tag{B.21}$$

Similarly, we also have

$$D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{T_3}) \geq \Omega(1) \geq \frac{\varepsilon}{50} \tag{B.22}$$

$$D_{\mathrm{KL}}(R^{(2)} \parallel R^{(2)}_{T_1}) \geq \frac{\varepsilon}{50} \tag{B.23}$$

$$D_{\mathrm{KL}}(R^{(2)} \parallel R^{(2)}_{T_3}) \geq \Omega(1) \geq \frac{\varepsilon}{50} \tag{B.24}$$

By Equation B.20, Equation B.22 and Equation B.24, we have $\widehat{T} \neq T_3$. Namely, $\widehat{T}$ is either T_1 or T_2. If $\widehat{T} = T_1$, by Equation B.20 and Equation B.23, we have

$$D_{\mathrm{KL}}(R^{(2)} \parallel R^{(2)}_{\widehat{T}}) > D_{\mathrm{KL}}(R^{*} \parallel R^{*}_{\widehat{T}})$$

and hence $R^* = R^{(1)}$. If $\widehat{T} = T_2$, by Equation B.20 and Equation B.21, we have

$$D_{\mathrm{KL}}(R^{(1)} \parallel R^{(1)}_{\widehat{T}}) > D_{\mathrm{KL}}(R^{*} \parallel R^{*}_{\widehat{T}})$$

and hence $R^* = R^{(2)}$. By the definition of $\widehat{R}$, both cases imply $\widehat{R} = R^*$.

□

Theorem 4.2.4. *Suppose P is an unknown Gaussian distribution such that there exists a directed tree T^* that P is T^*-structured, i.e. $P = P_{T^*}$. Given n i.i.d. samples drawn from P. For any small $\varepsilon > 0$, if $n = o(d/\varepsilon)$, then for any estimator $\widehat{T}$, the maximum probability of success over a hard family of distribution $\mathcal{P}$ is bounded, such that*

$$\inf_{\widehat{T}} \sup_{P \in \mathcal{P}} \Pr\left(D_{\mathrm{KL}}(P \parallel P_{\widehat{T}}) \leq \varepsilon\right) \leq 2/3$$

Proof. We will prove the statement by reducing problem B.2.4 to our problem. We first split the d variables into $m = d/3$ groups of 3 variables and for each group we select $R^{(1)}$ or $R^{(2)}$ defined in Equation B.19 (replacing ε with ε/d) uniformly and independently and notice that $D_{\mathrm{KL}}(R^{(1)} \parallel R^{(2)}) = O(\varepsilon/d)$ by a straightforward calculation. By fact B.2.5, it implies that if $n = o(\frac{d}{\varepsilon})$ then no algorithm can determine which distribution the samples are drawn from correctly for at least $51m/100$ groups with probability $\frac{2}{3}$.

Suppose there is an algorithm that takes these n i.i.d. samples as input and returns a directed tree $\widehat{T}$ such that

$$D_{\mathrm{KL}}(P \parallel P_{\widehat{T}}) \leq \varepsilon \tag{B.25}$$

with probability $\frac{2}{3}$. If we manage to show that we can use $\widehat{T}$ to determine which distribution the samples are drawn from correctly for $51m/100$ groups then it implies $n = \Omega(\frac{d}{\varepsilon})$.

We construct the reduction as follows. For the i-th group of variables, we consider its subtree $\widehat{T}_i$ of $\widehat{T}$ and declare $\widehat{R}_i$ to be the distribution for this group where $\widehat{R}_i$ is defined to be $\arg\min_{R \in \{R^{(1)}, R^{(2)}\}} D_{\mathrm{KL}}(R \parallel R_{\widehat{T}_i})$. To see the correctness, we have the following. Since each group is independent, Equation B.25 can be decomposed into

$$\sum_{i=1}^{m} D_{\mathrm{KL}}(P_i \parallel (P_i)_{\widehat{T}_i}) \leq \varepsilon$$

where P_i is the random pick of $R^{(1)}$ or $R^{(2)}$ for the i-th group. Therefore, at least $51m/100$ of the terms $D_{\mathrm{KL}}(P_i \parallel (P_i)_{\widehat{T}_i}) \leq \frac{10\varepsilon}{m}$. By Lemma 24, for these $51m/100$ groups, $\widehat{R}_i$ is correctly determined, i.e. $\widehat{R}_i = P_i$ and hence the reduction is completed.

□

B.2.5 Learning Polytrees given Skeleton

In this section, we sketch how to obtain a sample-efficient algorithm for learning bounded-degree gaussian *polytrees* by adapting the recent results from [Cho+23], using the guarantees of the estimator $\widehat{I}$, assuming that the skeleton is known. Let a *m-polytree* denote a polytree with maximum in-degree m. Our main result in this section is the following:

Theorem B.2.6. *There exists an algorithm which, given n samples from a gaussian m-polytree P over $\mathbb{R}^d$, accuracy parameter $\varepsilon > 0$, failure probability δ, maximum in-degree m, and the explicit description of the ground truth skeleton of P, outputs a m-polytree $\widehat{P}$ such that $D_{\mathrm{KL}}(P\|\widehat{P}) \leq \varepsilon$ with success probability at least $1-\delta$, as long as:*

$$n \geq \widetilde{O}\left(\frac{d}{\varepsilon}\log\frac{1}{\delta}\right).$$

Moreover, the algorithm runs in time polynomial in n and d.

Note that the guarantee in Theorem B.2.6 is entirely independent of any faithfulness parameter, in contrast to Theorem 4.3.3. The algorithm and its analysis is exactly the same as in [Cho+23], with the only change being that we use Equation B.5 for the estimator $\widehat{I}$.

B.3 Proofs of Section 4.3

B.3.1 Sample Conditional Correlation Coefficient as CI Tester

`PC-Tree` relies on sample (conditional) correlation coefficient as (conditional) independence tester. Specifically, denote the sample covariance matrix to be $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X^{(i)}X^{(i)^\top}$, for any two nodes $j,k \in V$ and any subset $S \subseteq V \setminus \{j,k\}$, which could be $\emptyset$, the sample correlation coefficient is defined by

$$\widehat{\rho}_{jk\,|\,S} := \frac{\widehat{\Sigma}_{jk} - \widehat{\Sigma}_{jS}\widehat{\Sigma}_{SS}^{-1}\widehat{\Sigma}_{Sk}}{\sqrt{(\widehat{\Sigma}_{jj} - \widehat{\Sigma}_{jS}\widehat{\Sigma}_{SS}^{-1}\widehat{\Sigma}_{Sj})(\widehat{\Sigma}_{kk} - \widehat{\Sigma}_{kS}\widehat{\Sigma}_{SS}^{-1}\widehat{\Sigma}_{Sk})}}$$

Algorithm 14: ORIENT algorithm

Input: Skeleton $\widehat{T}$, separation sets S

Output: CPDAG $\widehat{\widehat{T}}$.

1. For all pairs of nonadjacent nodes j, k with common neighbour ℓ:
   a) If $\ell \notin S(j,k)$, then directize $j - \ell - k$ in $\widehat{T}$ by $j \to \ell \leftarrow k$
2. In the resulting PDAG $\widehat{T}$, orient as many as possible undirected edges by applying following rules:
   - **R1** Orient $k - \ell$ into $k \to \ell$ whenever there is an arrow $j \to k$ such that j and ℓ are not adjacent
   - **R2** Orient $j - k$ into $j \to k$ whenever there is a chain $j \to \ell \to k$
   - **R3** Orient $j - k$ into $j \to k$ whenever there are two chains $j - \ell \to k$ and $j - i \to k$ such that ℓ and i are not adjacent
   - **R4** Orient $j - k$ into $j \to k$ whenever there are two chains $j - \ell \to i$ and $\ell - i \to k$ such that ℓ and i are not adjacent
3. Return $\widehat{T}$ as $\widehat{\widehat{T}}$.

Then the conditional independence tester for hypothesis $H_0 : X_j \perp\!\!\!\perp X_k \,|\, X_S$ is given by a cutoff on the sample correlation coefficient:

$$\text{Output} = \begin{cases} \text{accept } H_0 & \text{if } |\widehat{\rho}_{jk\,|\,S}| \geq c/2 \\ \text{reject } H_0 & \text{if } |\widehat{\rho}_{jk\,|\,S}| < c/2 \end{cases}. \tag{B.26}$$

Here the choice of $c/2$ is for theoretical purpose. Since correlation coefficient is normalized between $[-1, 1]$, in practice, the tester can be implemented by choosing a cutoff that is small enough, e.g. 0.05. The analysis of `PC-Tree` crucially relies on the following lemma on the estimation error of the sample (conditional) correlation coefficients:

Lemma 25. *Let $X \in \mathbb{R}^d \sim \mathcal{N}(0, \Sigma)$, for any $j, k \in V$ and any subset $S \subseteq V \setminus \{j, k\}$ with $|S| \leq q$, $\delta \in (0, 1)$, if $n \gtrsim q + 1/\delta^2$, then*

$$\Pr(|\widehat{\rho}_{jk\,|\,S} - \rho_{jk\,|\,S}| \geq \delta) \leq \exp(-C_0(n-q)\delta^2)\,,$$

for some universal constant $C_0 > 0$.

It is clear to see that as long as the (conditional) correlation coefficients are estimated accurately enough, the CI tests are correct due to c-strong Tree-faithfulness.

Lemma 25 is more general than needed to analyze `PC-Tree` algorithm. Since Lemma 25 reveals the dependence on the size of conditioning set S, while `PC-Tree` only requires $|S| \leq 1$.

B.3.2 Proof of Lemma 25

Lemma 25. *Let $X \in \mathbb{R}^d \sim \mathcal{N}(0, \Sigma)$, for any $j, k \in V$ and any subset $S \subseteq V \setminus \{j, k\}$ with $|S| \leq q$, $\delta \in (0, 1)$, if $n \gtrsim q + 1/\delta^2$, then*

$$\Pr(|\widehat{\rho}_{jk\,|\,S} - \rho_{jk\,|\,S}| \geq \delta) \leq \exp(-C_0(n-q)\delta^2)\,,$$

for some universal constant $C_0 > 0$.

Proof. The proof is a combination of the following lemmas. We start with analyzing sample marginal correlation of bivariate normal distribution, then extend to conditional correlation.

Lemma 26. *Let $W = (X, Y) \sim \mathcal{N}(0, \Sigma)$ where $\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} \in \mathbb{R}^{2\times 2}$, and $\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$. Let the sample covariance matrix and correlation be*

$$\frac{1}{n}\sum_{\ell=1}^{n} w^{(\ell)} w^{(\ell)\top} = \begin{pmatrix} \widehat{\sigma}_X^2 & \widehat{\sigma}_{XY} \\ \widehat{\sigma}_{XY} & \widehat{\sigma}_Y^2 \end{pmatrix}, \quad \text{and} \quad \widehat{\rho} = \frac{\widehat{\sigma}_{XY}}{\widehat{\sigma}_X \widehat{\sigma}_Y}\,.$$

For $\delta \in (0, 1)$, if $n \gtrsim 1/\delta^2$, then

$$\Pr(|\widehat{\rho} - \rho| \geq \delta) \leq \exp(-C_0 n \delta^2)\,,$$

for some constant $C_0 > 0$.

Now look at sample conditional correlation, suppose we want to estimate $\rho_{jk\,|\,S}$ with $|S| = q' \leq q$. Recall the sample covariance matrix is $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X^{(i)} X^{(i)\top}$. Denote $I = \{j, k\}$, then the estimator is given by 2×2 matrix

$$\widehat{\Sigma}_{II\,|\,S} := \widehat{\Sigma}_{II} - \widehat{\Sigma}_{II,S}\widehat{\Sigma}_{SS}^{-1}\widehat{\Sigma}_{S,II}\,.$$

We borrow a classic result regarding the distribution of $\widehat{\Sigma}_{II\,|\,S}$:

Lemma 27 ([And58], Theorem 4.3.4)**.** *The sample covariance matrix $\widehat{\Sigma}_{II\,|\,S}$ is distributed as $\frac{1}{n}\sum_{\ell=1}^{n-q'} u^{(\ell)} u^{(\ell)\top}$, where $\{u^{(\ell)}\}_{\ell=1}^{n-q'}$ are independently distributed according to $\mathcal{N}(0, \Sigma_{II\,|\,S})$.*

Then applying the bivariate result from Lemma 26 with covariance matrix $\Sigma_{II\,|\,S}$ and sample size $n - q' \leq n - q$ completes the proof. □

It remains to prove the lemma used in proof above.

Proof of Lemma 26. Let $Z_X = X/\sigma_X$, $Z_Y = Y/\sigma_Y$, then $Z_X, Z_Y \sim \mathcal{N}(0,1)$ and $\rho_{Z_X,Z_Y} = \rho = \mathrm{Cov}(Z_X, Z_Y) \in [-1,1]$. Denote the corresponding samples to be $z_X = (z_X^{(1)}, \ldots, z_X^{(n)})$ and $z_Y = (z_Y^{(1)}, \ldots, z_Y^{(n)})$, therefore

$$\widehat{\rho} = \frac{\widehat{\sigma}_{XY}}{\widehat{\sigma}_X \widehat{\sigma}_Y} = \frac{\widehat{\sigma}_{XY}/(\sigma_X\sigma_Y)}{(\widehat{\sigma}_X/\sigma_X) \times (\widehat{\sigma}_Y/\sigma_Y)} = \frac{\langle z_X, z_Y\rangle}{\|z_X\|\|z_Y\|}\,.$$

Then the deviation

$$\begin{aligned}
|\widehat{\rho} - \rho| &= \left| \frac{\langle z_X, z_Y\rangle}{\|z_X\|\|z_Y\|} - \mathrm{Cov}(Z_X, Z_Y) \right| \\
&\leq \left| \frac{\langle z_X, z_Y\rangle/n}{\|z_X\|\|z_Y\|/n} - \frac{\mathrm{Cov}(Z_X, Z_Y)}{\|z_X\|\|z_Y\|/n} + \frac{\mathrm{Cov}(Z_X, Z_Y)}{\|z_X\|\|z_Y\|/n} - \mathrm{Cov}(Z_X, Z_Y) \right| \\
&\leq \left| \frac{1}{\|z_X\|\|z_Y\|/n} - 1 \right| \left| \langle z_X, z_Y\rangle/n - \mathrm{Cov}(Z_X, Z_Y) \right| + \left| \langle z_X, z_Y\rangle/n - \mathrm{Cov}(Z_X, Z_Y) \right| \\
&\quad + \left| \mathrm{Cov}(Z_X, Z_Y) \right| \left| \frac{1}{\|z_X\|\|z_Y\|/n} - 1 \right|.
\end{aligned}$$

We apply the following lemma to bound the errors:

Lemma 28. *If* $(X,Y) \sim \mathcal{N}\Big(0, \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}\Big)$ *for* $|r| \leq 1$, *then the sample variance* $\widehat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^n {X^{(i)}}^2$, $\widehat{\sigma}_Y^2 = \frac{1}{n}\sum_{i=1}^n {Y^{(i)}}^2$ *and sample covariance* $\widehat{\sigma}_{XY} = \frac{1}{n}\sum_{i=1}^n X^{(i)}Y^{(i)}$ *have the following bounds: for* $\zeta < 1$, *if* $n \geq \frac{2048 \log 7}{\zeta^2}$, *then*

$$\begin{aligned}
&\Pr(|\widehat{\sigma}_X^2 - 1| \geq \zeta) \leq \exp(-n\zeta^2/16) \\
&\Pr(|\widehat{\sigma}_Y^2 - 1| \geq \zeta) \leq \exp(-n\zeta^2/16) \\
&\Pr(|\widehat{\sigma}_{XY} - r| \geq \zeta) \leq \exp(-n\zeta^2/2048)\,.
\end{aligned}$$

Using Lemma 28, with probability at least $1 - 3\exp(-n\zeta^2/2048)$, we have $|\|z_X\|^2/n - 1| \leq \zeta$, $|\|z_Y\|^2/n - 1| \leq \zeta$, $|\langle z_X, z_Y\rangle/n - \mathrm{Cov}(Z_X, Z_Y)| \leq \zeta$. Then

$$\left| \frac{1}{\|z_X\|\|z_Y\|/n} - 1 \right| = \frac{|\|z_X\|\|z_Y\|/n - 1|}{\|z_X\|\|z_Y\|/n} \leq \frac{\zeta}{1-\zeta}\,.$$

Choose $\zeta = \frac{\delta}{3+\delta}$, then $\left|\frac{1}{\|z_X\|\|z_Y\|/n} - 1\right| \le \delta/3$, $\left|\langle z_X, z_Y\rangle/n - \mathrm{Cov}(Z_X, Z_Y)\right| \le \delta/(3+\delta) \le \delta/3$. Lastly,

$$|\widehat{\rho} - \rho| \le \frac{\delta}{3} \times \frac{\delta}{3} + \frac{\delta}{3} + \frac{\delta}{3} \le \delta\,,$$

with probability at least

$$\begin{aligned} 1 - 3\exp(-n\zeta^2/2048) &= 1 - \exp\Big(- n \times \frac{\delta^2}{(3+\delta)^2}/2048 + \log 3\Big) \\ &\ge 1 - \exp\Big(- n \times \frac{\delta^2}{16 \times 2048} + \log 3\Big) \\ &\ge 1 - \exp(-C_0 n\delta^2)\,, \end{aligned}$$

for some constant $C_0 > 0$ as long as $n \gtrsim 1/\delta^2$. □

Proof of Lemma 28. We only show variance bound for X. Since $\widehat{\sigma}_X^2 \sim \chi_n^2/n$, using the concentration of χ^2 distribution, we have

$$\Pr(|\widehat{\sigma}_X^2 - 1| \ge \zeta) = \Pr(|\chi_n^2 - n|/n \ge \zeta) \le \exp(-n\zeta^2/16)\,.$$

Now we show bound for covariance. Since bivariate Gaussian (X, Y) can be reparameterized by

$$\begin{aligned} X &= U + W \\ Y &= V + W \end{aligned}$$

where U,V,W are mutually independent with $\mathrm{Var}(U) = \mathrm{Var}(V) = 1 - r$, $\mathrm{Var}(W) = r$. Therefore,

$$\begin{aligned} \widehat{\sigma}_{XY} &= \frac{1}{n}\sum_{i=1}^n U^{(i)}V^{(i)} + \frac{1}{n}\sum_{i=1}^n U^{(i)}W^{(i)} + \frac{1}{n}\sum_{i=1}^n V^{(i)}W^{(i)} + \frac{1}{n}\sum_{i=1}^n W^{(i)2} \\ &= \frac{1-r}{2n}\Big[\sum_{i=1}^n \Big(\frac{U'^{(i)} + V'^{(i)}}{\sqrt{2}}\Big)^2 - \sum_{i=1}^n \Big(\frac{U'^{(i)} - V'^{(i)}}{\sqrt{2}}\Big)^2\Big] \\ &\quad + \frac{\sqrt{r(1-r)}}{2n}\Big[\sum_{i=1}^n \Big(\frac{U'^{(i)} + W'^{(i)}}{\sqrt{2}}\Big)^2 - \sum_{i=1}^n \Big(\frac{U'^{(i)} - W'^{(i)}}{\sqrt{2}}\Big)^2\Big] \\ &\quad + \frac{\sqrt{r(1-r)}}{2n}\Big[\sum_{i=1}^n \Big(\frac{V'^{(i)} + W'^{(i)}}{\sqrt{2}}\Big)^2 - \sum_{i=1}^n \Big(\frac{V'^{(i)} - W'^{(i)}}{\sqrt{2}}\Big)^2\Big] + \frac{r}{n}\sum_{i=1}^n W'^{(i)2} \\ &\overset{\mathcal{D}}{\sim} \frac{1-r}{2n}(\chi_{n11}^2 - \chi_{n12}^2) + \frac{\sqrt{r(1-r)}}{2n}(\chi_{n21}^2 - \chi_{n22}^2) + \frac{\sqrt{r(1-r)}}{2n}(\chi_{n31}^2 - \chi_{n32}^2) + \frac{r}{n}\chi_{n4}^2 \end{aligned}$$

where U', V', W' are standard normal random variables, thus $\sum_{i=1}^{n}(U'^{(i)} \pm V'^{(i)})^2/2$, $\sum_{i=1}^{n}(U'^{(i)} \pm W'^{(i)})^2/2$, $\sum_{i=1}^{n}(V'^{(i)} \pm W'^{(i)})^2/2$ are χ^2_n random variables. Since $r \leq 1$,

$$\begin{aligned}
\Pr(|\widehat{\sigma}_{XY} - r| \geq \zeta) &\leq \Pr\Big(\frac{1-r}{2} \times \frac{1}{n}|\chi^2_{n11} - \chi^2_{n12}| \geq \zeta/4\Big) \\
&\quad + \Pr\Big(\frac{\sqrt{r(1-r)}}{2} \times \frac{1}{n}|\chi^2_{n21} - \chi^2_{n22}| \geq \zeta/4\Big) \\
&\quad + \Pr\Big(\frac{\sqrt{r(1-r)}}{2} \times \frac{1}{n}|\chi^2_{n31} - \chi^2_{n32}| \geq \zeta/4\Big) \\
&\quad + \Pr\Big(r \times |\chi^2_{n41}/n - 1| \geq \zeta/4\Big) \\
&\leq \Pr\Big(|\chi^2_{n11}/n - 1| \geq \zeta/8\Big) + \Pr\Big(|\chi^2_{n12}/n - 1| \geq \zeta/8\Big) \\
&\quad + \Pr\Big(|\chi^2_{n21}/n - 1| \geq \zeta/8\Big) + \Pr\Big(|\chi^2_{n22}/n - 1| \geq \zeta/8\Big) \\
&\quad + \Pr\Big(|\chi^2_{n31}/n - 1| \geq \zeta/8\Big) + \Pr\Big(|\chi^2_{n32}/n - 1| \geq \zeta/8\Big) \\
&\quad + \Pr\Big(|\chi^2_{n41}/n - 1| \geq \zeta/4\Big) \\
&\leq 7\exp(-n\zeta^2/32^2) \leq \exp(-n\zeta^2/2048)\,.
\end{aligned}$$

The last inequality holds when $n \geq 2048 \log 7/\zeta^2$. □

B.3.3 Proof of Theorem 4.3.3

Theorem 4.3.3. *For any $T \in \widetilde{\mathcal{T}}$, assuming P is c-strong tree-faithful to T, applying Algorithm 4 with sample correlation for CI testing, if the sample size*

$$n \gtrsim \frac{1}{c^2}\Big(\log d + \log(1/\delta)\Big)\,,$$

then $\Pr(\widehat{T} = \mathrm{sk}(T)) \geq 1 - \delta$, *and* $\Pr(\textsc{Orient}(\widehat{T}, S) = \overline{T}) \geq 1 - \delta$

Proof. We firstly show the correctness of Algorithm 4. We make following notation of sets of nodes:

- $W = \{(j,k) : 1 \leq j < k \leq d\}$ is the set of all pairs of nodes in $[d]$;
- E is the true edge set;
- $A = \{(j,k) : j$ and k are d-separated by $\emptyset\}$;
- $B = \{(j,k) : \exists \ell \in [d] \setminus \{j,k\}, j$ and k are d-separated by $\ell\}$

- $C = \{(j,k) : \exists \ell \in [d] \setminus \{j,k\}, j \to \ell \leftarrow k \text{ is a } v\text{-structure}\}$
- $D = \{(j,k) : \exists \ell \in [d] \setminus \{j,k\}, j - \ell - k \text{ is a unshielded triple but not a } v\text{-structure}\}$

We claim that

1. E and $A \cup B$ are disjoint;
2. $W = E \cup A \cup B$;
3. $C \subseteq A$;
4. $D \subseteq B$.

It is easy to see the first claim, since for any pair of nodes connected by an edge, they cannot be d-separated by any set, and vice versa.

For the second claim, it suffices to show that for any pair of nodes not adjacent, it is in either A or B. First of all, for any two nodes j and k not adjacent, there will be one and only one path, denoted as ϕ, with length at least two between them. By property of polytree:

- If there is a collider on ϕ, then the path is blocked by $\emptyset$, so $(j,k) \in A$;
- If there is no collider on ϕ, then any node on ϕ will block the path, thus there exists $\ell \in [d] \setminus \{j,k\}$ such that i and j are d-separated by ℓ, so $(j,k) \in B$.

For the third claim, since $j \to \ell \leftarrow k$ is the only path between (j,k), which is blocked by $\emptyset$, thus $C \subseteq A$. For the forth claim, since $j - \ell - k$ is the only path between (j,k), either one of $j \to \ell \to k$ and $j \leftarrow \ell \leftarrow k$ and $j \leftarrow \ell \to k$ will be blocked by ℓ, thus $D \subseteq B$.

We now claim if the CI tests in Step 2 of Algorithm 4 are correct for

- all pairs $(j,k) \in E$ with $\ell \in [d] \cup \{\emptyset\} \setminus \{j,k\}$;
- all pairs $(j,k) \in A$ with $\ell = \emptyset$;
- all pairs $(j,k) \in C$ with ℓ being the collider;
- all pairs $(j,k) \in B$ with ℓ being the corresponding separation node(s) ,

then

1. the returned $\widehat{T}$ has the correct edge set E thus is the correct skeleton;
2. for any $(j,k) \in C$, $\ell \notin S(j,k)$;
3. for any $(j,k) \in D$, $\ell \in S(j,k)$.

For the first claim, if the CI tests conducted in Step 2 are correct for E, then pairs in E will pass all the CI tests and be included into $\widehat{E}$ (which is ensured by adjacency-faithfulness in Tree-faithfulness). But pairs in A will not pass marginal independence tests, and pairs in B will not pass some CI tests with corresponding ℓ (which is ensured by Markov property). Therefore, the returned $\widehat{T}$ is the correct skeleton. The second claim is ensured by orientation-faithfulness in Tree-faithfulness, and the third claim is ensured by Markov property and $D \subseteq B$.

Once the returned $\widehat{T}$ is the correct skeleton, Algorithm 14 will use the returned separation sets to determine v-structure for each possible unshielded triple. Note that $\{\text{All unshielded triples}\} = C \cup D$. For any $(j,k) \in C$, $\ell \notin S(j,k)$, thus it will be oriented as a v-structure; For any $(j,k) \in D$, $\ell \in S(j,k)$; thus it will remain as non-v-structure. Then ORIENT step is correct, which leads to correct CPDAG.

Finally we show the sample complexity of Algorithm 4 with CI tester Equation B.26. Note that correct CI tests implies correct estimation. Therefore,

$$
\begin{aligned}
&\Pr(\widehat{T} \neq \mathrm{sk}(T)) \\
&\leq \Pr\Bigg(\bigcup_{\substack{(j,k)\in E \\ \ell\in[d]\cup\{\emptyset\}\setminus\{j,k\}}} \text{or} \bigcup_{\substack{(j,k)\in A \\ \ell=\emptyset}} \text{or} \bigcup_{\substack{(j,k)\in C \\ \ell=\text{collider}}} \text{or} \bigcup_{\substack{(j,k)\in B \\ \ell \text{ d-separates } (j,k)}} |\widehat{\rho}_{ij\,|\,\ell} - \rho_{ij\,|\,\ell}| \geq c/2 \Bigg) \\
&\leq \binom{d}{2} \times (1 + (d-2)) \times \sup_{\substack{(j,k)\in E \\ \ell\in[d]\cup\{\emptyset\}\setminus\{j,k\}}} \text{or} \sup_{\substack{(j,k)\in A \\ \ell=\emptyset}} \text{or} \sup_{\substack{(j,k)\in C \\ \ell=\text{collider}}} \text{or} \sup_{\substack{(j,k)\in B \\ \ell \text{ d-separates } (j,k)}} \Pr(|\widehat{\rho}_{ij\,|\,\ell} - \rho_{ij\,|\,\ell}| \geq c/2) \\
&\leq \exp(3\log d) \times \sup_{\substack{(j,k)\in E \\ \ell\in[d]\cup\{\emptyset\}\setminus\{j,k\}}} \text{or} \sup_{\substack{(j,k)\in A \\ \ell=\emptyset}} \text{or} \sup_{\substack{(j,k)\in C \\ \ell=\text{collider}}} \text{or} \sup_{\substack{(j,k)\in B \\ \ell \text{ d-separates } (j,k)}} \Pr(|\widehat{\rho}_{ij\,|\,\ell} - \rho_{ij\,|\,\ell}| \geq c/2) \\
&\leq \exp\Big(-C_0(n-1)c^2 + 3\log d \Big).
\end{aligned}
$$

The first inequality is because it suffices to have $|\widehat{\rho}_{ij\,|\,\ell} - \rho_{ij\,|\,\ell}| \leq c/2$ for correct CI test. By c-strong Tree-faithfulness, $|\rho_{ij\,|\,S}| \geq c$ for $\rho_{ij\,|\,S} \neq 0$. Therefore,

$$
\begin{cases}
\widehat{\rho}_{ij\,|\,S} > c/2 & \text{if } \rho_{ij\,|\,S} \neq 0 \\
\widehat{\rho}_{ij\,|\,S} \leq c/2 & \text{if } \rho_{ij\,|\,S} = 0
\end{cases}
$$

Thus the cutoff $= c/2$ implies correct CI tests. The last inequality is by Lemma 25 where $q = 1$ and the sample size requirement is satisfied by the stated sample complexity. Set RHS to be smaller than δ, we need sample complexity

$$n \gtrsim \frac{1}{c^2}\left(\log d + \log\frac{1}{\delta}\right),$$

which completes the proof. □

B.3.4 Proof of Theorem 4.3.4

Theorem 4.3.4. *Assuming c-strong tree-faithfulness, and $c^2 \leq 1/5$, $d \geq 4$, if the sample size is bounded as*

$$n \leq \frac{1-2\delta}{8} \times \frac{\log d}{c^2},$$

then for any estimator $\widehat{T}$ for $\overline{T}$,

$$\inf_{\widehat{T}} \sup_{\substack{T\in\widetilde{\mathcal{T}} \\ P \text{ is } c\text{-strong} \\ \text{tree-faithful to } T}} \Pr(\widehat{T} \neq \overline{T}) \geq \delta - \frac{\log 2}{\log d}.$$

Proof. We construct a hard ensemble to show the lower bound. The construction is as follows: consider a subset $\mathcal{T}' \subset \mathcal{T} \subset \widetilde{\mathcal{T}}$, where $\mathcal{T}'$ is all the directed trees rooted at the first node $k = 1$. $\mathcal{T}'$ has the same cardinality as all undirected trees with d nodes, and the elements in it have different skeletons and no v-structures. Since our target is MEC, which is determined by its skeleton and v-structures, we have at least as many MECs as undirected trees, which leads to cardinality $|\mathcal{T}'| = d^{d-2}$ using Cayley's formula. Thus the size of the ensemble is lower bounded as

$$\log|\mathcal{T}'| = (d-2)\log d \geq \frac{1}{2} d \log d$$

The inequality holds when d is large enough, e.g. $d \geq 4$. Any directed tree has an important property: each node has at most one parent. Then we parameterize $\mathcal{T}'$ as follows

$$X_k = \beta X_{\mathrm{pa}(k)} + \eta_k, \quad \forall k \in [d] \tag{B.27}$$

where $\eta_k \sim \mathcal{N}(0,1)$ for all $k \in [d]$. Now we determine $\beta > 0$ to make sure the parametrization satisfies c-strong Tree-faithfulness.

In the subsequent lemma, we assert that the condition $\beta^2 = 2c^2 \asymp c^2$ is adequate for the validity of c-strong Tree-faithfulness, provided that c is sufficiently small:

Lemma 29. *If $\beta = \sqrt{2}c$ and $c^2 \leq 1/5$, then for any $T \in \mathcal{T}'$, the distribution defined in Equation B.27:*

1. *is c-strong Tree-faithful to T;*
2. *for all $k \in [d]$, $\mathrm{Var}(X_k) \leq 1 + \frac{\beta^2}{1-\beta^2}$.*

It remains to bound the KL divergence between any two instances in this ensemble. Before that, we claim that for any instance, we have $\mathrm{Cov}(X_k, X_j) > 0$ for all distinct $j, k \in [d]$. This is because for any pair of distinct nodes (j, k), there can be 3 possible paths between them:

- There is a directed path $j \to \phi_1 \to \cdots \to \phi_h \to k$ with length $h+1$, then $\mathrm{Cov}(X_j, X_k) = \mathbf{E}[X_j X_k] = \beta^{h+1}\,\mathbf{E}[X_j^2] > 0$;
- There is a directed path $k \to \phi_1 \to \cdots \to \phi_h \to j$ with length $h+1$, then $\mathrm{Cov}(X_j, X_k) = \mathbf{E}[X_j X_k] = \beta^{h+1}\,\mathbf{E}[X_k^2] > 0$;
- j, k share a common ancestor ℓ and there is a path $j \leftarrow \phi_1 \leftarrow \cdots \leftarrow \phi_h \leftarrow \ell \to \varphi_1 \to \cdots \to \varphi_g \to k$, then $\mathrm{Cov}(X_j, X_k) = \mathbf{E}[X_j X_k] = \beta^{h+g+2}\,\mathbf{E}[X_\ell^2] > 0$.

To compute the KL divergence between distributions P_0 and P_1 induced by any two $T_0, T_1 \in \mathcal{T}'$, let's first look at the covariance matrices of them Σ_0, Σ_1. Under our parametrization, they share the same determinant. To see this, let covariance matrix of η be $\Sigma_\eta = I_d$, for $\ell \in \{0, 1\}$, $\det(\Sigma_\ell) = \det(\Sigma_\eta) = \det(I_d) = 1$. Then the KL divergence is:

$$\begin{aligned} D_{\mathrm{KL}}(P_0 \| P_1) &= \mathop{\mathbf{E}}_{P_0} \log \frac{P_0}{P_1} \\ &= \mathop{\mathbf{E}}_{P_0} \log \frac{\exp\left(-\frac{1}{2}\sum_{k=1}^d (X_k - \beta\,\mathrm{pa}_{T_0}(k))^2 \right) / \sqrt{\det(\Sigma_0)}}{\exp\left(-\frac{1}{2}\sum_{k=1}^d (X_k - \beta\,\mathrm{pa}_{T_1}(k))^2 \right) / \sqrt{\det(\Sigma_1)}} \\ &= \frac{1}{2}\left[\mathop{\mathbf{E}}_{P_0} \sum_{k=1}^d (X_k - \beta\,\mathrm{pa}_{T_1}(k))^2 - d \right]. \end{aligned}$$

For all $k \in [d]$, let $\mathrm{pa}_{T_1}(k) = j$, then

$$\mathop{\mathbf{E}}_{P_0}(X_k - \beta\,\mathrm{pa}_{T_1}(k))^2 = \mathop{\mathbf{E}}_{P_0}[X_k^2] + \beta^2 \mathop{\mathbf{E}}_{P_0}[X_j^2] - 2\beta \mathop{\mathbf{E}}_{P_0}[X_k X_j]$$

$$\begin{aligned}
&\leq \mathop{\mathbf{E}}_{P_0}[X_k^2] + \beta^2 \mathop{\mathbf{E}}_{P_0}[X_j^2] \\
&\leq (1+\beta^2)\left(1 + \frac{\beta^2}{1-\beta^2}\right) \\
&= 1 + \frac{2\beta^2}{1-\beta^2}\,.
\end{aligned}$$

The first inequality is because all covariances are positive; the second one is due to the upper bound for all variances. Thus, we have

$$D_{\mathrm{KL}}(P_0 \| P_1) \leq \frac{1}{2}\left(d + \frac{2d\beta^2}{1-\beta^2} - d\right) = d\beta^2 \times \frac{1}{1-\beta^2} \leq 2d\beta^2 = 4dc^2$$

The last inequality holds when β^2 is small enough, e.g. $\beta^2 \leq 1/2$. The proof follows from applying Fano's inequality with KL divergence upper bound $4dc^2$ and cardinality of ensemble lower bound $\frac{1}{2}d\log d$. □

We end by proving the lemma used in the lower bound proof.

Proof of Lemma 29. Since for any $T \in \mathcal{T}'$, there is no v-structure because each node has at most one parent, thus it suffices to show the first part of Definition 4.3.2.

We first show all marginal variances are bounded, i.e. $1 \leq \mathrm{Var}(X_k) \leq 1 + \beta^2/(1-\beta^2)$ for all $k \in [d]$. Starting from the root node r, whose variance is $\mathrm{Var}(X_r) = \mathrm{Var}(\eta_r) = 1$, we can compute the variances of its children, they are all $\mathrm{Var}(X_\ell) = \mathrm{Var}(\eta_\ell) + \beta^2 \mathrm{Var}(X_r) = 1 + \beta^2$ for all $\ell \in \mathrm{ch}(r)$. Proceed the calculation, $\mathrm{Var}(X_j) = \mathrm{Var}(\eta_j) + \beta^2 \mathrm{Var}(X_\ell) = 1 + \beta^2 + \beta^4$ for all $j \in \mathrm{ch}(\ell)$ and $\ell \in \mathrm{ch}(r)$. Therefore, because the longest path has length at most $d-1$,

$$1 \leq \mathrm{Var}(X_k) \leq 1 + \beta^2 + \beta^4 + \cdots + \beta^{2d} = 1 + \frac{\beta^2}{1-\beta^2} \times (1 - \beta^{2(d-1)}) \leq 1 + \frac{\beta^2}{1-\beta^2}\,, \quad \forall k \in [d]$$

Now we can show the marginal correlation is lower bounded for any adjacent nodes (j,k). Without loss of generality, let $j = \mathrm{pa}(k)$, then $X_k = \beta X_j + \eta_k$, and the correlation

$$\rho(X_j, X_k) = \frac{\mathbf{E}[X_j X_k]}{\sqrt{\mathrm{Var}(X_k)\,\mathrm{Var}(X_j)}} = \beta\sqrt{\frac{\mathbf{E}[X_j^2]}{1 + \beta^2\,\mathbf{E}[X_j^2]}}$$

Thus $\rho(X_j, X_k) \geq c \Leftrightarrow \beta^2\,\mathbf{E}[X_j^2] \geq \frac{c^2}{1-c^2}$. Since $\mathbf{E}(X_j^2) \geq 1$, then $\beta^2\,\mathbf{E}(X_j^2) \geq \beta^2 = 2c^2 \geq \frac{c^2}{1-c^2}$ when $c^2 \leq 1/2$. Now consider any pair of adjacent nodes (j,k), assuming $j = \mathrm{pa}(k)$, and any other node $\ell \in [d] \setminus \{j,k\}$, there are 4 cases on the relation between ℓ and (j,k):

1. ℓ is ancestor of j, i.e. a directed path ϕ: $\ell \to \phi_1 \to \cdots \to \phi_h \to j$;
2. j and ℓ share the same ancestor w, i.e. a directed path ϕ: $w \to \phi_1 \to \cdots \to \phi_h \to j$ and a directed path φ: $w \to \varphi_1 \to \cdots \to \varphi_g \to \ell$;
3. ℓ is a descendant of k, i.e. a directed path ϕ: $j \to k \to \phi_1 \to \cdots \to \phi_h \to \ell$;
4. ℓ is a descendant of j but not k, i.e. a directed path ϕ: $j \to \phi_1 \to \cdots \to \phi_h \to \ell$ not going through k;

where $h \geq 0$ in either case. We deal with them separately:

- For the first and second case, because $X_\ell \perp\!\!\!\perp \eta_k$, the conditional correlation is

$$\begin{aligned}\rho(X_k, X_j \,|\, X_\ell) &= \frac{\mathbf{E}[X_k X_j \,|\, X_\ell]}{\sqrt{\mathbf{E}(X_k^2 \,|\, X_\ell)\,\mathbf{E}(X_j^2 \,|\, X_\ell)}} \\ &= \frac{\beta\,\mathbf{E}[X_j^2 \,|\, X_\ell]}{\sqrt{\mathbf{E}(X_j^2 \,|\, X_\ell)(1+\beta^2\,\mathbf{E}(X_j^2 \,|\, X_\ell))}} \\ &= \sqrt{\frac{\beta^2\,\mathbf{E}[X_j^2 \,|\, X_\ell]}{1+\beta^2\,\mathbf{E}(X_j^2 \,|\, X_\ell)}}\end{aligned}$$

  Thus $\rho(X_k, X_j \,|\, X_\ell) \geq c \Leftrightarrow \beta^2\,\mathbf{E}(X_j^2 \,|\, X_\ell) \geq \frac{c^2}{1-c^2}$. Since $X_{\phi_h} \perp\!\!\!\perp \eta_j \,|\, X_\ell$, we have $\mathbf{E}(X_j^2 \,|\, X_\ell) = 1 + \beta^2\,\mathbf{E}(X_{\phi_h}^2 \,|\, X_\ell) \geq 1$, then $\beta^2\,\mathbf{E}(X_j^2 \,|\, X_\ell) \geq \beta^2 = 2c^2 \geq \frac{c^2}{1-c^2}$ when $c^2 \leq 1/2$.

- For the third case, denote $v = \mathbf{E}[X_j^2]$, let's compute the covariance matrix of (X_k, X_k, X_ℓ):

$$\begin{pmatrix} v & \beta v & \beta^{h+2} v \\ \beta v & \beta^2 v + 1 & \beta^{h+1}(\beta^2 v + 1) \\ \beta^{h+2} v & \beta^{h+1}(\beta^2 v + 1) & \beta^{2(h+2)} v + \beta^{2(h+1)} + \cdots + \beta^2 + 1 \end{pmatrix}$$

  Denote $V(v,h) = \beta^{2(h+2)} v + \beta^{2(h+1)} + \cdots + \beta^2 + 1$. The covariance matrix of (X_j, X_k) given X_ℓ

$$\begin{pmatrix} v & \beta v \\ \beta v & \beta^2 v + 1 \end{pmatrix} - \frac{1}{V(v,h)} \begin{pmatrix} \beta^{2(h+2)} v^2 & \beta^{2h+3} v(\beta^2 v + 1) \\ \beta^{2h+3} v(\beta^2 v + 1) & \beta^{2(h+1)}(\beta^2 v + 1)^2 \end{pmatrix}$$
$$= \frac{1}{V(v,h)} \left[\begin{pmatrix} v \cdot (\sum_{i=1}^{h+1} \beta^{2i} + 1) & \beta v \cdot (\sum_{i=1}^{h+1} \beta^{2i} + 1) \\ \beta v \cdot (\sum_{i=1}^{h+1} \beta^{2i} + 1) & (\beta^2 v + 1) \cdot (\sum_{i=1}^{h+1} \beta^{2i} + 1) \end{pmatrix} \right.$$

$$
\begin{aligned}
&- \begin{pmatrix} \beta^{2(h+2)}v^2 & \beta^{2h+3}v(\beta^2 v+1) \\ \beta^{2h+3}v(\beta^2 v+1) & \beta^{2(h+1)}(\beta^2 v+1)^2 \end{pmatrix} \Bigg] \\
&= \frac{1}{V(v,h)} \begin{pmatrix} (\beta^{2(h+1)}+\cdots+\beta^2+1)v & \beta v(\beta^{2h}+\cdots+\beta^2+1) \\ \beta v(\beta^{2h}+\cdots+\beta^2+1) & (\beta^{2h}+\cdots+\beta^2+1)(\beta^2 v+1) \end{pmatrix}
\end{aligned}
$$

Thus the conditional correlation is

$$
\begin{aligned}
\rho(X_j, X_k \,|\, X_\ell) &= \frac{\beta v \times \frac{1-\beta^{2(h+1)}}{1-\beta^2}}{\sqrt{v \times \frac{1-\beta^{2(h+2)}}{1-\beta^2} \times (1+\beta^2 v) \times \frac{1-\beta^{2(h+1)}}{1-\beta^2}}} \\
&= \sqrt{\frac{\beta^2 v}{1+\beta^2 v} \times \frac{1-\beta^{2(h+1)}}{1-\beta^{2(h+2)}}}
\end{aligned}
$$

Denote $f(h) = \frac{1-\beta^{2(h+1)}}{1-\beta^{2(h+2)}} = 1 - \frac{(1-\beta^2)\beta^{2(h+1)}}{1-\beta^{2(h+2)}}$, which is increasing in h with minimum value being $f(0) = \frac{1}{1+\beta^2}$. Therefore, $\rho(X_j, X_k \,|\, X_\ell) \geq c \Leftrightarrow \beta^2 v \geq \frac{c^2}{f(h)-c^2}$. Since $v = \mathbf{E}[X_j^2] \geq 1$ for all $j \in [d]$, then $\beta^2 v \geq \beta^2 = 2c^2 \geq \frac{c^2}{1/(1+2c^2)-c^2} \geq \frac{c^2}{f(h)-c^2}$ when $c^2 \leq 1/5$, which yields the bound.

- For the forth case, analogously, denote $v = \mathbf{E}[X_j^2]$, let's compute the covariance matrix of (X_k, X_k, X_ℓ):

$$
\begin{pmatrix} v & \beta v & \beta^{h+1}v \\ \beta v & \beta^2 v+1 & \beta^{h+2}v \\ \beta^{h+1}v & \beta^{h+2}v & \beta^{2(h+1)}+\beta^{2h}+\cdots+\beta^2+1 \end{pmatrix}
$$

Denote $W(v,h) = \beta^{2(h+1)} + \beta^{2h}v + + \cdots + \beta^2 + 1$. The covariance matrix of (X_j, X_k) given X_ℓ is

$$
\begin{aligned}
&\begin{pmatrix} v & \beta v \\ \beta v & \beta^2 v+1 \end{pmatrix} - \frac{1}{W(v,h)} \begin{pmatrix} \beta^{2(h+1)}v^2 & \beta^{2h+3}v^2 \\ \beta^{2h+3}v^2 & \beta^{2(h+2)}v^2 \end{pmatrix} \\
=&\frac{1}{W(v,h)} \Bigg[\begin{pmatrix} \beta^{2(h+1)}v^2+\beta^{2h}v+\cdots+\beta^2 v+v & \beta^{2(h+1)+1}v^2+\beta^{2h+1}v+\cdots+\beta^3 v+\beta v \\ \beta^{2(h+1)+1}v^2+\beta^{2h+1}v+\cdots+\beta^3 v+\beta v & (\beta^{2(h+1)+2}v^2+\beta^{2h+2}v+\cdots+\beta^4 v+\beta^2 v \\ & +\beta^{2(h+1)}v+\beta^{2h}+\cdots+\beta^2+1) \end{pmatrix} \\
&- \begin{pmatrix} \beta^{2(h+1)}v^2 & \beta^{2h+3}v^2 \\ \beta^{2h+3}v^2 & \beta^{2(h+2)}v^2 \end{pmatrix} \Bigg] \\
=&\frac{1}{W(v,h)} \begin{pmatrix} (\beta^{2h}+\beta^{2(h-1)}+\cdots+\beta^2+1)v & (\beta^{2h}+\beta^{2(h-1)}\cdots+\beta^2+1)\beta v \\ (\beta^{2h}+\beta^{2(h-1)}\cdots+\beta^2+1)\beta v & (\beta^{2h}+\beta^{2(h-1)}+\cdots+\beta^2+1)(\beta^2 v+1)+\beta^{2h+2}v \end{pmatrix}
\end{aligned}
$$

Thus the conditional correlation is

$$\rho(X_j, X_k \,|\, X_\ell) = \frac{\beta v \times \frac{1-\beta^{2(h+1)}}{1-\beta^2}}{\sqrt{v \times \frac{1-\beta^{2(h+1)}}{1-\beta^2} \times \left[(1+\beta^2 v) \times \frac{1-\beta^{2(h+1)}}{1-\beta^2} + \beta^{2h} \times \beta^2 v\right]}} = \sqrt{\frac{\beta^2 v}{1 + \left(1 + \frac{\beta^{2h}}{g(h)}\right)\beta^2 v}}$$

where $g(h) = \frac{1-\beta^{2(h+1)}}{1-\beta^2} \geq 1$ for $h \geq 0$. Since $\beta^2 = 2c^2 \leq 1$, then $1+\beta^{2h}/g(h) \leq 2$. Since $\rho(X_j, X_k \,|\, X_\ell) \geq c \Leftrightarrow \beta^2 v \geq \frac{c^2}{1-\left(1+\frac{\beta^{2h}}{g(h)}\right)c^2}$, and $v = \mathbf{E}[X_j^2] \geq 1$ for all $j \in [d]$, then $\beta^2 v \geq \beta^2 = 2c^2 \geq \frac{c^2}{1-2c^2} \geq \frac{c^2}{1-\left(1+\frac{\beta^{2h}}{g(h)}\right)c^2}$ when $c^2 \leq 1/5$, which completes the proof.

□

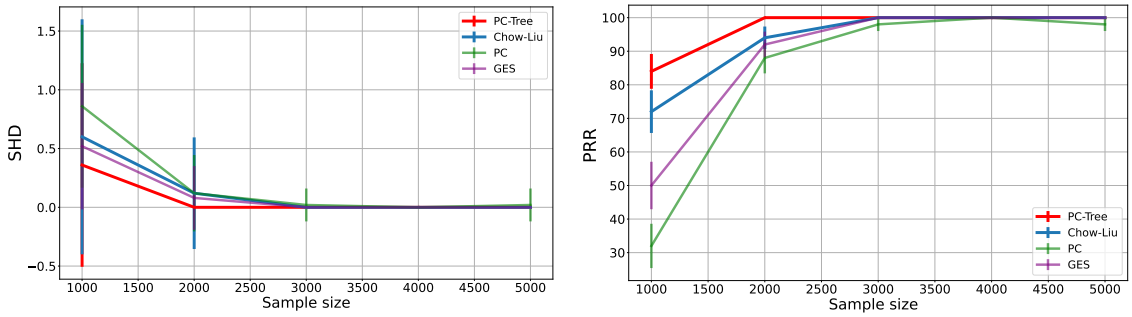
B.4 Experiments

Synthetic Data Generation We generate trees using package `networkx`, then randomly pick a node as root and orient it into a directed tree. We consider number of nodes $d \in \{10, 50, 100\}$. To generate the data as in Equation 8.9, we uniformly sample β_k from the interval $(-0.5, 0.1] \cup [0.1, 0.5)$ as our coefficient weight. For sample size $n = \{1000, 2000, 3000, 4000, 5000\}$, we generate our i.i.d. samples $X \in \mathbb{R}^{n \times d}$ according to Equation 8.9, where $\eta \sim \mathbb{N}(\mathbf{0}, I_{d \times d})$. Besides, we also present experiments on agnostic setting where $\eta \sim \mathcal{U}(-1, 1)$ is uniform distribution, or $\eta \sim$ Laplace$(0, 1)$ is Laplace distribution.

Baselines We have employed two baseline algorithms: the PC algorithm has been executed using the Python package `Causal-learn`, while the GES algorithm has been implemented with `py-tetrad`.

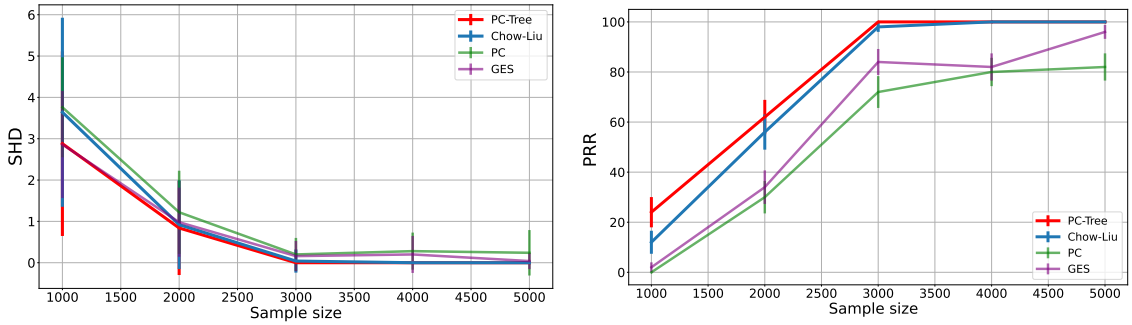
Evaluation For each experiment setup, we report the average (over 50 random instantiations) Structural Hamming Distance (SHD) between the ground truth and our estimated graph skeleton, and the Precise Recovery Rate (PRR), which is the

frequency of exact recovery of the tree skeleton. Results are reported in Fig. B.9-B.16. All experiments were conduced on an Intel Core i7-12800H 2.40GHz CPU.



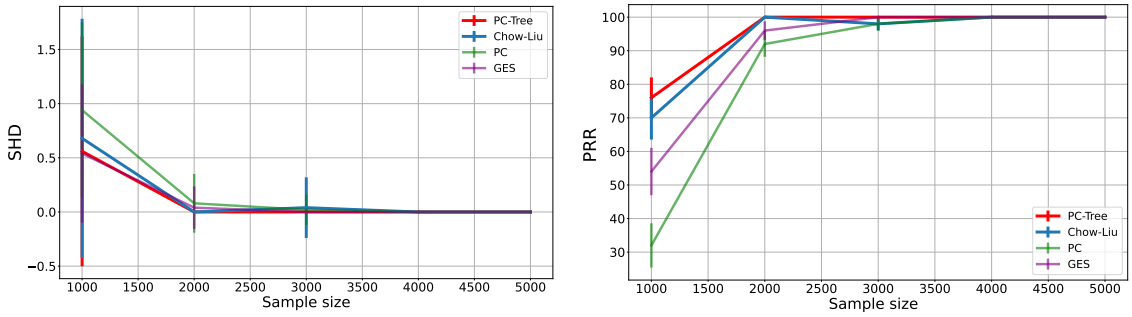
(a) SHD (b) PRR

Figure B.9: SHD and PRR for Gaussian η and $d = 10$.



(a) SHD (b) PRR

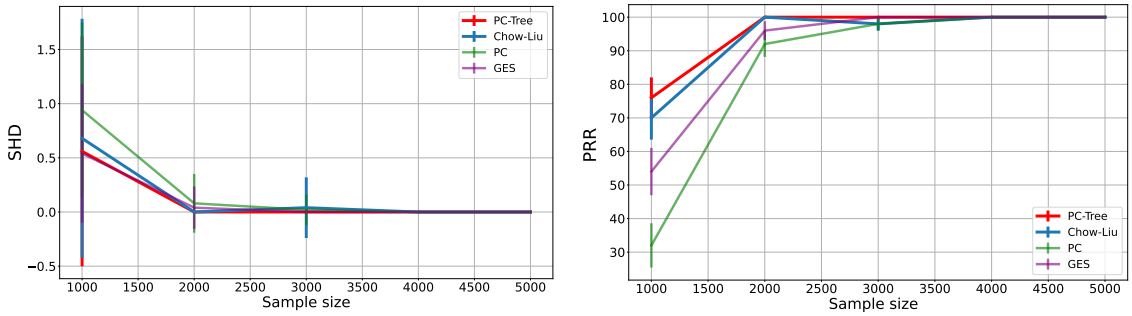
Figure B.10: SHD and PRR for Gaussian η and $d = 50$.



(a) SHD (b) PRR

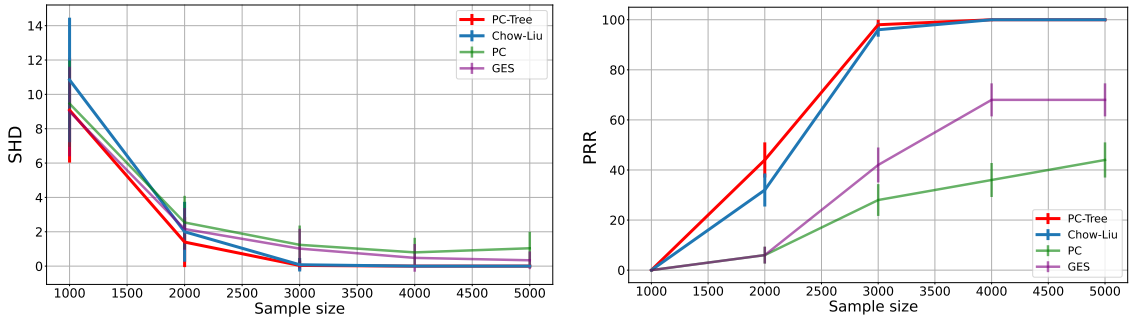
Figure B.11: SHD and PRR for Uniform η and $d = 10$.

Agnostic Learning Additionally, we investigated the algorithm's performance under conditions where the assumption is violated. Specifically, we examined the



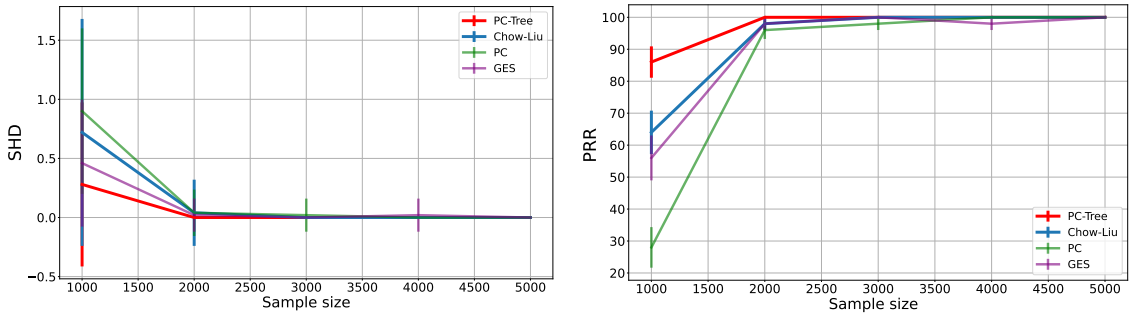
(a) SHD (b) PRR

Figure B.12: SHD and PRR for Uniform η and $d = 50$.



(a) SHD (b) PRR

Figure B.13: SHD and PRR for Uniform η and $d = 100$.

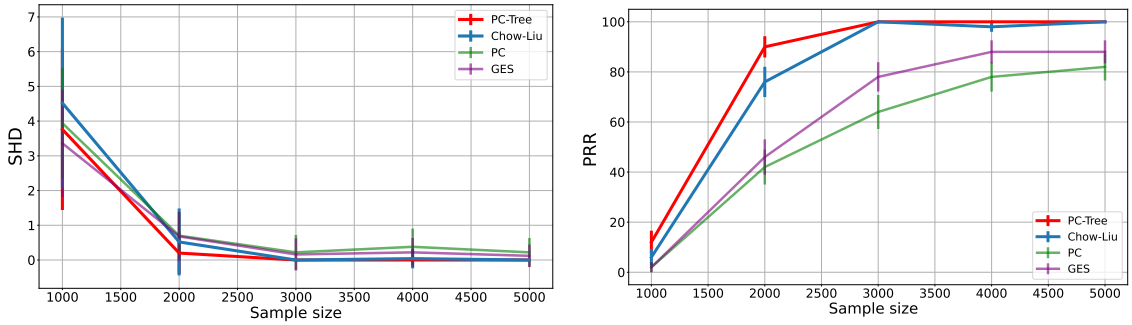


(a) SHD (b) PRR

Figure B.14: SHD and PRR for Laplace η and $d = 10$.

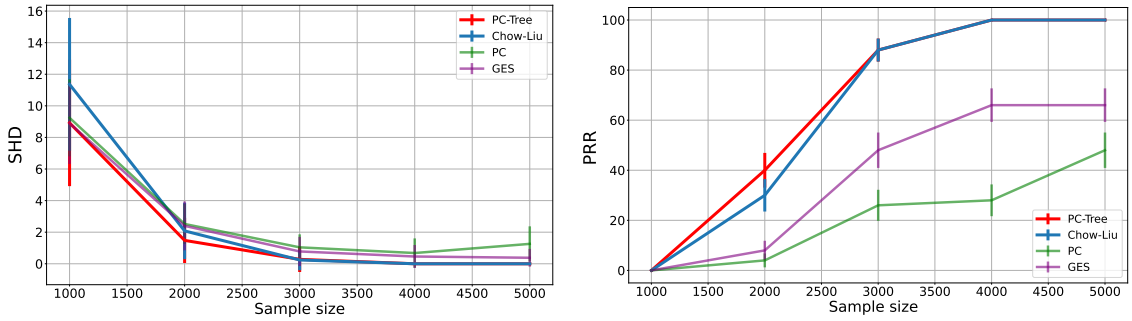
impact on our algorithm's performance when the coefficients β_k in Equation 8.9 are not independently and identically distributed (i.i.d.). To address this question, we conducted agnostic learning experiments and present the corresponding results.

See Fig. B.17 for results with non-iid β_k. Specifically, $\beta_k = \alpha_k + z$, where we sample α_k iid uniformly and z uniformly, applying the same z to all α_k. Here, z introduces dependence among β_k. When $z = 0$, β_k is i.i.d., and when $z \neq 0$, β_k is



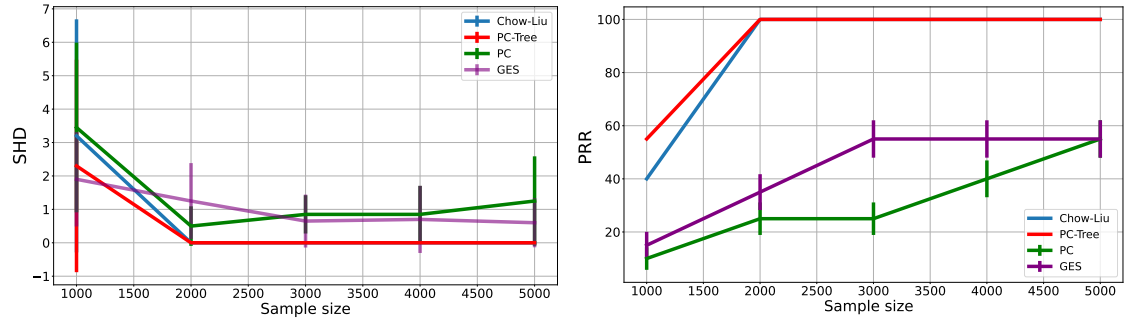
(a) SHD (b) PRR

Figure B.15: SHD and PRR for Laplace η and $d = 50$.



(a) SHD (b) PRR

Figure B.16: SHD and PRR for Laplace η and $d = 100$.



(a) SHD comparison (non-i.i.d.) (b) PRR comparison (non-i.i.d.)

Figure B.17: Performance comparison for PC-Tree, Chow-Liu, PC, and GES algorithms evaluated on SHD and PRR in (a) and (b) for non-iid β_k. The red, blue, green, and purple lines represent PC-Tree, Chow-Liu, PC, and GES, respectively.

non-i.i.d. For brevity, we only report the most relevant setting with $d = 100$ nodes and data are Gaussian. We simulated random directed trees and synthetic data via equation Equation 8.9. We can see the performance of both PC-tree and Chow-Liu are less affected even when β_k are non i.i.d: The Structural Hamming Distance

(SHD) becomes 0 in both i.i.d and non i.i.d. setting, and the Precise Recovery Rate (PRR) also outperforms other methods.

Appendix C

Supplementary Material - Chapter 5

C.1 Useful Lemma

Lemma 30 (Lebesgue Dominated Convergence Theorem (LDCT)). *Let $f_1, f_2, \ldots$ be a sequence of functions on a domain S. Suppose that this sequence converges to a function f pointwisely, i.e. for any $x \in S$, we have*

$$f_i(x) \to f(x)$$

and is dominated by a integrable function g, i.e.

$$|f_i(x)| \leq |g(x)| \qquad \text{for all } x \in S.$$

Then, we have

$$\lim_{i\to\infty} \int_S f_i(x)dx = \int_S f(x)dx.$$

Lemma 31. *For any zero-mean, twice diffierentiable and log-concave density functions f_1, f_2, let f, F be the functions*

$$f(y,\alpha) = \int_{-\infty}^{\infty} f_1(x) f_2(y-\alpha x)dx \quad \text{for all } y, \alpha \in \mathbb{R}$$
$$F(\alpha) = -\int_{-\infty}^{\infty} f(y,\alpha)\log f(y,\alpha)dy \quad \text{for all } \alpha \in \mathbb{R}.$$

Then, we have

$$\frac{\partial F(0)}{\partial \alpha} = 0 \quad \text{and} \quad \frac{\partial^2 F(0)}{\partial \alpha^2} = O\Big(\int_{-\infty}^{\infty} x^2 f_1(x)dx\Big).$$

Proof. We first compute

$$\frac{\partial f(y,\alpha)}{\partial \alpha} = -\int_{-\infty}^{\infty} x f_1(x) f_2'(y-\alpha x)dx \quad \text{by LDCT in Lemma 30}$$

and

$$\frac{\partial F(\alpha)}{\partial \alpha} = -\int_{-\infty}^{\infty} \frac{\partial f(y,\alpha)}{\partial \alpha}(1+\log f(y,\alpha))dy \quad \text{by LDCT in Lemma 30.}$$

Therefore, when $\alpha = 0$, we have

$$\frac{\partial f(y,0)}{\partial \alpha} = -\left(\int_{-\infty}^{\infty} x f_1(x)dx\right)\cdot f_2'(y) = 0 \quad \text{since we assume zero-mean}$$

and

$$\frac{\partial F(0)}{\partial \alpha} = -\int_{-\infty}^{\infty}\left(\frac{\partial f(y,0)}{\partial \alpha}\right)(1+\log f_2(y))dy = 0.$$

Moreover, we compute

$$\frac{\partial^2 f(y,\alpha)}{\partial \alpha^2} = \int_{-\infty}^{\infty} x^2 f_1(x) f_2''(y-\alpha x)dx \quad \text{by LDCT in Lemma 30}$$

and

$$\frac{\partial^2 F(\alpha)}{\partial \alpha^2} = -\int_{-\infty}^{\infty} \left(\frac{\partial^2 f(y,\alpha)}{\partial \alpha^2}(1+\log f(y,\alpha)) + \frac{\left(\frac{\partial f(y,\alpha)}{\partial \alpha}\right)^2}{f(y,\alpha)} \right) dy \quad \text{by LDCT in Lemma 30.}$$

Therefore, when $\alpha = 0$, we have

$$\begin{aligned} \frac{\partial^2 F(0)}{\partial \alpha^2} &= -\int_{-\infty}^{\infty} \left(\frac{\partial^2 f(y,0)}{\partial \alpha^2}(1+\log f_2(y)) + \frac{\left(\frac{\partial f(y,0)}{\partial \alpha}\right)^2}{f_2(y)} \right) dy \\ &= -\Big(\int_{-\infty}^{\infty} x^2 f_1(x)dx \Big) \Big(\int_{-\infty}^{\infty} f_2''(y)(1+\log f_2(y))dy \Big) \end{aligned} \tag{C.1}$$

Note that

$$(f_2(y)\log f_2(y))'' = f_2''(y)(1+\log f_{\eta_Y}(y)) + \frac{f_2'(y)^2}{f_2(y)}$$

and hence the term $\int_{-\infty}^{\infty} f_2''(y)(1+\log f_2(y))dy$ can be rewritten as

$$\int_{-\infty}^{\infty} f_2''(y)(1+\log f_2(y))dy = (f_2(y)\log f_2(y))' \,|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \frac{(f_2'(y))^2}{f_2(y)}dy. \tag{C.2}$$

By Lemma 32 and plugging Equation C.2 into Equation C.1, we have

$$\frac{\partial^2 F(0)}{\partial \alpha^2} = O\Big(\int_{-\infty}^{\infty} x^2 f_1(x)dx \Big).$$

□

Lemma 32. *For any twice diffierentiable and log-concave density function f, we have*

$$f'(x)(1+\log f(x)) \to 0 \text{ as } x \to \pm\infty \text{ and } \int_{-\infty}^{\infty} \frac{(f'(x))^2}{f(x)}dx \le O(1).$$

Proof. Since f is log-concave,we express f as

$$f(x) = e^{-3g(x)} \quad \text{for all } x \in \mathbb{R}$$

for some convex function g. Here, we introduce the factor 3 for simplicity in the following calculation. Note that

$$f'(x)(1+\log f(x)) = -3g'(x)e^{-3g(x)}(1-3g(x)) = \Big((9g(x)-3)e^{-2g(x)} \Big) \Big(g'(x)e^{-g(x)} \Big) \tag{C.3}$$

and

$$\int_{-\infty}^{\infty} \frac{(f'(x))^2}{f(x)} dx = \int_{-\infty}^{\infty} 9(g'(x))^2 e^{-3g(x)} dx = \int_{-\infty}^{\infty} 9(g'(x)e^{-g(x)})^2 e^{-g(x)} dx \qquad \text{(C.4)}$$

If we manage to prove that $g'(x)e^{-g(x)} < O(1)$ as $x \to \infty$ (for simplicity, we argue the statement for $x \to \infty$ only; the case of $x \to -\infty$ follows similarly), then we have the following. For Equation C.3, we have

$$f'(x)(1 + \log f(x)) = O\Big((9g(x) - 3)e^{-2g(x)}\Big) \to 0 \quad \text{as } x \to \infty \text{ since } g(x) \to \infty.$$

For Equation C.4, note that g is convex and $g(x) \to \infty$ which implies $g(x) = \Omega(x)$ as $x \to \infty$. We have

$$\int_{-\infty}^{\infty} \frac{(f'(x))^2}{f(x)} dx = \int_{-\infty}^{\infty} O(e^{-g(x)}) dx = O(1).$$

In other words, we need to show $g'(x)e^{-g(x)} < O(1)$ as $x \to \infty$. If g' is bounded, then $g'e^{-g/2}$ is also bounded since $g(x) \to \infty$ as $x \to \infty$. WLOG, assume that $g'(0) = 1$ and $g(0) = 0$. Recall that g' is increasing. Let x_1 be the smallest positive value such that $g'(x_1) = 2e^{x_1}$ or ∞ if such value does not exist. For all $x \in [0, x_1)$, we have

$$g'(x) < 2e^x \quad \text{and} \quad g(x) = \int_0^x g'(y)dy \geq x \quad \text{which implies} \quad g'(x)e^{-g(x)/2} < 2.$$

If $x_1 \neq \infty$, let x_2 be the smallest positive value such that $g'(x_2) = 2e^{2(x_2 - x_1)e^{x_1} + x_1}$ or ∞ if such value does not exist. For all $x \in [x_1, x_2)$, we have

$$g'(x) < 2e^{2(x-x_1)e^{x_1}+x_1} \quad \text{and} \quad g(x) = \int_{x_1}^x g'(y)dy + \int_0^{x_1} g'(y)dy \geq 2(x - x_1)e^{x_1} + x_1$$

which implies

$$g'(x)e^{-g(x)/2} < 2.$$

It is easy to see that we can prove the statement

$$g'(x)e^{-g(x)/2} < 2 \quad \text{for all } x > 0$$

by induction when we define x_i in a similar manner. We will omit the detail.

□

Lemma 33 (Concentration bound for sub-Gaussian random variables)**.** *For any sub-Gaussian random variables X, Y (not necessarily independent), let $(X^{(1)}, Y^{(1)}), \ldots, (X^{(n)}, Y^{(n)})$ be i.i.d. samples of (X, Y). Define*

$$\sigma_X^2 := \mathbf{E}(X^2), \quad \sigma_Y^2 := \mathbf{E}(Y^2), \quad \rho_{XY} := \mathbf{E}(XY),$$
$$\widehat{\sigma}_X^2 := \frac{1}{n}\sum_{i=1}^{n}(X^{(i)})^2 \quad \textit{and} \quad \widehat{\rho}_{XY} := \frac{1}{n}\sum_{i=1}^{n} X^{(i)}Y^{(i)}.$$

For any $0 < t < 1$, when $n = \Omega(\frac{1}{t^2}\log\frac{1}{\delta})$, we have

$$\left|\sigma_X^2 - \widehat{\sigma}_X^2\right| \le t\sigma_X^2 \quad \textit{and} \quad |\rho_{XY} - \widehat{\rho}_{XY}| \le t\sigma_X\sigma_Y \quad \textit{with probability } 1-\delta.$$

Fact C.1.1 ([Bha+21])**.** *Let T_1 and T_2 be two spanning trees on d vertices such that their symmetric difference consists of the edges $E = \{e_1, e_2, \ldots, e_l\} \in T_1 \setminus T_2$ and $F = \{f_1, f_2, \ldots, f_l\} \in T_2 \setminus T_1$. Then E and F can be paired up, say $\langle e_i, f_i\rangle$, such that for all i, $T_1 \cup \{f_i\} \setminus \{e_i\}$ is a spanning tree.*

Lemma 34. *Let (X, Y, Z) be a random variable. Note that one can always write X, Y, Z as follows.*

$$\begin{aligned} X &= \eta_X \\ Y &= \beta X + \eta_Y \\ Z &= \lambda X + \gamma Y + \eta_Z \end{aligned} \tag{C.5}$$

for some coefficients β, λ, γ and some random variables η_X, η_Y, η_Z where $\mathbf{E}(X\eta_Y) = \mathbf{E}(X\eta_Z) = \mathbf{E}(Y\eta_Z) = 0$. Suppose we have n i.i.d. samples of (X, Y, Z). Define

$$\tilde{I}(X;Y) = \log(1 + \frac{\widehat{\beta}^2\widehat{\sigma}_{\eta_X}^2}{\widehat{\sigma}_{\eta_Y}^2}) \quad \textit{and} \quad \tilde{I}(Y;Z \mid X) = \log(1 + \frac{\widehat{\gamma}^2\widehat{\sigma}_{\eta_Y}^2}{\widehat{\sigma}_{\eta_Z}^2})$$

It is easy to check that $\widehat{I}(X;Y)$ and $\widehat{I}(Y;Z \mid X)$ can be rewritten in terms of $\widehat{\sigma}_X^2, \widehat{\sigma}_Y^2, \widehat{\sigma}_Z^2, \widehat{\rho}_{XY}, \widehat{\rho}_{XZ}, \widehat{\rho}_{YZ}$ which is independent of the form in Equation C.5, e.g.

$$\tilde{I}(X;Y) = -\log(1 - \frac{\widehat{\rho}_{XY}^2}{\widehat{\sigma}_X^2\widehat{\sigma}_Y^2}) \quad \textit{recall that} \quad \widehat{\sigma}_{\eta_X}^2 = \widehat{\sigma}_X^2, \quad \widehat{\beta} = \frac{\widehat{\rho}_{XY}}{\widehat{\sigma}_X^2} \quad \textit{and} \quad \widehat{\sigma}_{\eta_Y}^2 = \widehat{\sigma}_Y^2 - \widehat{\beta}^2\widehat{\sigma}_X^2.$$

Then, we have

$$\tilde{I}(X;Y) - \tilde{I}(X;Z) = \tilde{I}(X;Y \mid Z) - \tilde{I}(X;Z \mid Y).$$

Proof. Let $(\mathring{X}, \mathring{Y}, \mathring{Z})$ be the Gaussian random variable $\mathcal{N}(0, \Sigma)$ whose covariance matrix Σ is

$$\Sigma = \begin{bmatrix} \widehat{\sigma}_X^2 & \widehat{\rho}_{XY} & \widehat{\rho}_{XZ} \\ \widehat{\rho}_{XY} & \widehat{\sigma}_Y^2 & \widehat{\rho}_{YZ} \\ \widehat{\rho}_{XZ} & \widehat{\rho}_{YZ} & \widehat{\sigma}_Z^2 \end{bmatrix}.$$

Since $(\mathring{X}, \mathring{Y}, \mathring{Z})$ is Gaussian, by the standard calculation of (conidtional) mutual information for Gaussians, we have

$$I(\mathring{X}; \mathring{Y}) = \frac{1}{2} \log(1 + \frac{\widehat{\beta}^2 \widehat{\sigma}_{\eta_X}^2}{\widehat{\sigma}_{\eta_Y}^2}) = \frac{1}{2} \tilde{I}(X; Y)$$
$$I(\mathring{Y}; \mathring{Z} \mid \mathring{X}) = \frac{1}{2} \log(1 + \frac{\widehat{\gamma}^2 \widehat{\sigma}_{\eta_Y}^2}{\widehat{\sigma}_{\eta_Z}^2}) = \frac{1}{2} \tilde{I}(Y; Z \mid X).$$

Therefore, we have

$$\begin{aligned} \tilde{I}(X;Y) - \tilde{I}(X;Z) &= 2I(\mathring{X}; \mathring{Y}) - 2I(\mathring{X}; \mathring{Z}) \\ &= 2I(\mathring{X}; \mathring{Y} \mid \mathring{Z}) - 2I(\mathring{X}; \mathring{Z} \mid \mathring{Y}) \quad \text{by the chain rule for mutual information} \\ &= \tilde{I}(X; Y \mid Z) - \tilde{I}(X; Z \mid Y). \end{aligned}$$

□

C.2 Proofs

Theorem 5.2.1. *Let (X, Y) be the random variable of the form in Equation 5.1. Assume that α is bounded above by a constant, i.e. $\alpha = O(1)$. Then, we have*

$$I(X; Y) \leq O\Big(\sigma_X^2 \alpha^2\Big).$$

Proof. For any random variable R, let f_R be its pdf. Then, we have

$$f_X(x) = f_{\eta_X}(x) \quad \text{for all } x \in \mathbb{R}$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{\eta_X}(x) f_{\eta_Y}(y - \alpha x) dx \quad \text{for all } y \in \mathbb{R} \text{ since } \eta_X \text{ and } \eta_Y \text{ are independent}$$
$$f_{Y|X}(y \mid x) = f_{\eta_Y}(y - \alpha x) \quad \text{for all } x, y \in \mathbb{R}.$$

Recall that the definition of $I(X; Y)$ is

$$I(X;Y) = -\int_{-\infty}^{\infty} f_Y(y) \log f_Y(y) dy + \int_{-\infty}^{\infty} f_X(x) \int_{-\infty}^{\infty} f_{Y|X}(y \mid x) \log f_{Y|X}(y \mid x) dy dx$$

For the second term, we have

$$\begin{aligned}
&\int_{-\infty}^{\infty} f_X(x) \int_{-\infty}^{\infty} f_{Y|X}(y \mid x) \log f_{Y|X}(y \mid x) dy dx \\
&= \int_{-\infty}^{\infty} f_{\eta_X}(x) \int_{-\infty}^{\infty} f_{\eta_Y}(y - \alpha x) \log f_{\eta_Y}(y - \alpha x) dy dx \\
&= \int_{-\infty}^{\infty} f_{\eta_X}(x) \int_{-\infty}^{\infty} f_{\eta_Y}(y) \log f_{\eta_Y}(y) dy dx \\
&= \int_{-\infty}^{\infty} f_{\eta_Y}(y) \log f_{\eta_Y}(y) dy.
\end{aligned}$$

Namely, the expression for $I(X;Y)$ can be written as

$$I(X;Y) = -\int_{-\infty}^{\infty} f_Y(y) \log f_Y(y) dy + \int_{-\infty}^{\infty} f_{\eta_Y}(y) \log f_{\eta_Y}(y) dy.$$

Note that f_Y depends on α and hence we abuse $f_Y(y) = f_Y(y, \alpha)$ for all $y, \alpha \in \mathbb{R}$.
Let F be the function

$$F(\alpha) := -\int_{-\infty}^{\infty} f_Y(y, \alpha) \log f_Y(y, \alpha) dy \quad \text{which also implies} \quad F(0) = -\int_{-\infty}^{\infty} f_{\eta_Y}(y) \log f_{\eta_Y}(y) dy.$$

Hence, we have

$$I(X;Y) = F(\alpha) - F(0).$$

Furthermore, we would like to expand the Taylor expansion for F at $\alpha = 0$, i.e.

$$I(X;Y) = \frac{\partial F(0)}{\partial \alpha}\alpha + \frac{1}{2}\frac{\partial^2 F(\alpha_0)}{\partial \alpha^2}\alpha^2 \quad \text{for some } \alpha_0 \text{ between 0 and } \alpha. \tag{C.6}$$

By Lemma 31, we have

$$\frac{\partial F(0)}{\partial \alpha} = 0 \quad \text{and} \quad \frac{\partial^2 F(0)}{\partial \alpha^2} \leq O\Big(\int_{-\infty}^{\infty} x^2 f_{\eta_X}(x) dx\Big) = O(\sigma_X^2).$$

Since $\frac{\partial^2 F(\alpha)}{\partial \alpha^2}$ is continuous, we also have

$$\frac{\partial^2 F(\alpha_0)}{\partial \alpha^2} \leq O(\sigma_X^2) \quad \text{for some } \alpha_0 \text{ between 0 and } \alpha.$$

In other words, Equation C.6 can be expressed as

$$I(X;Y) \leq O\Big(\sigma_X^2 \alpha^2\Big).$$

□

Theorem 5.2.2 (Mutual Information Tester)**.** *Suppose we are given n i.i.d. samples of (X, Y) of the form in Equation 5.1. For any sufficiently small $\varepsilon, \delta > 0$ that $I(X;Y) \geq \varepsilon$ when $I(X;Y) \neq 0$, if $n = \Omega(\frac{1}{\varepsilon} \log \frac{1}{\delta})$, there exists a constant C such that the estimator defined in Equation 5.4 satisfies the following with probability $1 - \delta$:*

- *If $I(X;Y) = 0$, then $\tilde{I}(X;Y) \leq \frac{\varepsilon}{100C}$.*
- *If $I(X;Y) \geq \varepsilon$, then $\tilde{I}(X;Y) \geq \frac{\varepsilon}{50C}$.*

Proof. By Theorem 5.2.1, we first have

$$I(X;Y) \leq C \cdot \frac{\sigma_X^2 \alpha^2}{\sigma_Y^2} \quad \text{for some large constant } C > 1. \tag{C.7}$$

Here, since σ_Y^2 is bounded, we introduce this term for analytic purposes.

By Lemma 33 and a straightforward calculation, we have

$$|\widehat{\alpha} - \alpha| \leq \sqrt{\frac{\varepsilon}{101C}} \frac{\sigma_Y}{\sigma_X}, \quad \left|\widehat{\sigma}_X^2 - \sigma_X^2\right| \leq \sqrt{\frac{\varepsilon}{100C}} \sigma_X^2 \quad \text{and} \quad \left|\widehat{\sigma}_{\eta_Y}^2 - \sigma_{\eta_Y}^2\right| \leq \sqrt{\frac{\varepsilon}{100C}} \sigma_Y^2. \tag{C.8}$$

Note that $\sigma_{\eta_Y}^2 = \sigma_Y^2 - \alpha^2 \sigma_X^2$ and hence the error depends on σ_Y^2 for $\left|\widehat{\sigma}_{\eta_Y}^2 - \sigma_{\eta_Y}^2\right|$. Now, we express

$$\tilde{I}(X;Y) = -\log\left(1 - \frac{\widehat{\rho}_{XY}^2}{\widehat{\sigma}_X^2 \widehat{\sigma}_Y^2}\right) = \log\left(1 + \frac{\widehat{\alpha}^2 \widehat{\sigma}_X^2}{\widehat{\sigma}_{\eta_Y}^2}\right) = \log\left(1 + \widehat{\alpha}^2 \frac{\sigma_X^2}{\sigma_Y^2} \cdot \frac{\widehat{\sigma}_X^2}{\sigma_X^2} \cdot \frac{\sigma_Y^2}{\widehat{\sigma}_{\eta_Y}^2}\right).$$

We will now consider the two cases: 1) $I(X;Y) = 0$ and 2) $I(X;Y) \geq \varepsilon$.

For 1) $I(X;Y) = 0$: When $I(X;Y) = 0$, it means X and Y are independent and hence $\alpha = 0$ and $\sigma_Y = \sigma_{\eta_Y}$. By Equation C.8, we have

$$\widehat{\alpha}^2 \frac{\sigma_X^2}{\sigma_Y^2} \leq \frac{\varepsilon}{101C}, \quad \frac{\widehat{\sigma}_X^2}{\sigma_X^2} \leq 1 + \sqrt{\frac{\varepsilon}{100C}} \quad \text{and} \quad \frac{\sigma_Y^2}{\widehat{\sigma}_{\eta_Y}^2} \leq \frac{1}{1 - \sqrt{\frac{\varepsilon}{100C}}}.$$

Hence, we now bound $\tilde{I}(X;Y)$

$$\tilde{I}(X;Y) \leq \log\left(1 + \frac{\varepsilon}{101C} \cdot \frac{1 + \sqrt{\frac{\varepsilon}{100C}}}{1 - \sqrt{\frac{\varepsilon}{100C}}}\right) \leq \frac{\varepsilon}{100C} \quad \text{for small } \varepsilon > 0.$$

For 2) $I(X;Y) \geq \varepsilon$**:** When $I(X;Y) \geq \varepsilon$, we have

$$\frac{\sigma_X \alpha}{\sigma_Y} \geq \sqrt{\frac{\varepsilon}{C}} \quad \text{by Equation C.7 which implies} \quad \frac{\sigma_X \widehat{\alpha}}{\sigma_Y} \geq \frac{\sigma_X \alpha}{\sigma_Y} - \sqrt{\frac{\varepsilon}{101C}} > 0.$$

Furthermore, we have

$$\frac{\sigma_X^2 \widehat{\alpha}^2}{\sigma_Y^2} \geq (\frac{\sigma_X \alpha}{\sigma_Y} - \sqrt{\frac{\varepsilon}{101C}})^2, \quad \frac{\widehat{\sigma}_X^2}{\sigma_X^2} \geq 1 - \sqrt{\frac{\varepsilon}{100C}} \quad \text{and} \quad \frac{\sigma_Y^2}{\widehat{\sigma}_{\eta_Y}^2} \geq \frac{1}{1+\sqrt{\frac{\varepsilon}{100C}}}.$$

Hence, we now bound $\tilde{I}(X;Y)$

$$\begin{aligned}
\tilde{I}(X;Y) &\geq \log\left(1 + \left(\frac{\sigma_X \alpha}{\sigma_Y} - \sqrt{\frac{\varepsilon}{101C}}\right)^2 \cdot \frac{1-\sqrt{\frac{\varepsilon}{100C}}}{1+\sqrt{\frac{\varepsilon}{100C}}}\right) \\
&\geq \log\left(1 + \underbrace{\left(\frac{\sigma_X^2 \alpha^2}{2\sigma_Y^2} - \frac{\varepsilon}{101C}\right)}_{\text{by } (a-b)^2 \geq \frac{1}{2}a^2 - b^2} \cdot \underbrace{\frac{9}{10}}_{\text{for small } \varepsilon}\right)
\end{aligned}$$

For small $a = \left(\frac{\sigma_X^2 \alpha^2}{2\sigma_Y^2} - \frac{\varepsilon}{101C}\right) \cdot \frac{9}{10} > 0$, we have $\log(1+a) \geq \frac{1}{2}a$. We have

$$\tilde{I}(X;Y) \geq \frac{1}{2}\left(\frac{\sigma_X^2 \alpha^2}{2\sigma_Y^2} - \frac{\varepsilon}{101C}\right) \cdot \frac{9}{10} \geq \frac{\varepsilon}{50C}.$$

□

Theorem 5.3.1. *Let (X,Y,Z) be the random variable of the form in Equation 5.5 such that $\lambda = 0$ and Y and ξ_Z are independent. Assume that γ is bounded above by a constant, i.e. $\gamma = O(1)$. Then, we have*

$$I(Y;Z \mid X) \leq O\left(\sigma_{\xi_Y}^2 \gamma^2\right).$$

Proof. For any random variable R, let f_R be its pdf. Then, we have

$$\begin{aligned}
f_X(x) &= f_{\xi_X}(x) \quad \text{for all } x \in \mathbb{R} \\
f_{Z|X}(z \mid x) &= \int_{-\infty}^{\infty} f_{\xi_Y}(y') f_{\xi_Z}(z - \gamma\beta x - \gamma y') dy' \quad \text{for all } x, z \in \mathbb{R} \\
f_{Z|X,Y}(z \mid x, y) &= f_{\xi_Z}(z - \gamma y) \quad \text{for all } x, y, z \in \mathbb{R}
\end{aligned}$$

Recall that the definition of $I(Y;Z \mid X)$ is

$$I(Y;Z \mid X) = -\int_{-\infty}^{\infty} f_X(x) \int_{-\infty}^{\infty} f_{Z|X}(z \mid x) \log f_{Z|X}(z \mid x) dz dx$$

$$+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\int_{-\infty}^{\infty} f_{Z|X,Y}(z\mid x,y)\log f_{Z|X,Y}(z\mid x,y)dzdxdy$$

For the second term, we have

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\int_{-\infty}^{\infty} f_{Z|X,Y}(z\mid x,y)\log f_{Z|X,Y}(z\mid x,y)dzdxdy\\
&\quad=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\int_{-\infty}^{\infty} f_{\xi_Z}(z-\gamma y)\log f_{\xi_Z}(z-\gamma y)dzdxdy\\
&\quad=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\int_{-\infty}^{\infty} f_{\xi_Z}(z)\log f_{\xi_Z}(z)dzdxdy\\
&\quad=\int_{-\infty}^{\infty} f_{\xi_Z}(z)\log f_{\xi_Z}(z)dz
\end{aligned}$$

Also, for the first term, we can rewrite it as

$$\begin{aligned}
&\int_{-\infty}^{\infty} f_X(x)\int_{-\infty}^{\infty} f_{Z|X}(z\mid x)\log f_{Z|X}(z\mid x)dzdx\\
&\quad=\int_{-\infty}^{\infty} f_X(x)\int_{-\infty}^{\infty}\Big(\int_{-\infty}^{\infty} f_{\xi_Y}(y')f_{\xi_Z}(z-\gamma\beta x-\gamma y')dy'\Big)\log\Big(\int_{-\infty}^{\infty} f_{\xi_Y}(y')f_{\xi_Z}(z-\gamma\beta x-\gamma y')dy'\Big)d\\
&\quad=\int_{-\infty}^{\infty} f_X(x)\int_{-\infty}^{\infty}\Big(\int_{-\infty}^{\infty} f_{\xi_Y}(y')f_{\xi_Z}(z-\gamma y')dy'\Big)\log\Big(\int_{-\infty}^{\infty} f_{\xi_Y}(y')f_{\xi_Z}(z-\gamma y')dy'\Big)dzdx\\
&\quad=\int_{-\infty}^{\infty}\Big(\int_{-\infty}^{\infty} f_{\xi_Y}(y')f_{\xi_Z}(z-\gamma y')dy'\Big)\log\Big(\int_{-\infty}^{\infty} f_{\xi_Y}(y')f_{\xi_Z}(z-\gamma y')dy'\Big)dz
\end{aligned}$$

Let F be the function

$$F(\gamma):=-\int_{-\infty}^{\infty}\Big(\int_{-\infty}^{\infty} f_{\xi_Y}(y')f_{\xi_Z}(z-\gamma y')dy'\Big)\log\Big(\int_{-\infty}^{\infty} f_{\xi_Y}(y')f_{\xi_Z}(z-\gamma y')dy'\Big)dz$$

which also implies

$$F(0)=\int_{-\infty}^{\infty} f_{\xi_Z}(z)\log f_{\xi_Z}(z)dz.$$

Hence, we have

$$I(Y;Z\mid X)=F(\gamma)-F(0).$$

Furthermore, we would like to expand the Taylor expansion for F at $\gamma=0$, i.e.

$$I(X;Y)=\frac{\partial F(0)}{\partial\gamma}\gamma+\frac{1}{2}\frac{\partial^2 F(\gamma_0)}{\partial\gamma^2}\gamma^2\quad\text{for some }\gamma_0\text{ between 0 and }\gamma.\tag{C.9}$$

By Lemma 31, we have

$$\frac{\partial F(0)}{\partial\gamma}=0\quad\text{and}\quad\frac{\partial^2 F(0)}{\partial\gamma^2}\le\Big(\int_{-\infty}^{\infty} y^2 f_{\xi_Y}(y)dy\Big)\Big(\int_{-\infty}^{\infty}\frac{(f'_{\xi_Z}(z))^2}{f_{\xi_Z}(z)}dz\Big)$$

Since $\frac{\partial^2 F(\gamma)}{\partial \gamma^2}$ is continuous, we also have

$$\frac{\partial^2 F(\gamma_0)}{\partial \gamma^2} \leq O(\sigma^2_{\xi_Y}) \quad \text{for some } \gamma_0 \text{ between 0 and } \gamma.$$

In other words, Equation C.9 can be expressed as

$$I(Y;Z \mid X) \leq O\Big(\sigma^2_{\xi_Y}\gamma^2\Big).$$

□

Theorem 5.3.2 (Conditional Mutual Information Tester)**.** *Suppose we are given n i.i.d. samples of (X,Y,Z) of the form in Equation 5.5. For any sufficiently small $\varepsilon, \delta > 0$ that $I(Y;Z \mid X) \geq \varepsilon$ when $I(Y;Z \mid X) \neq 0$, if $n = \Omega(\frac{1}{\varepsilon}\log\frac{1}{\delta})$, there exists a constant C such that the estimator defined in Equation 5.6 satisfies the following with probability $1-\delta$:*

- *If $I(Y;Z \mid X) = 0$, then $\tilde{I}(Y;Z \mid X) \leq \frac{\varepsilon}{100C}$.*
- *If $I(Y;Z \mid X) \geq \varepsilon$, $\lambda = 0$ and Y and ξ_Z are independent, then $\tilde{I}(Y;Z \mid X) \geq \frac{\varepsilon}{50C}$.*

Proof. By Theorem 5.3.1, we first have

$$I(Y;Z \mid X) \leq C \cdot \frac{\sigma^2_{\xi_Y}\gamma^2}{\sigma^2_Z} \quad \text{for some large constant } C > 1. \tag{C.10}$$

Here, since σ^2_Z is bounded, we introduce this term for analytic purposes. By Lemma 33 and a straightforward calculation, we have

$$|\widehat{\gamma} - \gamma| \leq \sqrt{\frac{\varepsilon}{101C}}\frac{\sigma_{\xi_Y}}{\sigma_Z}, \quad \left|\widehat{\sigma}^2_{\xi_Y} - \sigma^2_{\xi_Y}\right| \leq \sqrt{\frac{\varepsilon}{100C}}\sigma^2_Y \quad \text{and} \quad \left|\widehat{\sigma}^2_{\xi_Z} - \sigma^2_{\xi_Z}\right| \leq \sqrt{\frac{\varepsilon}{100C}}\sigma^2_Z. \tag{C.11}$$

Note that $\sigma^2_{\xi_Y} = \sigma^2_Y - \beta^2\sigma^2_X$ and $\sigma^2_{\xi_Z} = \sigma^2_Z - \gamma^2\sigma^2_Y$ and hence the error depends on σ^2_Y (resp. σ^2_Z) for $\left|\widehat{\sigma}^2_{\eta_Y} - \sigma^2_{\eta_Y}\right|$ (resp. $\left|\widehat{\sigma}^2_{\eta_Z} - \sigma^2_{\eta_Z}\right|$). Now, we express

$$\tilde{I}(Y;Z \mid X) = \log\left(1 + \frac{\widehat{\gamma}^2\widehat{\sigma}^2_{\xi_Y}}{\widehat{\sigma}^2_{\xi_Z}}\right) = \log\left(1 + \widehat{\gamma}^2\frac{\sigma^2_{\xi_Y}}{\sigma^2_Z} \cdot \frac{\widehat{\sigma}^2_{\xi_Y}}{\sigma^2_{\xi_Y}} \cdot \frac{\sigma^2_Z}{\widehat{\sigma}^2_{\xi_Z}}\right).$$

We will now consider the two cases: 1) $I(Y;Z \mid X) = 0$ and 2) $I(Y;Z \mid X) \geq \varepsilon$.

For 1) $I(Y;Z \mid X) = 0$**:** When $I(Y;Z \mid X) = 0$, it means $\gamma = 0$ and $\sigma_Z = \sigma_{\xi_Z}$. By Equation C.11, we have

$$\widehat{\gamma}^2 \frac{\sigma^2_{\xi_Y}}{\sigma^2_Z} \leq \frac{\varepsilon}{101C}, \quad \frac{\widehat{\sigma}^2_{\xi_Y}}{\sigma^2_{\xi_Y}} \leq 1 + \sqrt{\frac{\varepsilon}{100C}} \frac{\sigma^2_Y}{\sigma^2_{\xi_Y}} \quad \text{and} \quad \frac{\sigma^2_Z}{\widehat{\sigma}^2_{\eta_Z}} \leq \frac{1}{1 - \sqrt{\frac{\varepsilon}{100C}}}.$$

Hence, we now bound $\tilde{I}(X;Y)$

$$\tilde{I}(Y;Z \mid X) \leq \log\left(1 + \frac{\varepsilon}{101C} \cdot \frac{1 + \sqrt{\frac{\varepsilon}{100C}} \frac{\sigma^2_Y}{\sigma^2_{\xi_Y}}}{1 - \sqrt{\frac{\varepsilon}{100C}}}\right) \leq \frac{\varepsilon}{100C} \quad \text{for small } \varepsilon > 0.$$

For 2) $I(Y;Z \mid X) \geq \varepsilon$**:** When $I(Y;Z \mid X) \geq \varepsilon$, we have

$$\frac{\sigma_{\xi_Y}\gamma}{\sigma_Z} \geq \sqrt{\frac{\varepsilon}{C}} \quad \text{by Equation C.10 which implies} \quad \frac{\widehat{\sigma}_{\xi_Y}\gamma}{\widehat{\sigma}_Z} \geq \frac{\sigma_{\xi_Y}\gamma}{\sigma_Z} - \sqrt{\frac{\varepsilon}{101C}} > 0.$$

Furthermore, we have

$$\frac{\widehat{\sigma}^2_{\xi_Y}\gamma^2}{\widehat{\sigma}^2_Z} \geq (\frac{\sigma_{\xi_Y}\gamma}{\sigma_Z} - \sqrt{\frac{\varepsilon}{101C}})^2, \quad \frac{\widehat{\sigma}^2_{\xi_Y}}{\sigma^2_{\xi_Y}} \geq 1 - \sqrt{\frac{\varepsilon}{100C}} \frac{\sigma^2_Y}{\sigma^2_{\xi_Y}} \quad \text{and} \quad \frac{\sigma^2_Z}{\widehat{\sigma}^2_{\eta_Z}} \geq \frac{1}{1 + \sqrt{\frac{\varepsilon}{100C}}}.$$

Hence, we now bound $\tilde{I}(X;Y)$

$$\begin{aligned}
\tilde{I}(Y;Z \mid X) &\geq \log\left(1 + \left(\frac{\sigma_{\xi_Y}\gamma}{\sigma_Z} - \sqrt{\frac{\varepsilon}{101C}}\right)^2 \cdot \frac{1 - \sqrt{\frac{\varepsilon}{100C}} \frac{\sigma^2_Y}{\sigma^2_{\xi_Y}}}{1 + \sqrt{\frac{\varepsilon}{100C}}}\right) \\
&\geq \log\left(1 + \underbrace{\left(\frac{\sigma^2_{\xi_Y}\gamma^2}{2\sigma^2_Z} - \frac{\varepsilon}{101C}\right)}_{\text{by } (a-b)^2 \geq \frac{1}{2}a^2 - b^2} \cdot \underbrace{\frac{9}{10}}_{\text{for small } \varepsilon}\right)
\end{aligned}$$

For small $a = \left(\frac{\sigma^2_{\xi_Y}\gamma^2}{2\sigma^2_Z} - \frac{\varepsilon}{101C}\right) \cdot \frac{9}{10} > 0$, we have $\log(1+a) \geq \frac{1}{2}a$. We have

$$\tilde{I}(Y;Z \mid X) \geq \frac{1}{2}\left(\frac{\sigma^2_{\xi_Y}\gamma^2}{2\sigma^2_Z} - \frac{\varepsilon}{101C}\right) \cdot \frac{9}{10} \geq \frac{\varepsilon}{50C}.$$

□

Theorem 5.3.3. *Let $T^* \in \mathcal{T}$ be a directed tree. Suppose we are given n i.i.d. samples of $(X_1, \ldots, X_d)$ of the form in Equation 5.2. For any sufficiently small $\varepsilon, \delta > 0$ that $I(Y;Z \mid X) \geq \varepsilon$ when $I(Y;Z \mid X) \neq 0$ for any three nodes X, Y, Z in T^*, if $n = \Omega(\frac{1}{\varepsilon}\log\frac{d}{\delta})$, the tree outputted by Algorithm 5, $\widehat{T}$, is equal to the skeleton of T^* with probability $1 - \delta$.*

Proof. We first consider the edge difference between $\widehat{T}$ and T^*. By fact C.1.1, we can pair up the edges in $\widehat{T}\backslash T^*$ with the edges in $T^*\backslash\widehat{T}$ such that $T^*\cup\{(W,Z)\}\backslash\{(X,Y)\}$ is also a spanning tree for any $(W,Z)\in\widehat{T}\backslash T^*$ and $(X,Y)\in T^*\backslash\widehat{T}$. Let $\widehat{T}\backslash T^*$ be $\{(W_1,Z_1),\dots,(W_k,Z_k)\}$ and $T^*\backslash\widehat{T}$ be $\{(X_1,Y_1),\dots,(X_k,Y_k)\}$ such that (W_i,Z_i) pairs up with (X_i,Y_i) for $i=1,\dots,k$. Because of that, there exists a path in T^* from W_i to Z_i containing X_i and Y_i. Without loss of generality, we assume that the order of them is $W_i - X_i \to Y_i \rightsquigarrow Z_i$ in T^*. Here, $-$ means an undirected path in T^*, $\to$ means a directed edge in T^* and $\rightsquigarrow$ means a directed path in T^*.

Since $\widehat{T}$ is the output of Algorithm 5, we have

$$\sum_{i=1}^{k}\tilde{I}(X_i;Y_i)-\sum_{i=1}^{k}\tilde{I}(W_i;Z_i)\le 0$$

by the definition of the maximum spanning tree. We first expand the LHS as

$$\begin{aligned}
\sum_{i=1}^{k}\tilde{I}(X_i,Y_i)-\sum_{i=1}^{k}\tilde{I}(W_i,Z_i) &= \sum_{i=1}^{k}\Big(\tilde{I}(X_i,Y_i)-\tilde{I}(X_i;Z_i)+\tilde{I}(X_i;Z_i)-\tilde{I}(W_i;Z_i)\Big)\\
&= \sum_{i=1}^{k}\Big(\tilde{I}(X_i;Y_i\mid Z_i)-\tilde{I}(X_i;Z_i\mid Y_i)+\tilde{I}(X_i;Z_i\mid W_i)-\tilde{I}(W_i;Z_i\mid X_i)\Big) \quad \text{by Lemma 34}\\
&= \underbrace{\sum_{i=1}^{k}\Big(\tilde{I}(X_i;Y_i\mid Z_i)+\tilde{I}(X_i;Z_i\mid W_i)\Big)}_{:=A}-\underbrace{\sum_{i=1}^{k}\Big(\tilde{I}(X_i;Z_i\mid Y_i)+\tilde{I}(W_i;Z_i\mid X_i)\Big)}_{:=B}.
\end{aligned}$$

In other words, we have $A\le B$.

Recall that $W_i - X_i \to Y_i \rightsquigarrow Z_i$ in T^* and hence it implies $I(X_i;Z_i\mid Y_i)=0$. Similarly, we have $I(W_i;Z_i\mid X_i)=0$. By Theorem 5.3.2 and union bound with $\Theta(\frac{1}{\varepsilon}\log\frac{d}{\delta})$ samples, we have

$$\tilde{I}(X_i;Z_i\mid Y_i)\le\frac{\varepsilon}{100C}\qquad\text{and}\qquad\tilde{I}(W_i;Z_i\mid X_i)\le\frac{\varepsilon}{100C}\qquad\text{for all } i=1,\dots,k.$$

Plugging them into each term in B, we can bound B by $2k\cdot\frac{\varepsilon}{100C}=\frac{k\varepsilon}{50C}$. Namely, we have

$$A=\sum_{i=1}^{k}\Big(\tilde{I}(X_i;Y_i\mid Z_i)+\tilde{I}(X_i;Z_i\mid W_i)\Big)\le\frac{k\varepsilon}{50C}. \tag{C.12}$$

By Theorem 5.3.2 (note that Y_i,Z_i are a descendant of X_i and hence it satisfies the extra conditions in the statement) and union bound with $\Theta(\frac{1}{\varepsilon}\log\frac{d}{\delta})$ samples, we have

$$\frac{\varepsilon}{50C}\le\tilde{I}(X_i;Y_i\mid Z_i)\quad\text{and}\quad\frac{\varepsilon}{50C}\le\tilde{I}(X_i;Z_i\mid W_i)$$

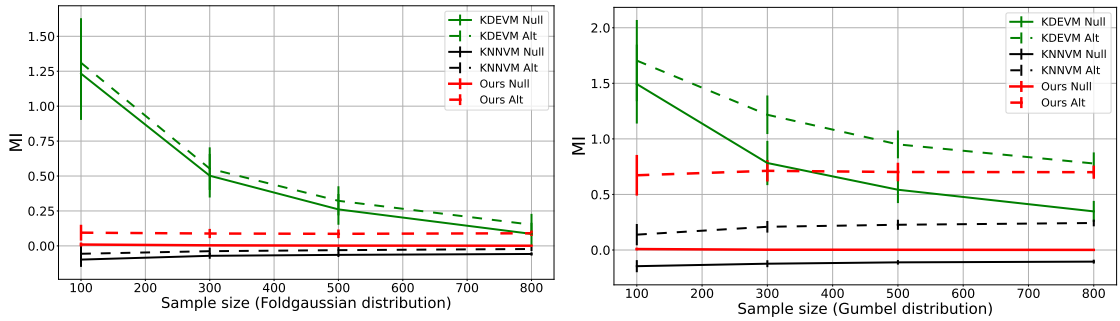
for all $i = 1, \ldots, k$. In other words,

$A \geq \dfrac{k\varepsilon}{25C}$ which contradicts Equation C.12 unless $k = 0$, i.e. the skeleton of $T^* = \widehat{T}$.

□

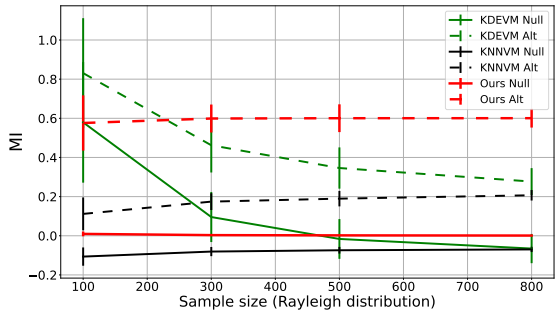
C.3 Experiments

Mutual Information Tester Additionally, we conducted more experiments with our mutual information tester, as shown in Fig. C.1 (a)–(d), when the data distributions are in $\mathcal{G}$, and (e)–(f), where the distributions are not in $\mathcal{G}$.

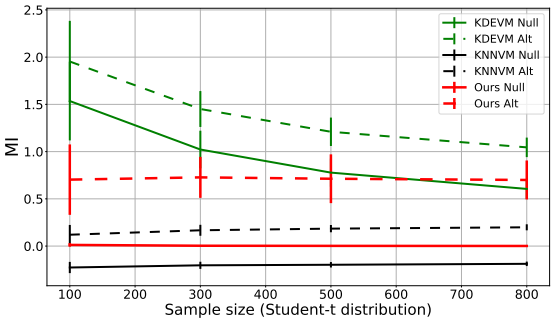


(a) Foldgaussian distribution

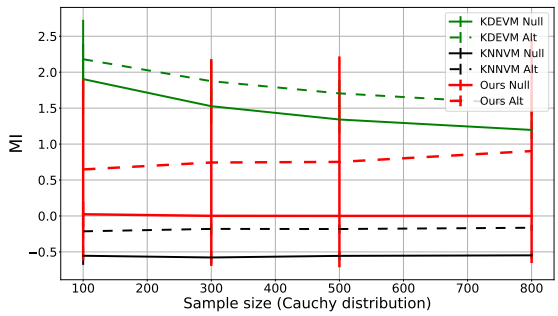
(b) Gumbel distribution



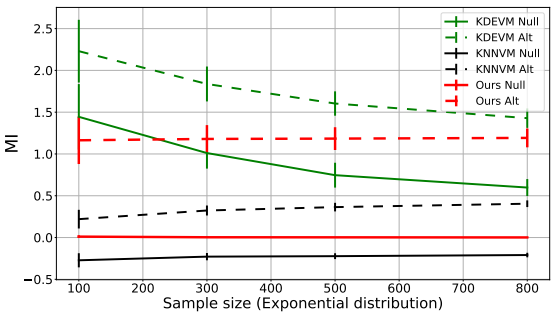
(c) Rayleigh distribution



(d) Student-t distribution



(e) Cauchy distribution



(f) Exponential distribution

Figure C.1: MI tester on null (solid line) and alternative (dashed line) hypothesis. The red, green, and black lines represent our methods, KDEVM, and KNNVM, respectively. For plots (a) - (d), the distributions are in $\mathcal{G}$, while (e) and (f) are not in $\mathcal{G}$.

Appendix D

Supplementary Material - Chapter 6

D.1 Omitted proofs from Section 6.2

We will require the below well-known results for the statistic Z, show in [Can+21, Lemma 4.1]. For completeness, we provide the proof below:

Lemma 2. *For the random variable Z defined in Eq. (6.1), obtained from two independent sets of n samples (i.e. $2n$ total samples) from P, the following holds:*

$$\mathbb{E}[Z] = \langle \mathbb{E}[\bar{\mathbf{X}}], \mathbb{E}[\bar{\mathbf{Y}}] \rangle = \|\mu_S\|_2^2 \tag{6.2}$$

$$\mathrm{Var}[Z] \leqslant \frac{\|\mathbf{\Sigma}_S\|_F^2}{n^2} + \frac{2}{n}\|\mathbf{\Sigma}_S\|_F \|\mu_S\|_2^2 \tag{6.3}$$

Proof. We will prove this for any d-dimensional distribution $X \sim P$. Suppose the $\mu = \mathbb{E}[X]$ and denote Σ its covariance matrix. Draw $\mathbf{X} = \{x^{(1)}, \ldots, x^{(n)}\}$, $\mathbf{Y} = \{y^{(1)}, \ldots, y^{(n)}\}$ i.i.d. $2n$ samples from P; let $\bar{\mathbf{X}} := \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i$, $\bar{\mathbf{Y}} := \frac{1}{n}\sum_{i=1}^n \mathbf{Y}_i$.

$$Z = \langle \bar{\mathbf{X}}, \bar{\mathbf{Y}} \rangle = \frac{1}{n^2}\sum_{i=1}^{d}\sum_{k=1}^{n}\sum_{l=1}^{n} X_{l,i} Y_{k,i}.$$

$$\mathbb{E}[Z] = \|\mu\|_2^2.$$

The proof follows from the fact that $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$ are independent, thus $\mathbf{X}_i \perp\!\!\!\perp \mathbf{X}_j$ regardless of i and j. Note that $\mathrm{Var}[Z] = \mathbb{E}[Z^2] - \mathbb{E}^2[Z]$, we start by computing the second moment of the statistic:

$$\begin{aligned}
\mathbb{E}[Z^2] &= \mathbb{E}\left[\left(\frac{1}{n^2}\sum_{i=1}^{d}\sum_{k=1}^{n}\sum_{l=1}^{n} X_{l,i}Y_{k,i}\right)\left(\frac{1}{n^2}\sum_{j=1}^{d}\sum_{k'=1}^{n}\sum_{l'=1}^{n} X_{l',j}Y_{k',j}\right)\right] \\
&= \frac{1}{n^4}\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{n}\sum_{l=1}^{n}\sum_{k'=1}^{n}\sum_{l'=1}^{n} \mathbb{E}[X_{l,i}Y_{k,i}X_{l',j}Y_{k',j}] \\
&= \frac{1}{n^4}\sum_{i=1}^{d}\sum_{j=1}^{d}\left(\sum_{l=1}^{n}\sum_{l'=1}^{n}\mathbb{E}[X_{l,i}X_{l',j}]\sum_{k=1}^{n}\sum_{k'=1}^{n}\mathbb{E}[Y_{k,i}Y_{k',j}]\right) \\
&= \frac{1}{n^4}\sum_{i=1}^{d}\sum_{j=1}^{d}\left(\sum_{l=1}^{n}\sum_{l'=1}^{n}\mathbb{E}[X_{l,i}X_{l',j}]\right)\left(\sum_{k=1}^{n}\sum_{k'=1}^{n}\mathbb{E}[Y_{k,i}Y_{k',j}]\right) \\
&= \frac{1}{n^4}\sum_{i=1}^{d}\sum_{j=1}^{d}\left(\sum_{l=1}^{n}\sum_{l'=1}^{n}\mathbb{E}[X_{l,i}X_{l',j}]\right)^2 \\
&= \frac{1}{n^4}\sum_{i=1}^{d}\sum_{j=1}^{d}\left(\sum_{l=1}^{n}\mathbb{E}[X_{l,i}X_{l,j}] + \sum_{l\neq l'}\mathbb{E}[X_{l,i}]\mathbb{E}[X_{l',j}]\right)^2
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n^4}\sum_{i=1}^{d}\sum_{j=1}^{d}\left(\sum_{l=1}^{n}\operatorname{Cov}(X_{l,i},X_{l,j}) + \mathbb{E}[X_{l,i}]\mathbb{E}[X_{l,j}] + n(n-1)\mu_i\mu_j\right)^2 \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d}\left(\frac{1}{n}\operatorname{Cov}(X_i,X_j) + \mu_i\mu_j\right)^2 \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d}\left(\frac{1}{n^2}\Sigma_{i,j}^2 + \frac{2}{n}\Sigma_{i,j}\mu_i\mu_j + \mu_i^2\mu_j^2\right) \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d}\left(\frac{1}{n^2}\Sigma_{i,j}^2 + \frac{2}{n}\Sigma_{i,j}\mu_i\mu_j\right) + \|\mu\|_2^4
\end{aligned}
$$

Then substitute Eq. (6.2) to complete the computation of Eq. (6.3)

$$
\begin{aligned}
\operatorname{Var}[Z] = \mathbb{E}[Z^2] - \mathbb{E}^2[Z] &= \sum_{i=1}^{d}\sum_{j=1}^{d}\left(\frac{1}{n^2}\Sigma_{i,j}^2 + \frac{2}{n}\Sigma_{i,j}\mu_i\mu_j\right) \\
&= \frac{1}{n^2}\sum_{1\leqslant i,j\leqslant d}\Sigma_{i,j}^2 + \frac{2}{n}\sum_{i,j}\Sigma_{i,j}\mu_i\mu_j \\
&= \frac{\|\Sigma\|_F^2}{n^2} + \frac{2}{n}\sum_{i,j}\Sigma_{i,j}\mu_i\mu_j \\
&\leqslant \frac{\|\Sigma\|_F^2}{n^2} + \frac{2}{n}\sqrt{\sum_{i,j}\Sigma_{i,j}^2}\sqrt{\sum_{i,j}\mu_i^2\mu_j^2} \quad \text{(By Cauchy-Schwarz)} \\
&= \frac{\|\Sigma\|_F^2}{n^2} + \frac{2}{n}\|\Sigma\|_F\cdot\|\mu\|_2^2 \qquad \square
\end{aligned}
$$

Lemma 3 (Truncated vs non-truncated parameters)**.** *Let $\mu_S, \mathbf{\Sigma}_S$ be the mean and covariance of the truncated Gaussian $\mathcal{N}(\mu, \mathbf{I}_d; S)$ with a measure of at least $1-\varepsilon$. Then the following holds:*

$$\|\mu_S - \mu\|_2 \leq \mathcal{O}(\varepsilon\cdot\sqrt{\log(1/\varepsilon)}) \text{ and } \|\mathbf{\Sigma}_S - \mathbf{I}_d\|_F \leqslant \mathcal{O}(\sqrt{d}).$$

Proof. We establish each statement separately.

Bound on the mean $\|\mu_S - \mu\|_2$: Consider the region $\bar{S}$, which contributes the most to the change in the mean or covariance matrix in terms of the Frobenius norm. Let $\mathbf{v}_S$ be the unit vector in the direction of the truncated mean μ_S. For any unit vector $\mathbf{v}$, the region that impacts the expectation $\mathbb{E}[\mathbf{v}^T\mathbf{x}]$ or $\operatorname{Var}[\mathbf{v}^T\mathbf{x}]$ the most corresponds to truncating the ε-tail of $\mathbf{v}^T\mathbf{x}$. The change in the mean in this direction can be bounded by $\mathcal{O}(\varepsilon\sqrt{\log(1/\varepsilon)})$, and similarly, the variance of $\mathbf{v}^T\mathbf{x}$ changes by at most $\mathcal{O}(\varepsilon\log(1/\varepsilon))$, as can be shown by relatively standard and elementary computations on a single-dimensional standard Gaussian. Thus, for the

mean shift, we have

$$\|\mu_S - \mu\|_2 = \|\mu_S\| = \mathcal{O}(\varepsilon\sqrt{\log(1/\varepsilon)})$$

Even if the region $\bar{S}$ fully truncates its ε mass in the direction of $\mathbf{v}_S$, the mean shift in that direction is at most $\mathcal{O}(\varepsilon\sqrt{\log(1/\varepsilon)})$.

Bound on the covariance $\|\mathbf{\Sigma}_S - \mathbf{I}_d\|_F$: Next, we turn to the covariance matrix. For any unit vector $\mathbf{v}$, the variance of $\mathbf{v}^T\mathbf{x}$ in the truncated distribution can be expressed as:

$$\mathrm{Var}[\mathbf{v}^T\mathbf{x}] = \mathbb{E}[(\mathbf{v}^T\mathbf{x})^2] - (\mathbb{E}[\mathbf{v}^T\mathbf{x}])^2.$$

For the truncated Gaussian, the variance of $\mathbf{v}^T\mathbf{x}$ differs from 1 by at most $\mathcal{O}(\varepsilon\log(1/\varepsilon))$, i.e.,[1],

$$\mathrm{Var}[\mathbf{v}^T\mathbf{x}] - 1 = \mathbb{E}[(\mathbf{v}^T\mathbf{x})^2] - \mathbb{E}[\mathbf{v}^T\mathbf{x}]^2 - \mathbf{v}^T\mathbf{I}_d\mathbf{v}$$

where

$$\mathbb{E}[\mathbf{v}^T\mathbf{x}]^2 = (\mathbf{v}^T\mu_S)\cdot(\mu_S^T\mathbf{v})$$
$$\mathbb{E}[(\mathbf{v}^T\mathbf{x})(\mathbf{v}^T\mathbf{x})^T] - \mathbf{v}^T\mathbf{I}_d\mathbf{v} = \mathbf{v}^T(\mathbb{E}[\mathbf{x}\mathbf{x}^T] - \mathbf{I}_d)\mathbf{v} = \mathbf{v}^T(\Sigma_S + \mu_S\mu_S^T - I_d)\mathbf{v}$$

Thus

$$|\,\mathrm{Var}[\mathbf{v}^T\mathbf{x}] - 1| = |\mathbf{v}^T(\Sigma_S - I_d)\mathbf{v}| \leqslant \mathcal{O}(\varepsilon\log 1/\varepsilon).$$

Now, recall the relationship between the spectral norm and the Frobenius norm: if the spectral norm of $\mathbf{\Sigma}_S - \mathbf{I}_d$ is bounded by $\mathcal{O}(\varepsilon\log(1/\varepsilon)) = \mathcal{O}(1)$, then the Frobenius norm satisfies $\mathcal{O}(\sqrt{d})$.

□

Theorem 6.2.1. *There exists an algorithm (Algorithm 6) that, given i.i.d. samples from truncated Gaussian distribution P with an unknown support set $S \subset \mathbb{R}^d$, can distinguish the following two cases based on the truncation mass parameter $\varepsilon \in (0,1)$ and the accuracy parameter $\alpha > 0$:*

- ***(Completeness)** If P is a truncated Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_d; S)$ and the truncation mass satisfies $1 - \mathcal{N}(0, \mathbf{I}_d; S) \leqslant \varepsilon$, the algorithm will output "ACCEPT" with probability at least $2/3$.*

[1]We believe one can prove a bound of $\mathcal{O}(\varepsilon\log 1/\varepsilon)$ with a more sophisticated analysis; however, this weaker bound suffices for our purposes.

- ***(Soundness)*** *If P is a truncated Gaussian distribution $\mathcal{N}(\mu, \mathbf{I}_d; S)$ where $\|\mu\|_2 \geqslant \alpha \geqslant c_1 \cdot \varepsilon\sqrt{\log \frac{1}{\varepsilon}}$ for some constant $c_1 > 0$ and the truncation mass satisfies $1 - \mathcal{N}(0, \mathbf{I}_d; S) \leqslant \varepsilon$, the algorithm will output "REJECT" with probability at least $2/3$.*

The algorithm requires $\mathcal{O}\left(\frac{\sqrt{d}}{\alpha^2}\right)$ samples from P.

Proof. When $\mathbf{X}, \mathbf{Y}$ come from the truncated Gaussian distribution $\mathcal{N}(\mu, \mathbf{I}_d, S)$, by Lemma 2, we know that for random variable Z

$$Z = \langle \bar{\mathbf{X}}, \bar{\mathbf{Y}} \rangle,$$

the following condition holds:

$$\begin{aligned}
\mathbb{E}[Z] &= \langle \mathbb{E}[\overline{\mathbf{X}}], \mathbb{E}[\overline{\mathbf{Y}}] \rangle = \|\mu\|_2^2 \\
\mathrm{Var}[Z] &\leqslant \frac{\|\Sigma_S\|_F^2}{n^2} + \frac{2}{n}\|\Sigma_S\|_F \|\mu_S\|_2^2.
\end{aligned}$$

By Lemma 3,

$$\|\mu_S - \mu\|_{I_d} \leqslant \mathcal{O}\left(\varepsilon \cdot \sqrt{\log(1/\varepsilon)}\right) \text{ and } \|\Sigma_S - \mathbf{I}_d\|_F \leqslant \mathcal{O}(\sqrt{d}). \tag{D.1}$$

Completeness: In the completeness case, we want to show the expectation of Z is small when $\|\mu\|_2 = 0$.

$$\begin{aligned}
\mathbb{E}[Z] &= \|\mu_S\|_2^2 \leqslant O(\varepsilon^2 \cdot \log(1/\varepsilon)) \leqslant O(c_1^2 \cdot \alpha^2) = O(\alpha^2). \\
\mathrm{Var}[Z] &\leqslant \frac{\|\Sigma_S\|_F^2}{n^2} + \frac{2}{n}\|\Sigma_S\|_F \|\mu_S\|_2^2 \\
&\leqslant \frac{(\|\Sigma_S - \mathbf{I}_d\|_F + \|\mathbf{I}_d\|_F)^2}{n^2} + \frac{2}{n}(\|\Sigma_S - \mathbf{I}_d\|_F + \|\mathbf{I}_d\|_F)(\|\mu_S - \mu\|_2 + \|\mu\|_2)^2 \\
&\lesssim \frac{(\sqrt{d} + \|\mathbf{I}_d\|_F)^2}{n^2} + \frac{2}{n}(\sqrt{d} + \|\mathbf{I}_d\|_F)\left(\varepsilon \cdot \sqrt{\log(1/\varepsilon)} + 0\right)^2 \\
&= O\left(\frac{d}{n^2}\right) + O\left(\frac{\sqrt{d}}{n}\varepsilon^2 \log 1/\varepsilon\right) \\
&\lesssim \underbrace{\frac{d}{n^2}}_{\ll \alpha^4} + \underbrace{\frac{\sqrt{d}}{n} \cdot \alpha^2}_{\ll \alpha^4}
\end{aligned}$$

$$\lesssim \alpha^4$$

Since $n \gtrsim \frac{\sqrt{d}}{\alpha^2}$ and $\alpha \gtrsim \varepsilon\sqrt{\log(1/\varepsilon)}$, both terms are much smaller than α^4. By Chebyshev's inequality, we have:

$$\Pr\left[Z - \|\mu_S\|_2^2 \geqslant \frac{1}{2}\|\mu_S\|_2^2\right] \leqslant \frac{4\,\mathrm{Var}[Z]}{\mathbb{E}^2[Z]}$$

Using the bounds on $\mathbb{E}[Z]$ and $\mathrm{Var}\,[Z]$, this gives:

$$\Pr[Z \geqslant \Omega(\alpha^2)] \leqslant \frac{1}{9}.$$

Thus, the algorithm outputs "ACCEPT" with high probability in the completeness case.

Soundness: In the soundness case, we have that $\|\mu\|_2 \geqslant c_1\alpha \geqslant \varepsilon \cdot \sqrt{\log(1/\varepsilon)}$. We now show that Z is large in this case:

$$\mathbb{E}[Z] = \|\mu_S\|_2^2 \geqslant (\|\mu\|_2 - \|\mu-\mu_S\|_2)^2 = \left(\|\mu\|_2 - \mathcal{O}\left(\varepsilon \cdot \sqrt{\log\frac{1}{\varepsilon}}\right)\right)^2 \geqslant (\|\mu\|_2 - \mathcal{O}(\alpha))^2 \geqslant \Omega(\|\mu\|_2^2). \tag{D.2}$$

Similarly, the variance of Z is bounded as:

$$\begin{aligned}
\mathrm{Var}[Z] &\leqslant \frac{\|\Sigma_S\|_F^2}{n^2} + \frac{2}{n}\|\Sigma_S\|_F\|\mu_S\|_2^2 \\
&\leqslant \frac{(\|\Sigma_S - \mathbf{I}_d\|_F + \|\mathbf{I}_d\|_F)^2}{n^2} + \frac{2}{n}(\|\Sigma_S - \mathbf{I}_d\|_F + \|\mathbf{I}_d\|_F)\|\mu_S\|_2^2 \\
&\lesssim \frac{(\sqrt{d}+\sqrt{d})^2}{n^2} + \frac{2}{n}(\sqrt{d}+\sqrt{d})\|\mu_S\|_2^2 \\
&= O\left(\frac{d}{n^2}\right) + O\left(\frac{\sqrt{d}}{n}\|\mu_S\|_2^2\right) \\
&\lesssim \frac{d}{n^2} + \frac{\sqrt{d}}{n}\mathbb{E}[Z] \lesssim \mathbb{E}[Z]^2
\end{aligned}$$

using that $n \geqslant \Omega\left(\frac{\sqrt{d}}{\alpha^2}\right)$ and recalling that $\mathbb{E}[Z] = \|\mu_S\|_2^2 \geqslant \Omega(\alpha^2)$ via Equation D.2 in the last step. By Chebyshev's inequality, we get:

$$\Pr\left[\|\mu_S\|_2 - Z \geqslant \frac{1}{2}\mathbb{E}[Z]\right] \leqslant \frac{4\,\mathrm{Var}[Z]}{\mathbb{E}[Z]^2} \Rightarrow \Pr[Z \leqslant O(\|\mu_S\|_2)] \leqslant \frac{1}{9}$$

Thus, with high probability:

$$\Pr[Z \leqslant O(\alpha^2)] \leqslant \frac{1}{9}.$$

Hence, the algorithm outputs "REJECT" with high probability in the soundness case. □

Theorem 6.2.2. *The sample complexity for truncated mean testing when* $\varepsilon \lesssim \alpha \lesssim \varepsilon \cdot \sqrt{\log \frac{1}{\varepsilon}}$ *is* $\Theta(d)$.

Proof. This is a consequence of sample complexity lower bound of $\Omega(d)$ from Lemma 4 and the robust mean estimation [DK23, Proposition 1.20] upper bound of $O(d)$. □

D.2 Proof of Theorem 6.3.1

Lemma 5. *Let* Z_1 *be the statistics in Algorithm 7 Line 4, and* $\mu'_S = \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d,S)}[\mathbf{x}]$ *(truncated mean under zero mean). Let* μ_S *be the truncated mean of the unknown Gaussian* P*, we can show that*

$$\mathbb{E}[Z_1] = \|\mu_S - \mu'_S\|_2^2.$$

$$\mathrm{Var}[Z_1] \leqslant O(\alpha^4 + \alpha^2 \cdot \|\mu_S - \mu'_S\|_2^2).$$

Proof. By linearity of expectation and independence between X_is and Y_is,

$$\mathbb{E}[Z_1] = \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu'_S\right)^T\right]\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n Y_i - \mu'_S\right)\right] = (\mu - \mu'_S)^T(\mu - \mu'_S).$$

Think of $\tilde{X}_i = X_i - \mu'_S$ as a random variable ($\tilde{Y}_i = Y_i - \mu'_S$ as the other random variable), denote $\tilde{\Sigma}$ the covariance of a single $\tilde{X}_i$, and $\tilde{\mu}$ the mean. We have that as in the proof of Lemma 2,

$$\mathrm{Var}\left[\left(\frac{1}{n}\sum_{i=1}^n \tilde{X}_i\right)^T\left(\frac{1}{n}\sum_{i=1}^n \tilde{Y}_i\right)\right] \leqslant \frac{\|\tilde{\Sigma}\|_F^2}{n^2} + \frac{2}{n}\|\tilde{\Sigma}\|_F\|\tilde{\mu}\|_2^2.$$

We know that shifting the location of a random variable does not affect the covariance matrix, and thus $\tilde{\Sigma} = \Sigma_S$, and $\tilde{\mu} = \mu_S - \mu'_S$, which means

$$\begin{aligned}
\mathrm{Var}[Z_1] &\leqslant \frac{\|\Sigma_S\|_F^2}{n^2} + \frac{2}{n}\|\Sigma_S\|_F\|\mu_S - \mu'_S\|_2^2 \\
&\leqslant \frac{(\|\Sigma_S - \mathbf{I}_d\|_F + \|\mathbf{I}_d\|_F)^2}{n^2} + \frac{2}{n}(\|\Sigma_S - \mathbf{I}_d\|_F + \|\mathbf{I}_d\|_F)\|\mu_S - \mu'_S\|_2^2 \\
&\lesssim \frac{d}{n^2} + \frac{\sqrt{d}}{n}\|\mu_S - \mu'_S\|_2^2,
\end{aligned}$$

where the last step follows from Lemma 3. Letting $n = O\left(\frac{\sqrt{d}}{\alpha^2}\right)$, we conclude our proof. □

Lemma 6 (Gap of Mean under Truncation)**.** *Let* $\mathbb{E}_{\mathbf{y}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d;S)}[\mathbf{y}] = \mu'_S$ *and* $\mathbb{E}_{\mathbf{y}\sim\mathcal{N}(\mu'',\mathbf{I}_d;S)}[\mathbf{x}] = \mu''_S$*, where* $\|\mu''\|_2^2 \geq \alpha^2$*. Additionally, assume that* $\mathcal{N}(\mu'', \mathbf{I}_d; S) \geq 1-\beta$ *for some constant* β*. Then, it holds that*

$$\|\mu'_S - \mu''_S\|_2^2 \geq \Omega(\alpha^2).$$

Proof. Consider the negative log-likelihood function, $\bar{\ell}(\mathbf{0})$, with the mean set to $\mathbf{0}$ as the input parameter. This function is defined for a population drawn from a truncated normal distribution with an unknown mean μ. From Equation 2.4, we can express the gradient of the negative log-likelihood with respect to the mean evaluated at $\mathbf{0}$, as follows:

$$\nabla\bar{\ell}(\mathbf{0}) = -\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mu,\mathbf{I}_d,S)}[\mathbf{x}] + \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d,S)}[\mathbf{z}] = \mu_S - \mu'_S.$$

Likewise, when evaluating the gradient at μ, we have

$$\nabla\bar{\ell}(\mu) = -\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mu,\mathbf{I}_d,S)}[\mathbf{x}] + \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mu,\mathbf{I}_d,S)}[\mathbf{z}] = \mathbf{0}.$$

So, $\nabla\bar{\ell}(\mathbf{0})$ represents the difference between the truncated mean of the underlying distribution and that of the distribution with mean $\mathbf{0}$. From Lemma 1, let $\lambda_0 = \frac{1}{2^{13}}\left(\frac{\beta}{C}\right)^4 \min\left\{\frac{1}{4}, \frac{1}{16\|\mu\|_2^2+1}\right\}$, we know that $\bar{\ell}(\cdot)$ is λ_0-strongly convex, and λ_0 is a constant if β is a constant and $\|\mu\|_2^2 \leqslant \frac{1}{16}$. Therefore, by leveraging the properties of strong convexity and applying the CauchySchwarz inequality, we obtain the following result:

$$\sqrt{\|\mu-\mathbf{0}\|_2^2 \cdot \|\nabla\bar{l}(\mu) - \nabla\bar{l}(\mathbf{0})\|_2^2} \geqslant \langle\nabla\bar{l}(\mu) - \nabla\bar{l}(\mathbf{0}), \mu-\mathbf{0}\rangle \geqslant \frac{\lambda_0}{2}\|\mu\|_2^2$$

By simplifying the expression and substituting μ with any $\|\mu''\|_2^2 \geqslant \alpha^2$, we can show that:

$$\|\mu''_S - \mu'_S\|_2^2 \geqslant \Omega(\alpha^2).$$

□

Theorem 6.3.1 (Known truncation tester)**.** *There exists an algorithm (Algorithm 7) that takes i.i.d. samples from truncated normal Gaussian* P *and given oracle access to* $S \subset \mathbb{R}^d$*, the effective support of* P*, distinguishing the cases for parameters (mass of truncation)* $0 < \varepsilon \leqslant 1-\beta$*, where* β *is a constant and (accuracy)* $\frac{1}{4} \geqslant \alpha > 0$*:*

- ***(Completeness)*** *P is a truncated Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_d, S)$ and $1 - \mathcal{N}(\mathbf{0}, \mathbf{I}_d; S) \leqslant \varepsilon$. In this case, the algorithm will output yes with probability at least $2/3$.*
- ***(Soundness)*** *P is a truncated Gaussian distribution $\mathcal{N}(\mu, \mathbf{I}_d, S)$ where $\|\mu\|_2 \geqslant \alpha$ and $1 - \mathcal{N}(\mathbf{0}, \mathbf{I}_d; S) \leqslant \varepsilon$. In this case, the algorithm will output no with probability at least $2/3$.*

The algorithm will take $\mathcal{O}\left(\frac{\sqrt{d}}{\alpha^2}\right)$ samples from P.

Proof. Suppose $\|\mu\|_2^2 = \alpha^2 \leqslant \frac{1}{16}$, then we have

$$\min\left\{\frac{1}{4}, \frac{1}{16\|\mu\|_2^2 + 1}\right\} = \frac{1}{4}.$$

And so λ_0 is a constant.

Completeness: When P is a truncated Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I}_d, S)$. By Lemma 5, we have that when $n \geqslant \Omega\left(\frac{\sqrt{d}}{\lambda_0^2\alpha^2}\right) = \Omega\left(\frac{\sqrt{d}}{\alpha^2}\right)$,

$$\mathbb{E}[Z_1] = \mathbf{0} \text{ and } \operatorname{Var}[Z_1] \leqslant O(\alpha^4).$$

By Chebyshev's inequality, for n being a large enough multiple of $\frac{\sqrt{d}}{\alpha^2}$, we have that,

$$\Pr\left[Z_1 \geqslant \frac{3}{2}c_2 \cdot \alpha^2\right] \leqslant O\left(\frac{\operatorname{Var}[Z_1]}{c_2 \cdot \alpha^4}\right) \leqslant \frac{1}{10}.$$

Soundness: When P is a truncated Gaussian $\mathcal{N}(\mu, \mathbf{I}_d, S)$ and $\|\mu\|_2^2 \geqslant \alpha^2$. By Lemma 6, we have that the gap under truncation is:

$$\|\mu_S - \mu'_S\|_2^2 \geqslant \Omega(\alpha^2) = c_2 \cdot \alpha^2.$$

By Lemma 5, we have that when $n \geqslant \Omega\left(\frac{\sqrt{d}}{\lambda_0^2\alpha^2}\right) = \Omega\left(\frac{\sqrt{d}}{\alpha^2}\right)$,

$$\mathbb{E}[Z_1] = \|\mu - \mu_S\|_2^2 \text{ and } \operatorname{Var}[Z_1] \leqslant O(\alpha^4 + \alpha^2\|\mu_S - \mu'_S\|_2^2).$$

By Chebyshev's inequality, we have that,

$$\Pr\left[Z_1 \leqslant \frac{3}{2}c_1 \cdot \alpha^2\right] \leqslant \Pr\left[Z_1 \leqslant \|\mu - \mu_S\|_2^2 + \frac{1}{2}c_2 \cdot \alpha^2\right] \leqslant \frac{\operatorname{Var}[Z_1]}{\left(\frac{1}{2}c_2 \cdot \alpha^2\right)^2} \leqslant O\left(\frac{\alpha^4 + \alpha^2\|\mu_S - \mu'_S\|_2^2}{c_2^2 \cdot \alpha^4}\right).$$

Let n be a large enough multiple of $\frac{\sqrt{d}}{\alpha^2}$, then

$$\Pr\left[Z_1 \leqslant \frac{3}{2}c_1 \cdot \alpha^2\right] \leqslant O\left(\frac{\alpha^4 + \alpha^2\|\mu_S - \mu'_S\|_2^2}{c_2^2 \cdot \alpha^4}\right) = O\left(\frac{\alpha^4 + c_2 \cdot \alpha^4}{c_2^2 \cdot \alpha^4}\right) \leqslant \frac{1}{10}.$$

□

D.3 Proof of Lemma 4

Lemma 4 (Sample Complexity Lower Bound for Mean Testing with Unknown Truncation When $\varepsilon \lesssim \alpha \lesssim \varepsilon\sqrt{\log(1/\varepsilon)}$). *No algorithm can distinguish between* $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ *and a family of truncated normal distribution of the form:* $\mathcal{N}(\mathbf{v}, \mathbf{I}_d, S)$ *with measure* ε *on the truncation set* $\bar{S} = \mathbb{R}^d \backslash S$*, for any* $\varepsilon < 1$ *and some* $\|\mathbf{v}\|_2 = \alpha = \Theta(\varepsilon\sqrt{\log(1/\varepsilon)})$*, using fewer than* $\Omega(d/\varepsilon)$ *samples with a probability greater than* $2/3$.

Proof. We begin by constructing a one-dimensional truncated normal distribution $A = \mathcal{N}(\alpha, 1, S)$, where the truncated mass is ε. This means $\Pr_{x\sim\mathcal{N}(\alpha,1)}[x \in S] = 1-\varepsilon$. We can determine the $(1-\varepsilon)$-quantile as:

$$b = \alpha + \sqrt{2}\,\mathrm{erf}^{-1}(1-2\varepsilon).$$

which defines the truncation set as $S := (-\infty, b]$.

Let $\alpha(\varepsilon) := \alpha = \Theta\left(\varepsilon\sqrt{\log\frac{1}{\varepsilon}}\right)$. For any ε, we can find a constant $c_2 = \Theta(1)$ such that $\mathbb{E}[A] = 0$:

$$\mathbb{E}_{X\sim A}[X] = \alpha - \frac{\exp\left(-\frac{1}{2}(b-\alpha)^2\right)}{\sqrt{2\pi}(1-\varepsilon)} = 0,$$

which is equivalent to:

$$\alpha = \frac{\exp(-(\mathrm{erf}^{-1}(1-2\varepsilon))^2)}{\sqrt{2\pi}(1-\varepsilon)} = \Theta\left(\varepsilon\sqrt{\log\frac{1}{\varepsilon}}\right).$$

Next, we compute an upper bound on the chi-squared divergence between the truncated distribution A and the standard normal distribution $\mathcal{N}(0,1)$. We find that

$$\begin{aligned}
\chi^2(A, \mathcal{N}(0,1)) &= \left(\int_{-\infty}^{b} \frac{\left(\exp\left(-\frac{(x-\alpha)^2}{2}\right)/\sqrt{2\pi}(1-\varepsilon)\right)^2}{\exp\left(-\frac{x^2}{2}\right)/\sqrt{2\pi}} dx\right) - 1 \\
&= \frac{1}{\sqrt{2\pi}(1-\varepsilon)^2}\left(\int_{-\infty}^{b} \exp(-(x-\alpha)^2 + x^2/2) dx\right) - 1 \\
&= \frac{1}{\sqrt{2\pi}(1-\varepsilon)^2}\left(\int_{-\infty}^{b} \exp\left(-\frac{x^2}{2} + 2x\alpha - \alpha^2\right) dx\right) - 1 \\
&= \frac{1}{\sqrt{2\pi}(1-\varepsilon)^2}\left(\int_{-\infty}^{b} \exp\left(-\left(\frac{x}{\sqrt{2}}\right)^2 + 2x\alpha - \left(\sqrt{2}\alpha\right)^2 + \alpha^2\right) dx\right) - 1 \\
&= \frac{\exp(\alpha^2)}{\sqrt{2\pi}(1-\varepsilon)^2}\left(\int_{-\infty}^{b} \exp\left(-\left(\frac{x}{\sqrt{2}} - \sqrt{2}\alpha\right)^2\right) dx\right) - 1
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\exp(\alpha^2)}{\sqrt{2\pi}(1-\varepsilon)^2}\left(\int_{-\infty}^{b}\exp\left(-\frac{(x-2\alpha)^2}{2}\right)dx\right)-1\\
&= \frac{\mathcal{N}(2\alpha,1;S)}{(1-\varepsilon)^2}\cdot\exp(\alpha^2)-1\\
&\leqslant \frac{\exp(\alpha^2)}{(1-\varepsilon)^2}-1\\
&\leqslant (1+O(\varepsilon))\cdot(1+O(\alpha^2))-1\\
&\leqslant O(\varepsilon)+O(\alpha^2).
\end{aligned}
$$

We now apply Proposition 2.4.1 [DKS16, Proposition 7.1], and obtain a lower bound of

$$\Omega\left(\frac{d}{\varepsilon+\alpha^2}\right)=\Omega\left(\frac{d}{\varepsilon}\right).$$

concluding the proof. □

Appendix E

Supplementary Material - Chapter 7

E.1 Deferred Proofs from section 7.2

Below we provide the proof of Fact fact 7.2.1 for completeness.

Fact 7.2.1. *Let $A = \{a_{ij}\}$ and $B = \{b_{ij}\}$ be two $d \times d$ matrices such that $\forall i, j : |a_{ij} - b_{ij}| \leq \delta$. Then, $\|A - B\|_F \leq \delta \cdot d$.*

Proof. By definition of the Frobenious norm we have:

$$\|A - B\|_F^2 = \sum_{i=1}^{d} \sum_{j=1}^{d} (a_{ij} - b_{ij})^2 \leq d^2 \delta^2$$

Thus,

$$\|A - B\|_F \leq \delta \cdot d$$

□

E.2 Proofs from Section 7.3

This section provides the formal proofs that were deferred in favor of readability. For convenience, we will restate the statements before proving them.

Lemma 7. *Let $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ be the normal distribution with mean $\boldsymbol{\mu}^*$ and covariance matrix $\boldsymbol{\Sigma}^*$. Suppose that assumption 7.1.1 holds for some constant $\alpha > 0$, and let $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_d)$ be the estimated mean from the censored Gaussian in Line 6 of Algorithm 1. For all $\varepsilon > 0$, using $\tilde{\mathcal{O}}(\frac{d^2}{\alpha\varepsilon^2})$ samples[1] we have that:*

$$\forall i \in [d] : |\mu_i^* - \hat{\mu}_i| \leq (\varepsilon/d)\sigma_i \leq (\varepsilon/d)\sqrt{\lambda_{max}(\boldsymbol{\Sigma})}$$

where σ_i denotes the standard deviation of coordinate i (i.e $\sigma_i = \sqrt{\Sigma_{ii}^*}$, where Σ_{ii}^* is the i-th diagonal entry of the covariance matrix $\boldsymbol{\Sigma}^*$).

Proof. Fix a coordinate $i \in [d]$. If coordinate i appears in a censored sample, then the value would follow the distribution $\mathcal{N}(\mu_i^*, \Sigma_{ii}^*) = \mathcal{N}(\mu_i^*, \sigma_i^2)$. In order to apply Theorem 7.3.1 for $d = 1$ and $\varepsilon' = \varepsilon/d$, we need coordinate i to be present in at least $\tilde{O}(1/\varepsilon^2)$ censored samples for every $i \in [d]$. assumption 7.1.1 implies that coordinate i is present in each censored sample with probability at least α. Since in every batch

[1] We note that the $\tilde{\mathcal{O}}_\alpha$ notation here hides both $\log d$ and $\log(1/\delta)$ factors.

of $\mathcal{O}(1/\alpha(\varepsilon')^2)$ samples, there is a constant probability that the required number of $\mathcal{O}(1/(\varepsilon')^2)$ appearances of coordinate i is met, the error probability can be reduced to $1/d^2$ a the cost of an extra log factor in the sample complexity. Therefore, by union bound over all d coordinates, the statement holds with probability at least $1 - 1/d$ using $\mathcal{O}(\log d/\alpha(\varepsilon')^2) = \tilde{\mathcal{O}}(\frac{d^2}{\alpha\varepsilon^2})$ samples. □

Lemma 8. *Let $\hat{\mathbf{\Sigma}}$ be the matrix with entries $\hat{\Sigma}_{ij} = \hat{\Sigma}_{12}^{ij}$, where $\hat{\Sigma}_{12}^{ij}$ denotes the value of the off diagonal entries of the 2×2 matrix $\hat{\mathbf{\Sigma}}^{\mathbf{ij}}$.By $\hat{\mathbf{\Sigma}}^{ij}$ we denote the estimation of a 2×2 covariance matrix that we get when we restrict the input data to coordinates i and j. Then the following holds: Using $\tilde{\mathcal{O}}(\frac{d^2}{\alpha\varepsilon^2})$ samples to get the above estimates,we have that:*

$$\|\mathbf{\Sigma}^* - \hat{\mathbf{\Sigma}}\|_F \leq \varepsilon\lambda_{max}$$

where λ_{max} is the maximum eigenvalue of Σ^*

Proof. Consider any pair of coordinates $i, j \in [d]$. Given the sample size of $\frac{1}{\alpha\varepsilon'^2}$, assumption 7.1.1 and fact 7.2.1, there will be at least $\frac{1}{\varepsilon'^2}$ samples with non-censored entries in both coordinates i, j for any such pair. Therefore, we can apply Theorem 7.3.1 for $d = 2$ and $\varepsilon' = \varepsilon/d$ to get that:

$$\|\mathbf{\Sigma}^{*(ij)} - \hat{\mathbf{\Sigma}}^{(\mathbf{ij})}\|_F \leq \varepsilon'\lambda_{max}\left(\mathbf{\Sigma}^{*(ij)}\right) = \varepsilon\lambda_{max}\left(\mathbf{\Sigma}^{*(ij)}\right)/d \leq \varepsilon\lambda_{max}\left(\mathbf{\Sigma}^*\right)/d$$

Therefore, all the entries of the 2×2 matrix on the lhs have absolute value at most ε/d and thus this is also an upper bound on the maximum difference of corresponding off diagonal entries of the $d \times d$ matrices $\mathbf{\Sigma}^*$ and $\hat{\mathbf{\Sigma}}$ as constructed by Algorithm 8 (see line 9). The same upper bound holds for the diagonal entries, since they are more accurately estimated using 1d subproblems (line 2 of Algorithm 8). Therefore, the maximum entry-wise difference overall between the $d \times d$ matrices $\mathbf{\Sigma}^*$ and $\hat{\mathbf{\Sigma}}$ as constructed by Algorithm 8 is $\varepsilon\lambda_{max}/d$. We can now apply fact 7.2.1 to get $\|\mathbf{\Sigma}^* - \hat{\mathbf{\Sigma}}\|_F \leq \varepsilon\lambda_{max}$. □

Lower bound on the sample complexity We now explain why some dependence on the eigenvalues of $\mathbf{\Sigma}^*$ is necessary. To see this, consider the case where $\lambda_{\min}(\mathbf{\Sigma}^*) = 0$. In this case, all the samples come from a subspace of dimension at most $d - 1$. Consider one such subspace and its translation, by an infinitesimal

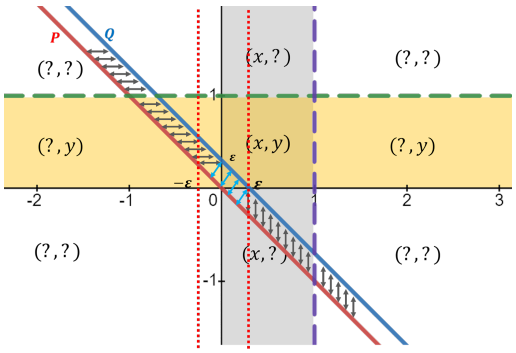


Figure E.1: In this example, the self-censoring missingness mechanism is as follows: Each coordinate x, y of the sample is seen if and only if $x \in [0,1]$ or $y \in [0,1]$ respectively.

amount ε, along the eigenvector corresponding to the 0 eigenvalue. It is not hard to see that if we take two such parallel subspaces arbitrarily close to each other, then with high probability no finite amount of censored samples can distinguish between these 2 cases, which would be required for keeping this Mahalanobis error bounded.

An illustrative example is in Fig. E.1: The censoring mechanism is as follows:

$$P^{\mathbb{S}} : \mathcal{N}(\mathbf{0}, \mathbf{I})|_{x+y=0}$$
$$Q^{\mathbb{S}} : \mathcal{N}(\mathbf{0}, \mathbf{I})|_{x+y=\varepsilon}$$

We will now present the following lemma, which formally justifies the above remark. It also acts as a warm-up for our subsequent lower bound.

Lemma 35. *For any sufficiently small value of $\varepsilon > 0$, given $m = o(1/\varepsilon)$ censored samples according to the missingness model $\mathbb{S}$ from either $P^{\mathbb{S}}$ or $Q^{\mathbb{S}}$, no algorithm can distinguish whether the samples are coming from $P^{\mathbb{S}}$ or $Q^{\mathbb{S}}$ with probability larger than $2/3$.*

Proof. We will define a coupling between the distributions $P^{\mathbb{S}}$ and $Q^{\mathbb{S}}$ such that with probability at least $1 - O(\varepsilon)$, the same missingness pattern and the same values for the seen coordinates are observed. This implies our lower bound of $\Omega(1/\varepsilon)$ samples for distinguishing the two distributions with constant probability.

More specifically, we define the following coupling matching the probability mass that points on the line $\{x + y = 0\}$ have due to $P^{\mathbb{S}}$ to the mass that $Q^{\mathbb{S}}$ imposes on the line $\{x + y = \varepsilon\}$:

$$(x,-x) \leftrightarrow \begin{cases} (x+\varepsilon,-x) & x \le -\varepsilon & \text{observed: } (?,-x) \\ (x+\varepsilon/2,-x+\varepsilon/2) & -\varepsilon/2 \le x \le \varepsilon/2 & \\ (x,-x+\varepsilon) & x \ge \varepsilon & \text{observed: } (x,?) \end{cases} \tag{E.1}$$

For the segments of the support of $P^{\mathbb{S}}$ unaccounted for in the above equation, we match their mass arbitrarily in a valid way to finish the coupling. Note that for the first and third branch of Eq. (E.1), the point $(x,-x)$ has strictly larger probability density from $P^{\mathbb{S}}$ than $(x+\varepsilon,-x)$ and $(x,-x+\varepsilon)$ respectively have from $Q^{\mathbb{S}}$, while for any $x \in \mathbb{R}$, the point $(x+\varepsilon/2,-x+\varepsilon/2)$ has the same density in $Q^{\mathbb{S}}$ as $(x,-x)$ has in $P^{\mathbb{S}}$.

To see this, consider the coordinate system: $\left\{w := \frac{x-y}{\sqrt{2}}, z := \frac{x+y}{\sqrt{2}}\right\}$ (which is a rotation of the original one by $\pi/4$ rad). Since the distribution $\mathcal{N}(\mathbf{0},\mathbf{I})$ is isotropic, can now write the distributions $P^{\mathbb{S}}$ and $Q^{\mathbb{S}}$ as follows:

$$P^{\mathbb{S}} : \mathcal{N}(\mathbf{0},\mathbf{I})|_{z=0}$$
$$Q^{\mathbb{S}} : \mathcal{N}(\mathbf{0},\mathbf{I})|_{z=\varepsilon/\sqrt{2}}$$

Due to the fact that the marginals over w and z are independent, we have that $w \sim \mathcal{N}(0,1)$ in both of the cases above.

Given the above coupling, it follows that whenever the true sample has first coordinate in the set $(-\infty,-\varepsilon)\cup(\varepsilon,+\infty)$ the observed sample would be exactly the same (see Eq. (E.1)) either with $P^{\mathbb{S}}$ or $Q^{\mathbb{S}}$.

Therefore, with probability at least $1-2\,\mathrm{erf}(\varepsilon\sqrt{2}) \ge 1-2\sqrt{2}\,\mathrm{erf}(\varepsilon)$, the censored sample that we get will not give us any information for our task distinguishing $P^{\mathbb{S}}$ from $Q^{\mathbb{S}}$. Thus, any algorithm that is able to distinguish $P^{\mathbb{S}}$ from $Q^{\mathbb{S}}$ need to draw $\Omega(\frac{1}{\mathrm{erf}(\varepsilon)})$ samples. This is $\Omega(1/\varepsilon)$ samples as ε gets arbitrarily close to 0. □

Lemma 9. *Given $m = o(1/\sqrt{\lambda_{min}})$ censored samples according to the missingness model $\mathbb{S}$ and $\varepsilon = \Omega(\sqrt{\lambda_{min}})$. No algorithm can estimate the true mean with accuracy $O(\varepsilon)$ and probability larger than $2/3$.*

Proof. We now consider two families of distributions parameterized by $\lambda \in [0,1]$. We define the two families in the rotated (z,w) coordinate system, as in Lemma 35,

as follows:

$$\begin{aligned} P_\lambda &: (z,w) \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \lambda & 0 \\ 0 & 1 \end{bmatrix}\right) \\ Q_\lambda &: (z,w) \sim \mathcal{N}\left(\begin{bmatrix} \varepsilon \\ 0 \end{bmatrix}, \begin{bmatrix} \lambda & 0 \\ 0 & 1 \end{bmatrix}\right) \end{aligned} \tag{E.2}$$

Note that the total variation distance between P_λ and Q_λ is:

$$\begin{aligned} d_{TV}(P_\lambda, Q_\lambda) &= d_{TV}(P_\lambda|_{w=0}, Q_\lambda|_{w=0}) \\ &= 2\Phi\left(\frac{\varepsilon}{2\sqrt{\lambda}}\right) - 1 = \operatorname{erf}\left(\frac{\varepsilon}{2\sqrt{2}\sqrt{\lambda}}\right) = O\left(\frac{\varepsilon}{\sqrt{\lambda}}\right). \end{aligned} \tag{E.3}$$

Consider the distributions P_λ and Q_λ defined in Eq. (E.2) for $\lambda = \lambda_{min}$. The main idea of the proof is that we can apply Lemma 35 for the case where $\varepsilon = \Theta(\sqrt{\lambda})$. Note that the distance between the means of these two distributions is $\varepsilon/\sqrt{2}$. Thus, any algorithm that can estimate the mean with accuracy at most $\varepsilon\sqrt{2}/4$, should be able to distinguish them.

We will use the same coupling as in Eq. (E.1) between the two distributions P_λ and Q_λ and use it to bound the probability that we will observe different censored samples p_c and q_c respectively. We observe that (similarly to the setting of Lemma 35) any sample from P_λ falling outside the band: $B = \{(x,y) : -\varepsilon \le x \le \varepsilon\}$, has an identical censored sample to the censored sample of the corresponding point in Q_λ via the coupling.

We now upper bound the probability that sample from P_λ falls in the band B:

$$\Pr_{(x,y)\sim P_\lambda}\left[-\varepsilon \le x \le \varepsilon\right] \le \Pr_{(z,w)\sim P_\lambda}\left[-\varepsilon\sqrt{2} \le w \le \varepsilon\sqrt{2}\right] = O(\varepsilon) \tag{E.4}$$

Thus, we have:

$$\Pr[p_c \neq q_c] \le \Pr_{(x,y)\sim P_\lambda}\left[-\varepsilon \le x \le \varepsilon\right] = O(\varepsilon) \tag{E.5}$$

In addition, due to Eq. (E.3), there exists a different coupling for which the following holds:

$$\Pr[p_c \neq q_c] \le \Pr_{p\sim P_\lambda, q\sim Q_\lambda}\left[p \neq q\right] = O\left(\frac{\varepsilon}{\sqrt{\lambda}}\right). \tag{E.6}$$

By Eq. (E.5) and Eq. (E.6), we get that no algorithm with $o(\max\{1/\varepsilon, \sqrt{\lambda}/\varepsilon\})$ samples can distinguish P_λ from Q_λ with probability at least $2/3$.

This implies the statement.

□

Theorem 7.1.2. *Suppose we can observe samples from $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ censored through a self-censoring missingness model $\mathbb{S}$. If assumption 7.1.1 is satisfied for some constant value of the parameter α, there exists a polynomial-time algorithm that recovers estimated $\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*$ with arbitrary accuracy. Specifically, for all $\varepsilon > 0$, and given that the eigenvalues of $\boldsymbol{\Sigma}^*$ lie in the interval $[\lambda_{min}, \lambda_{max}]$, the algorithm uses $\tilde{\mathcal{O}}\left(\frac{d^2(\lambda_{max}/\lambda_{min})^2}{\alpha\varepsilon^2}\right)$ samples and produces estimates that satisfy the following:*

$$\left\|\boldsymbol{\Sigma}^{*-1/2}(\boldsymbol{\mu}^* - \hat{\boldsymbol{\mu}})\right\|_2 \leq \mathcal{O}(\varepsilon);$$
$$\text{and } \left\|\boldsymbol{I} - \boldsymbol{\Sigma}^{*-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{*-1/2}\right\|_F \leq \mathcal{O}(\varepsilon).$$

Note that the sample complexity is proportional to $1/\alpha$. Furthermore, under the above conditions, we have $\mathrm{d_{TV}}(\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*), \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})) \leq \mathcal{O}(\varepsilon)$.

Proof of Theorem 7.1.2. By Lemma 7 and Lemma 8, we conclude that we can use $\tilde{O}(\frac{d^2}{\alpha\varepsilon^2})$ samples to get the following guarantees:

$$\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|_2 \leq \varepsilon\sqrt{\lambda_{max}}$$

and

$$\|\boldsymbol{\Sigma}^* - \hat{\boldsymbol{\Sigma}}\|_F \leq \varepsilon\lambda_{max}$$

Thus,we get the following:

$$\left\|\boldsymbol{\Sigma}^{*-1/2}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})\right\|_2 \leq \frac{1}{\sqrt{\lambda_{min}}}\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|_2 \leq \varepsilon\sqrt{\lambda_{max}/\lambda_{min}}$$

$$\begin{aligned}\left\|\boldsymbol{I} - \boldsymbol{\Sigma}^{*-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{*-1/2}\right\|_F &= \left\|\boldsymbol{\Sigma}^{*-1/2}(\boldsymbol{\Sigma}^* - \hat{\boldsymbol{\Sigma}})\boldsymbol{\Sigma}^{*-1/2}\right\|_F \\ &\leq \frac{1}{\lambda_{min}}\|\boldsymbol{\Sigma}^* - \hat{\boldsymbol{\Sigma}}\|_F \\ &\leq \varepsilon\lambda_{max}/\lambda_{min}\end{aligned}$$

Thus, by substituting $\varepsilon'' = \varepsilon\lambda_{max}/\lambda_{min}$, we get sample complexity of $\tilde{O}(\frac{d^2(\lambda_{max}/\lambda_{min})^2}{\alpha\varepsilon^2})$ for the following guarantees:

$$\left\|\boldsymbol{\Sigma}^{*-1/2}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})\right\|_2 \leq \mathcal{O}(\varepsilon)$$

$$\left\|\boldsymbol{I} - \boldsymbol{\Sigma}^{*-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{*-1/2}\right\|_F \leq \mathcal{O}(\varepsilon)$$

□

E.3 Section 7.4 Omitted Proofs

Lemma 10. *For any $\boldsymbol{\mu} \in \mathbb{R}^d$, it holds that: $\ell(\boldsymbol{\mu}) \geq \ell(\boldsymbol{\mu}^*)$.*

Proof. We first verify that the gradient vanishes at $\boldsymbol{\mu} = \boldsymbol{\mu}^*$. First, observe that

$$\boldsymbol{\nabla}\ell(\boldsymbol{\mu}) = -\sum_{A\subseteq[d]} \int_{\mathbf{x}\in\mathbb{R}^{|A|}} \frac{\int_{\mathbf{y}} \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}) \cdot \mathbb{1}_{\{\}}\mathbb{S}(\mathbf{y}) = A] \cdot \delta(\mathbf{y}_A - \mathbf{x}) g_{\boldsymbol{\mu}}(\mathbf{y}) d\mathbf{y}}{g^{\mathbb{S}}_{\boldsymbol{\mu}}(A, \mathbf{x})} g^{\mathbb{S}}_{\boldsymbol{\mu}^*}(A, \mathbf{x}) d\mathbf{x}$$

Hence:

$$\begin{aligned}
\boldsymbol{\nabla}\ell(\boldsymbol{\mu}^*) &= -\sum_{A\subseteq[d]} \int_{\mathbf{x}\in\mathbb{R}^{|A|}} \int_{\mathbf{y}} \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}^*) \cdot \mathbb{1}_{\{\}}\mathbb{S}(\mathbf{y}) = A] \cdot \delta(\mathbf{y}_A - \mathbf{x}) g_{\boldsymbol{\mu}^*}(\mathbf{y}) d\mathbf{y} d\mathbf{x} \\
&= -\int_{\mathbf{y}} \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}^*) \cdot \sum_{A\subseteq[d]} \mathbb{1}_{\{\}}\mathbb{S}(\mathbf{y}) = A] \cdot \int_{\mathbf{x}\in\mathbb{R}^{|A|}} \delta(\mathbf{y}_A - \mathbf{x}) d\mathbf{x} \cdot g_{\boldsymbol{\mu}^*}(\mathbf{y}) d\mathbf{y} \\
&= -\int_{\mathbf{y}} \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}^*) \cdot g_{\boldsymbol{\mu}^*}(\mathbf{y}) d\mathbf{y} = 0.
\end{aligned}$$

One can also show this by using (7.3) for the gradient and then using the law of total expectation. We next prove ℓ is convex by showing that $\boldsymbol{\nabla}^2\ell$ is positive semidefinite for any value of μ.

$$\begin{aligned}
&\boldsymbol{\nabla}^2\ell(\boldsymbol{\mu}) \\
&= \mathop{\mathbb{E}}_{(A,\mathbf{x})} \left[-\frac{\int_{\mathbf{y}} (-\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}) \mathbb{1}_{\{\}}\mathbb{S}(\mathbf{y}) = A] \cdot \delta(\mathbf{y}_A - \mathbf{x}) g_{\boldsymbol{\mu}}(\mathbf{y}) d\mathbf{y}}{g^{\mathbb{S}}_{\boldsymbol{\mu}}(A, \mathbf{x})} \right] + \\
&\quad \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma})} \left[\left(\frac{\int_{\mathbf{y}} \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}) \mathbb{1}_{\{\}}\mathbb{S}(\mathbf{y}) = A] \cdot \delta(\mathbf{y}_A - \mathbf{x}) g_{\boldsymbol{\mu}}(\mathbf{y}) d\mathbf{y}}{\int_{\mathbf{y}} \mathbb{1}_{\{\}}\mathbb{S}(\mathbf{y}) = A] \cdot \delta(\mathbf{y}_A - \mathbf{x}) g_{\boldsymbol{\mu}}(\mathbf{y}) d\mathbf{y}} \right)^2 \right] \\
&= \mathop{\mathbb{E}}_{(A,\mathbf{x})} \left[\boldsymbol{\Sigma}^{-1} - \mathop{\mathrm{Cov}}_{\mathbf{y}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})} \left[\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}) \mid \mathbb{S}(\mathbf{y}) = A, \mathbf{y}_A = \mathbf{x} \right] \right].
\end{aligned} \tag{E.7}$$

Observe that for a linear thresholding missingness pattern, the set $\{\mathbf{y} : \mathbb{S}(\mathbf{y}) = A, \mathbf{y}_A = \mathbf{x}\}$ is convex for any set A and any $\mathbf{x} \in \mathbb{R}^{|A|}$. Using the fact (Corollary 2.1 of [KP77]) that the variance of a Gaussian is non-increasing when restricted to a convex set:

$$\mathop{\mathrm{Cov}}_{\mathbf{y}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}) \mid \mathbb{S}(\mathbf{y}) = A, \mathbf{y}_A = \mathbf{x}] \preceq \mathop{\mathrm{Cov}}_{\mathbf{y}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})] = \boldsymbol{\Sigma}^{-1}.$$

Plugging into (E.7), we get that $\boldsymbol{\nabla}^2\ell(\boldsymbol{\mu}) \succeq 0$ for any $\boldsymbol{\mu}$. □

Lemma 13 (Strong Convexity with Missing Entries)**.** *Given our missingness model and assumption 7.1.3 with $\beta = \frac{c}{d}$ for some integer $c \in \{1, \dots d\}$, we have that the function with general covariance $\ell(\boldsymbol{\mu})$ is λ-strongly convex for $\lambda = \alpha\beta/\lambda_{\max}(\boldsymbol{\Sigma})$.*

Proof. Equivalently, we need to show that the minimum eigenvalue of the Hessian of the function $\ell(\boldsymbol{\mu})$ is at least λ. From (E.7), we have

$$\boldsymbol{\nabla}^2\ell(\boldsymbol{\mu}) = \mathop{\mathbb{E}}_{(A,\mathbf{x})}\left[\boldsymbol{\Sigma}^{-1} - \mathop{\mathrm{Cov}}_{\mathbf{y}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}) \mid \mathbb{S}(\mathbf{y}) = A, \mathbf{y}_A = \mathbf{x}\right]\right] \tag{E.8}$$

Fix a unit vector $\mathbf{v} \in \mathbb{R}^d$. Let $H \subseteq [d]$ denote the set of indices that contains the $c = \beta d$ highest v_i^2 values. Therefore, $\sum_{i\in H} v_i^2 \geq \beta$. Using $\mathbf{v}$ as a test vector: enecccfckgvkkjlfnlclcggguvjldtdurlrlltnecnid

$$\mathbf{v}^\top\nabla^2(\ell(\boldsymbol{\mu}))\mathbf{v} = \mathbf{v}^\top\boldsymbol{\Sigma}^{-1}\mathbf{v} - \mathbf{v}^\top\cdot\left(\mathop{\mathbb{E}}_{(A,\mathbf{x})}\mathop{\mathrm{Cov}}_{\mathbf{y}}(\boldsymbol{\Sigma}^{-1}\mathbf{y} \mid \mathbb{S}(\mathbf{y}) = A, \mathbf{y}_A = \mathbf{x})\right)\cdot\mathbf{v}$$

With probability $1-\alpha(H)$, some coordinate in H is not in A. Under this event, we upper-bound $\mathrm{Cov}_{\mathbf{y}}(\boldsymbol{\Sigma}^{-1}\mathbf{y} \mid \mathbb{S}(\mathbf{y}) = A, \mathbf{y}_A = \mathbf{x})$ by $\mathrm{Cov}_{\mathbf{y}}(\boldsymbol{\Sigma}^{-1}\mathbf{y}) = \boldsymbol{\Sigma}^{-1}$ using the facts that the missingness is linear thresholding and the aforementioned Corollary 2.1 of [KP77]. Under the complement event that $H \subseteq A$ is fully observed, we use the upper-bound $\mathrm{Cov}_{\mathbf{y}}(\boldsymbol{\Sigma}^{-1}\mathbf{y} \mid \mathbf{y}_H = \mathbf{x}_H)$, again by the same facts. By assumption 7.1.3, $\alpha(A) \geq \alpha$ and so:

$$\begin{aligned}\mathbf{v}^\top\nabla^2(\ell(\boldsymbol{\mu}))\mathbf{v} &\geq \alpha\left(\mathbf{v}^\top\boldsymbol{\Sigma}^{-1}\mathbf{v} - \mathbf{v}^\top\cdot\mathop{\mathrm{Cov}}_{\mathbf{y}}(\boldsymbol{\Sigma}^{-1}\mathbf{y} \mid \mathbf{y}_H = \mathbf{x}_H)\cdot\mathbf{v}\right)\\ &= \alpha\mathbf{v}^\top\left(I - \boldsymbol{\Sigma}^{-1}\mathop{\mathrm{Var}}_{\mathbf{y}}(\mathbf{y} \mid \mathbf{y}_H = \mathbf{x}_H)\right)\boldsymbol{\Sigma}^{-1}\mathbf{v}\end{aligned} \tag{E.9}$$

We use the standard facts that for any set A:

(i) $\mathrm{Var}_{\mathbf{y}}(\mathbf{y} \mid \mathbf{y}_A = \mathbf{x}_A) = \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{\Sigma}_{\bar{A}\bar{A}} - \boldsymbol{\Sigma}_{\bar{A}A}\boldsymbol{\Sigma}_{AA}^{-1}\boldsymbol{\Sigma}_{A\bar{A}}\end{bmatrix}$

(ii) $\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{AA}^{-1} + \boldsymbol{\Sigma}_{AA}^{-1}\boldsymbol{\Sigma}_{A\bar{A}}(\boldsymbol{\Sigma}_{\bar{A}\bar{A}} - \boldsymbol{\Sigma}_{\bar{A}A}\boldsymbol{\Sigma}_{AA}^{-1}\boldsymbol{\Sigma}_{A\bar{A}})^{-1}\boldsymbol{\Sigma}_{\bar{A}A}\boldsymbol{\Sigma}_{AA}^{-1} & -\boldsymbol{\Sigma}_{AA}^{-1}\boldsymbol{\Sigma}_{A\bar{A}}(\boldsymbol{\Sigma}_{\bar{A}\bar{A}} - \boldsymbol{\Sigma}_{\bar{A}A}\boldsymbol{\Sigma}_{AA}^{-1}\boldsymbol{\Sigma}_{A\bar{A}}) \\ -(\boldsymbol{\Sigma}_{\bar{A}\bar{A}} - \boldsymbol{\Sigma}_{\bar{A}A}\boldsymbol{\Sigma}_{AA}^{-1}\boldsymbol{\Sigma}_{A\bar{A}})^{-1}\boldsymbol{\Sigma}_{\bar{A}A}\boldsymbol{\Sigma}_{AA}^{-1} & (\boldsymbol{\Sigma}_{\bar{A}\bar{A}} - \boldsymbol{\Sigma}_{\bar{A}A}\boldsymbol{\Sigma}_{AA}^{-1}\boldsymbol{\Sigma}_{A\bar{A}})^{-1}\end{bmatrix}$

Using them repeatedly to simplify (E.9), we get:

$$\mathbf{v}^\top\nabla^2(\ell(\boldsymbol{\mu}))\mathbf{v} \geq \alpha\mathbf{v}^\top\left(I - \begin{bmatrix} 0 & -\boldsymbol{\Sigma}_{HH}^{-1}\boldsymbol{\Sigma}_{H\bar{H}} \\ 0 & I\end{bmatrix}\right)\boldsymbol{\Sigma}^{-1}\mathbf{v}$$

$$
\begin{aligned}
&= \alpha \mathbf{v}^\top \begin{bmatrix} I & \boldsymbol{\Sigma}_{HH}^{-1}\boldsymbol{\Sigma}_{H\bar{H}} \\ 0 & 0 \end{bmatrix} \boldsymbol{\Sigma}^{-1}\mathbf{v} \\
&= \alpha \mathbf{v}^\top \begin{bmatrix} \boldsymbol{\Sigma}_{HH}^{-1} & 0 \\ 0 & 0 \end{bmatrix} \mathbf{v} \\
&= \alpha \mathbf{v}_H^\top \boldsymbol{\Sigma}_{HH}^{-1}\mathbf{v}_H
\end{aligned}
$$

Finally, we use our choice of H and that $\lambda_{\min}(\boldsymbol{\Sigma}_{HH}^{-1}) = 1/\lambda_{\max}(\boldsymbol{\Sigma}_{HH}) \geq 1/\lambda_{\max}(\boldsymbol{\Sigma})$ to obtain our claim. □

We now describe our solutions to the following three problems as outlined in Section 7.4.

- **Initialization**: efficiently compute an initial feasible point from which to start the optimization. The pseudocode for `Initialize` appears in Algorithm 9;
- **Gradient estimation**: design a nearly unbiased sampler for $\boldsymbol{\nabla}\ell(\boldsymbol{\mu})$ using Langevin sampling. The `SampleGradient` pseudocode appears in Algorithm 12;
- **Efficient projection**: perform an efficient projection into a set of feasible points to make sure that PSGD converges. The pseudocode presents in Algorithm 11.

E.3.1 Gradient Estimation

Recall from the gradient expression in (7.2) that the main obstacle in computing $\nabla\ell(\boldsymbol{\mu})$ is the term $\mathbb{E}_{(A,\mathbf{x})}\,\mathbb{E}_{\mathbf{y}}[\mathbf{y} \mid \mathbb{S}(\mathbf{y}) = A, \mathbf{y}_A = \mathbf{x}]$. Here, $(A, \mathbf{x})$ is an observation generated from $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})^{\mathbb{S}}$, while $\mathbf{y}$ is sampled from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for some $\boldsymbol{\mu} \in \mathbb{R}^d$. So, to implement `SampleGradient`, we need an approximately unbiased estimator.

The most straightforward way is to apply rejection sampling: for an $(A, \mathbf{x})$ generated by $\mathcal{O}$, keep sampling $\mathbf{y}$ from the conditional[2] distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mid \mathbf{y}_A = \mathbf{x})$ until $\mathbb{S}(\mathbf{y}) = A$. If $\boldsymbol{\mu}^* = \boldsymbol{\mu}$, then the expected cost of the rejection sampling is $O(1/\gamma)$ by assumption 7.1.4, as $\mathbf{y}_A = \mathbf{x}$ implies $\mathbf{y}_C = \mathbf{x}_C$. The issue that arises is that the probability of $\mathbb{S}(\mathbf{y}) = A$ can decrease exponentially in the distance between $\boldsymbol{\mu}$ and $\boldsymbol{\mu}^*$, and so, rejection sampling becomes infeasible.

[2]We need to sample from the conditional distribution because $\mathbf{y} \equiv \mathbf{x}$ occurs with probability 0 in the $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ measure.

To sample the gradient when $\boldsymbol{\mu}$ is far from $\boldsymbol{\mu}^*$, we use the projected Langevin Monte Carlo algorithm [BEL18]. For an observation $(A, \mathbf{x})$, suppose $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and let $\boldsymbol{\mu}_{\text{cond}}$ and $\boldsymbol{\Sigma}_{\text{cond}}$ be the mean and covariance of $\mathbf{y}_{\bar{A}}$ conditioned on $\mathbf{y}_A = \mathbf{x}$. It is well-known that the conditional distribution is Gaussian with:

$$\boldsymbol{\mu}_{\text{cond}} = \boldsymbol{\mu}_{\bar{A}} + \boldsymbol{\Sigma}_{\bar{A},A}\boldsymbol{\Sigma}_{A,A}^{-1}(\mathbf{x} - \boldsymbol{\mu}_A) \tag{E.10}$$

$$\boldsymbol{\Sigma}_{\text{cond}} = \boldsymbol{\Sigma}_{\bar{A},\bar{A}} - \boldsymbol{\Sigma}_{\bar{A},A}\boldsymbol{\Sigma}_{A,A}^{-1}\boldsymbol{\Sigma}_{A,\bar{A}} \tag{E.11}$$

where $\bar{A}$ represents $[d] \setminus A$. Let[3] $\mathcal{K} = \{\mathbf{z} \in \mathbb{R}^{d-|A|} \mid \mathbb{S}(\mathbf{x} \circ \mathbf{z}) = A\}$. The iteration of the projected Langevin Monte Carlo algorithm takes the following form:

$$\mathbf{z}^{(t+1)} = \Pi_{\mathcal{K}\cap\mathcal{B}_\Sigma(\boldsymbol{\mu},R)}\left(\mathbf{z}^{(t)} - \frac{\eta}{2}\boldsymbol{\Sigma}_{\text{cond}}^{-1}(\mathbf{z}^{(t)} - \boldsymbol{\mu}_{\text{cond}}) + \sqrt{\eta}\cdot\boldsymbol{\zeta}^{(t)}\right) \tag{E.12}$$

where η is a step-size parameter, R is an appropriate radius parameter, and $\boldsymbol{\zeta}^{(0)}, \boldsymbol{\zeta}^{(1)}, \ldots$ are i.i.d. samples from the standard normal distribution in $(d - |A|)$-dimensions. We implicitly make the reasonable assumption here that Mahalanobis projection to the convex set $\Pi_{\mathcal{K}\cap\mathcal{B}_\Sigma(\boldsymbol{\mu},R)}(\cdot)$ can be performed efficiently. The pseudocode for `SampleGradient` appears in Algorithm 12.

E.3.2 Projection to Feasible Domain

Next, in each iteration of SGD in `MissingDescent` (Algorithm 10), we need to choose a projection set to make sure that PSGD converges. Specifically, we project a current guess back to a $\mathcal{B}_{\boldsymbol{\Sigma}}$ ball centered at $\boldsymbol{\mu}^{(0)}$.

E.3.3 Bounded Step Variance and Gradient Bias

Lemma 36. *$\mathcal{K} \cap \mathcal{B}_{\boldsymbol{\Sigma}}(\mathbf{0}, R_1) \supseteq \mathcal{B}_{\boldsymbol{\Sigma}}(\mathbf{c}, r)$ for some $\mathbf{c}$, where $R_1 = \sqrt{d} + O(\sqrt{\log(1/\gamma)}$ and $r = \Omega(\gamma/d^2)$*

Lemma 37. *For $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\|\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}} \le S$,*

$$\Pr[\mathbf{w} \notin \mathcal{B}_{\boldsymbol{\Sigma}}(\boldsymbol{\mu}, R_2) \cap K] \le \frac{\varepsilon}{16} \cdot \Pr[\mathbf{w} \in K]$$

where $R_2 = \operatorname{poly} d, S, \log(1/\gamma), \log(1/\varepsilon) > R_1$.

[3] $\mathbf{x}_A \circ \mathbf{z}$, where $\mathbf{z} \in \mathbb{R}^{d-|A|}$, denotes the vector $\mathbf{y} \in \mathbb{R}^d$ where $\mathbf{y}_A = \mathbf{x}_A$ and $\mathbf{y}_{\bar{A}} = \mathbf{z}$.

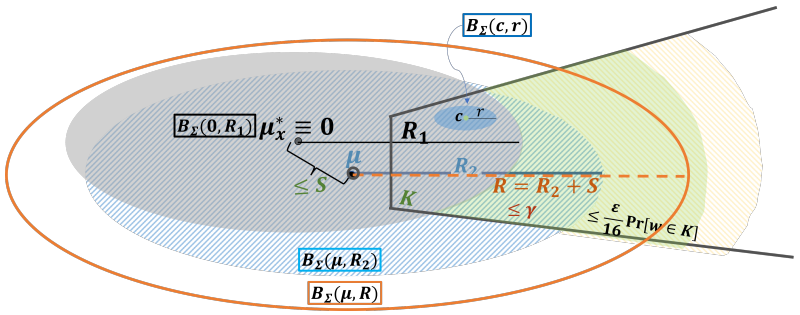


Figure E.2: An illustration of convex sets in Section 7.4.3.2.

Proof. Suppose $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. By standard concentration of gaussians:

$$\Pr[\|\mathbf{v}\|_{\mathbf{\Sigma}} > \sqrt{d} + O(\sqrt{\log(1/\delta)})] \leq \delta$$

Setting $\delta = \gamma/2$, we get that:

$$\Pr[\mathbf{v} \in \mathcal{B}_{\mathbf{\Sigma}}(\mathbf{0}, R_1) \cap \mathcal{K}] \geq \gamma/2. \tag{E.13}$$

Invoking Lemma 12 of [Che+22], we get that there exists $\mathbf{c}$ such that $\mathcal{B}_{\mathbf{\Sigma}}(\mathbf{0}, R_1) \cap \mathcal{K}$ contains $\mathcal{B}_{\mathbf{\Sigma}}(\mathbf{c}, r)$ for $r = \Omega(\gamma/d^2)$. □

Lemma 37. *For $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$, where $\|\boldsymbol{\mu}\|_{\mathbf{\Sigma}} \leq S$,*

$$\Pr[\mathbf{w} \notin \mathcal{B}_{\mathbf{\Sigma}}(\boldsymbol{\mu}, R_2) \cap K] \leq \frac{\varepsilon}{16} \cdot \Pr[\mathbf{w} \in K]$$

where $R_2 = \operatorname{poly} d, S, \log(1/\gamma), \log(1/\varepsilon) > R_1$.

Proof. Using (E.13):

$$\Pr[\mathbf{w} \in \mathcal{K}] \geq \Pr[\mathbf{w} \in \mathcal{B}_{\mathbf{\Sigma}}(\mathbf{0}, R_1) \cap \mathcal{K}] \geq \frac{\gamma}{2} \exp\left(-\frac{\|\boldsymbol{\mu}\|_{\mathbf{\Sigma}}^2 + 2R_1}{2}\right) \geq \frac{\gamma}{2} \exp\left(-\frac{S^2}{2} - R_1\right)$$

Call the lower-bound on the right γ'. Note that γ' may be exponentially smaller than γ for large S.

We define R_2 large enough so that $\Pr[\mathbf{w} \notin \mathcal{B}_{\mathbf{\Sigma}}(\boldsymbol{\mu}, R_2)] \leq \frac{\varepsilon}{16}\gamma'$. By concentration of gaussians, it suffices to take $R_2 = \sqrt{d} + O(\sqrt{\log(1/\gamma')}) = \sqrt{d} + O\left(\sqrt{\log \frac{1}{\gamma\varepsilon} + S^2 + R_1}\right)$. Note that the claim about R_2 follows. □

Theorem 7.4.2. *Assume $\|\boldsymbol{\mu}^* - \boldsymbol{\mu}\|_{\mathbf{\Sigma}} \leq S$. For $R = \tilde{O}(\sqrt{d} + S + \log(1/\gamma\varepsilon))$, if $M = \operatorname{poly} d, S, 1/\gamma, 1/\varepsilon$ and $\eta = \tilde{\Theta}(R^2/M)$, then*

$$\mathrm{d_{TV}}(\mathbf{z}^{(M)}, \mathcal{N}(W^{-1}\boldsymbol{\mu}_{cond}, I)) \leq \varepsilon.$$

Proof of Theorem 7.4.2. With R_2 as in Lemma 37, set $R = R_2 + S$, so that $\mathcal{L} = \mathcal{B}_{\mathbf{\Sigma}}(\boldsymbol{\mu}, R) \cap \mathcal{K}$. Since $R_2 > R_1$, Lemma 36 implies that $\mathcal{L}$ contains a ball of radius r. On the other hand, by Lemma 37, $\Pr[\mathbf{w} \notin \mathcal{L}] \leq \frac{\varepsilon}{4} \Pr[\mathbf{w} \in \mathcal{K}]$ for $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$, which implies that the truncation of $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ to $\mathcal{K}$ and to $\mathcal{L}$ are at most $\varepsilon/2$ far from each other in TV distance.

We can now use the main result of [BEL18] to approximately sample from the truncated gaussian $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}; \mathcal{L})$ with TV error $\varepsilon/2$. This work analyzes the projected Langevin Monte Carlo algorithm for sampling from a distribution μ on $\mathbb{R}^d$ whose density is proportional to $\exp(-f(\mathbf{x})) \cdot 1[\boldsymbol{x} \in \mathcal{M}]$ where $\mathcal{M}$ is a convex body containing the origin. Suppose for all $\mathbf{x}, \mathbf{y} \in \mathcal{M}$, $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq \beta \|\mathbf{x} - \mathbf{y}\|$ and $\|\nabla f(\mathbf{x})\| \leq \ell$. Consider the Langevin dynamics with $\bar{\mathbf{X}}_0 = \mathbf{0}$ and:

$$\bar{\mathbf{X}}_{k+1} = \Pi_{\mathcal{M}} \left(\bar{\mathbf{X}}_k - \frac{\eta}{2} \nabla f(\bar{\mathbf{X}}_k) + \sqrt{\eta} \boldsymbol{\zeta}_k \right)$$

where $\boldsymbol{\zeta}_0, \boldsymbol{\zeta}_1, \ldots$ are i.i.d. standard normal variables. If $\mathcal{M}$ contains a Euclidean ball of radius 1 and is contained in a Euclidean ball of radius R_{out}, then Theorem 1 of [BEL18] claims that $\mathrm{d}_{\mathrm{TV}}(\bar{\mathbf{X}}_N, \mu) \leq \varepsilon$ if $\eta = \tilde{\Theta}(R_{\text{out}}^2/N)$ and $N = \tilde{\Omega}(R_{\text{out}}^6 \max(d, R_{\text{out}}\ell, R_{\text{out}}\beta)^{12}/\varepsilon^{12})$.

In our context, Algorithm 12 already transforms by a Cholesky decomposition of $\mathbf{\Sigma}$ so as to transform Mahalanobis distance to Euclidean distance. We can then scale by r (from Lemma 36) to ensure that a Euclidean unit ball is contained inside the transformed $\mathcal{L}$. The radius of the outer ball is then $R_{\text{out}} \leq R/r \leq \tilde{O}(d^3 S/\gamma)$. We can bound the parameters β and ℓ (similarly to Section B.3 of [Che+22]):

$$\beta = O\left(\frac{\gamma^2}{d^4}\right) \qquad \ell = \tilde{O}\left(\frac{\gamma S}{d^{1.5}}\right)$$

So, invoking the result of [BEL18], for any particular $\mathbf{x}$, the running time of `SampleGradient` is $\operatorname{poly} d, S, 1/\gamma, 1/\varepsilon$. □

Corollary 7.4.3. *Let $\hat{\mathbf{g}}$ be the output of Algorithm 12 with inputs $\tilde{\mathbf{x}}$ and $\boldsymbol{\mu}$ and parameters R, M, η as in Theorem 7.4.2. Also, let $\mathbf{g} = -\mathbb{E}_{\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})}[\mathbf{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}) \mid \mathbb{S}(\mathbf{y}) = A, \mathbf{y}_A = \mathbf{x}]$. Then, we have that:*

$$\| \mathbb{E}[\hat{\mathbf{g}}] - \mathbf{g} \|_2 \leq \varepsilon \cdot \operatorname{poly} S, d, 1/\gamma, 1/\varepsilon, \lambda_{\max}, 1/\lambda_{\min} \tag{7.4}$$

Furthermore, we have the following bound

$$\mathbb{E}[\|\hat{\mathbf{g}}\|_2^2] \leq \operatorname{poly} d, 1/\gamma, S, 1/\lambda_{\min} \tag{7.5}$$

Proof. We first show Equation 7.4:

$$\begin{aligned}\|\mathbb{E}[\hat{\mathbf{g}}] - \mathbf{g}\|_2 &= \left\|\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_A \circ \mathbb{E}[W\mathbf{z}^{(M)}] - \tilde{\mathbf{x}}_A \circ \mathbb{E}[\mathbf{y}_{\bar{A}} \mid \mathbb{S}(\mathbf{y}) = A, \mathbf{y}_A = \mathbf{x}]\right)\right\| \\ &\leq \frac{1}{\lambda_{\min}(\boldsymbol{\Sigma})}\|\mathbb{E}[W\mathbf{z}^{(M)}] - \mathbb{E}[\mathbf{y}_{\bar{A}} \mid \mathbb{S}(\mathbf{y}) = A, \mathbf{y}_A = \mathbf{x}]\| \\ &\leq \frac{\varepsilon\sqrt{\lambda_{\max}(\boldsymbol{\Sigma})}}{\lambda_{\min}(\boldsymbol{\Sigma})}(R + O(\sqrt{\log(1/\gamma)}))\end{aligned}$$

The last inequality holds by the guarantee of Theorem 7.4.2, as well as the fact that $\mathbf{z}^{(M)}$ is contained within $\mathcal{B}_{\boldsymbol{\Sigma}}(\mu_{\text{cond}}, R)$ while $\mathbb{E}[\mathbf{y}_{\bar{A}} \mid \mathbf{y}_A = \mathbf{x}_A, \mathcal{A}(\mathbf{y}) = A]$ is within $\mathcal{B}_{\boldsymbol{\Sigma}}(\boldsymbol{\mu}_{\text{cond}}, O(\sqrt{\log(1/\gamma)}))$ by assumption 7.1.4 and Lemma 6 of [Das+18].

Given the above and the existence of the projection step in the gradient estimator, we can get the following bound on the centralized second moment of the gradient estimator:

$$\mathbb{E}[\|\hat{\mathbf{g}}\|_2^2] \leq \mathbb{E}[\|\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\|^2 \mid \mathbb{S}(\mathbf{y}) = A, \mathbf{y}_A = \mathbf{x}] + \varepsilon\left(\frac{R}{\sqrt{\lambda_{\min}}}\right)^2 \leq \operatorname{poly} d, 1/\gamma, S, 1/\lambda_{\min}$$

□

Bound on the Initialization

Lemma 38 (Empirical Parameters vs True Parameters)**.** *The empirical mean* $\mathbb{E}[w]$ *computed using* $\tilde{\mathcal{O}}(\frac{d\log(nd/\alpha\beta\delta)\log(1/\delta\beta)}{\varepsilon^2})$ *samples by sampling from the general missingness model with probability at least* $1-\delta$ *satisfy* $\|\mathbf{w}-\boldsymbol{\mu}^*\|_2 \leq \mathcal{O}\left(\sqrt{\frac{\lambda_{\max}}{\beta}}\log(1/\alpha)\right)$.

Proof. For each iteration of the for loop in Algorithm 9 uses $\tilde{\mathcal{O}}(\frac{\beta d\log(nd/\alpha\delta')\log(1/\delta')}{\varepsilon^2})$ samples due to the sample complexity bound in Lemma 5 of [Che+22] by applying Lemma 5 on Lemma 6 (1) using triangle inequality, and the bound holds with probability $1-\delta'$. Since the for loop makes $\lceil\frac{1}{\beta}\rceil$ iterations, we conclude that using $\tilde{\mathcal{O}}(\frac{d\log(nd/\alpha\delta')\log(1/\delta')}{\varepsilon^2})$ samples, our algorithm satisfies with probability at least $1-\delta'\cdot\lceil\frac{1}{\beta}\rceil$ using union bound. Let $\delta = \delta'\cdot\lceil\frac{1}{\beta}\rceil$. We have $\delta' = O(\delta\beta)$. Therefore, with $\tilde{\mathcal{O}}(\frac{d\log(nd/\alpha\beta\delta)\log(1/\delta\beta)}{\varepsilon^2})$ samples, with probability at least $1-\delta$, the output of Algorithm 9 satisfies that $\|\mathbf{w}-\boldsymbol{\mu}^*\|_2 \leq \mathcal{O}\left(\sqrt{\frac{\lambda_{\max}}{\beta}}\log(1/\alpha)\right)$. □

Appendix F

Supplementary Material - Chapter 8

F.1 p-tests and e-tests

Table F.1 summarizes key differences between p-tests and e-tests. While p-tests are linked to controlling false positive rates and p-values, e-tests are closely associated with likelihood ratios and betting game interpretations.

	p-tests (Martin-Löf)	e-tests (Levin)
Intuition	False positives / p-values	Betting games / likelihood ratios
Defining property	$\forall \epsilon > 0,\ P\left(\Lambda(X) \geq 1/\epsilon\right) \leq \epsilon$	$\mathbf{E}_{X\sim P}\left(\Lambda(X)\right) \leq 1$
Prototypical example	$1/P(\tau(X) \geq \tau(x))$	$Q(x)/P(x)$
Combination	$\sup_i w_i \Lambda_i$	$\sum_i w_i \Lambda_i$
Combination (log form)	$\sup_i\{\lambda_i + \log w_i\}$	$\log\left(\sum_i w_i 2^{\lambda_i}\right)$
Universal test (log form)	$\sup_i\{\lambda_i(x) - K(i \mid P^*)\}$	$-\log P(x) - K(x \mid P^*)$

Table F.1: Summary of types of test statistics (outlier scores)

Deriving e-Tests from p-Tests [VW21] describe a number of ways to *calibrate* any given p-test into an e-test. A natural choice is the Ramdas calibration, which turns the p-test Λ into the e-test

$$\Lambda'(x) := \frac{\Lambda(x) - \ln \Lambda(x) - 1}{\ln^2 \Lambda(x)}. \tag{F.1}$$

Example 7 (One-tailed p-value). For any feature statistic $\tau : \mathcal{X} \to \mathbb{R}$, the one-tailed p-value is given by

$$\Lambda_\tau(x) := P(\tau(X) \geq \tau(x)). \tag{F.2}$$

It is easily verified to be a p-test in probability form. The outlier score (8.2) is simply the p-value test (F.2) expressed in log form.

Example 8 (Likelihood ratio). For any alternative hypothesis described by a sub-probability distribution[1] Q, the likelihood ratio is given by

$$\Lambda_Q(x) := \frac{Q(x)}{P(x)}. \tag{F.3}$$

It is easily verified to be an e-test in ratio form; in fact, all e-tests can be written this way. By the Neyman-Pearson lemma [NP33], when Q is a probability distribution, Λ_Q is the optimal test to distinguish between P and Q.

[1]The algorithmic information theory literature refers to sub-probability distributions as *semimeasures*. They generalize probability measures by allowing their sum to be less than 1.

For the purposes of anomaly detection, τ might correspond to a feature that we expect would be higher for anomalies, while Q might correspond to a distribution that we expect would result from some kind of anomaly. In practice, we can design many tests, each specializing in a different kind of anomaly. The following lemma allows us to merge many such tests into one.

Lemma 39 (Combination tests)**.** *Let Λ_i (or λ_i) be a finite or countably infinite sequence of tests, with associated weights $w_i > 0$ summing to at most 1. Then:*

- *If Λ_i are p-tests in prob. form, so is $\inf_i \frac{\Lambda_i}{w_i}$.*
- *If Λ_i are p-tests in ratio form, so is $\sup_i w_i \Lambda_i$.*
- *If λ_i are p-tests in log form, so is $\sup_i \{\lambda_i + \log w_i\}$.*
- *If Λ_i are e-tests in prob. form, so is $\left(\sum_i \frac{w_i}{\Lambda_i}\right)^{-1}$.*
- *If Λ_i are e-tests in ratio form, so is $\sum_i w_i \Lambda_i$.*
- *If λ_i are e-tests in log form, so is $\log\left(\sum_i w_i 2^{\lambda_i}\right)$.*

Proof. [GRW06] state a similar result for finitely many independent p-tests, calling the combination test a *weighted Bonferroni procedure*. [VW21] state a similar result for equally weighted e-tests. It is fairly straightforward to extend these works to our setting; for completeness, the proof is as follows. First, given some p-tests Λ_i in ratio form, we verify Definition 8.2.1 for their combination $\sup_i w_i \Lambda_i$:

$$P\left(\sup_i w_i \Lambda_i(X) \geq \frac{1}{\epsilon}\right) \leq \sum_i P\left(\Lambda_i(X) \geq \frac{1}{w_i \epsilon}\right) \leq \sum_i w_i \epsilon \leq \epsilon.$$

Next, given e-tests Λ_i in ratio form, we verify Definition 8.2.1 for their combination $\sum_i w_i \Lambda_i$:

$$\mathbf{E}\left(\sum_i w_i \Lambda_i(X)\right) = \sum_i w_i \,\mathbf{E}(\Lambda_i(X)) \leq \sum_i w_i \leq 1.$$

Transforming the tests into probability and log form yields the remaining results. □

Example 9 (Two-tailed p-value). By combining the one-tailed p-values (F.2) for the feature statistics τ and $-\tau$, each with weight 0.5, we obtain the two-tailed p-value

$$\begin{aligned}\Lambda_{\pm\tau}(x) &:= \min\left\{\Lambda_\tau(x)/0.5,\ \Lambda_{-\tau}(x)/0.5)\right\} \\ &= 2\min\left\{\Pr(\tau(X) \geq \tau(x)),\ \Pr(\tau(X) \leq \tau(x)))\right\}.\end{aligned}$$

From now on, we express tests in log form except where stated otherwise. By Lemma 39, a combination test exceeds each of its component tests, up to the additive regularization term $\log w_i$ which does not depend on the sample x. This enables us to commit to a combination test a priori, while postponing the search over component tests until after observing x.

Unfortunately, the regularization term does depend on i, and becomes arbitrarily large as the number of tests becomes large or infinite. If we want the combination test to be competitive with its most promising component tests, it becomes important to choose the weights well [WR06]. The problem of choosing w_i is analogous to that of choosing Bayesian priors, and is philosophically challenged by formal impossibility results in the theory of inductive inference [Ada+19; Wol23]. Fortunately, computability poses a useful constraint on the set of permissible tests as well as priors [RH11]. It is suggestive that every computable sequence of weights w_i can be turned into a computable binary code with lengths $\lceil -\log w_i \rceil$ [CT06b]. Thus, minimizing the regularization penalty amounts to finding the shortest computable encoding for i.

F.2 Constructing the universal e-test

To develop a universal e-test, we first need to establish some fundamental computability concepts that will be essential for our construction.

We say a function $f : \{0,1\}^* \to \mathbb{R}$ is lower (or upper) semicomputable if it can be computably approximated from below (or above, respectively). For example, the Kolmogorov complexity K is upper, but not lower, semicomputable: by running all programs in parallel, we gradually find shorter programs that output x. We say f is computable if it is both lower and upper and semicomputable.

We say a p-test or e-test is semicomputable, if it is lower semicomputable when expressed as a function in either ratio or log form. A semicomputable p-test is called a Martin-Löf test [Mar66], while a semicomputable e-test is called a Levin test [Lev76]. Intuitively, semicomputable tests detect more anomalies as computation time increases. By restricting attention to semicomputable tests, it becomes possible to create universal combinations of them.

We return to the problem of optimizing the weights in Lemma 39: we would like

to dominate not only each of the individual component tests, but also each of the combination tests obtainable by some computable sequence of weights w_i. Note that the constraints on w_i amount to specifying a discrete sub-probability distribution. Among all lower semicomputable discrete sub-probability distributions,

$$m_i := 2^{-K(i|P^*)} \tag{F.4}$$

is *universal* in the sense that for all alternatives w and all i,

$$\log m_i \stackrel{+}{\geq} \log w_i - K(w \mid P^*).$$

This also holds when P is replaced by any other piece of prior knowledge.

Now, applying Lemma 39 with the universal weights (F.4), to any sequence of p-tests λ_i in log form, yields their universal combination

$$\lambda(x) := \sup_i \{\lambda_i(x) - K(i \mid P^*)\}.$$

In the case where $\{\lambda_i\}_{i=1}^{\infty}$ is a computable enumeration of some Martin-Löf tests, λ is itself a Martin-Löf test. It follows that if $\{\lambda_i\}_{i=1}^{\infty}$ enumerates *all* Martin-Löf tests, then λ is universal among them. Every program p that outputs i with access to P, determines the feature $\lambda_p := \lambda_i$. Moreover, it satisfies $|p| \geq K(i \mid P^*)$, with equality for the shortest such program. Therefore, we can rewrite the universal Martin-Löf test as

$$\rho(x) := \sup_p \{\lambda_p(x) - |p|\}. \tag{F.5}$$

With e-tests, the situation is even nicer because Equation (F.3) provides a one-to-one correspondence between e-tests Λ_Q and sub-probability distributions Q. Moreover, if we assume P to be computable, an e-test for it is semicomputable (i.e., is a Levin test) iff its corresponding Q is lower semicomputable. Applying Lemma 39 with the universal weights (F.4), to the likelihood ratios Λ_{Q_i}, yields their universal combination

$$\Lambda(x) := \sum_i m_i \Lambda_{Q_i}(x) = \frac{\sum_i m_i Q_i(x)}{P(x)}.$$

In the case where $\{Q_i\}_{i=1}^{\infty}$ enumerates all lower semicomputable sub-probability measures, Theorem 4.3.3 in [LV97] implies $\sum_i m_i Q_i(x) \stackrel{\times}{=} m(x \mid P^*)$, where $\stackrel{\times}{=}$ indicate

equality up to a multiplicative term. Switching to log form, we obtain the universal Levin test

$$\delta(x) := \log \frac{m(x \mid P^*)}{P(x)} = \log \frac{1}{P(x)} - K(x \mid P^*). \tag{F.6}$$

By the Kraft inequality, it is an e-test in log form. Note that (F.6) is the difference between a Shannon code length and a shortest program length for x. Intuitively, δ is high whenever the Shannon code derived from P is inefficient at compressing x, indicating that x possesses regularities that are atypical of P.

The algorithmic information theory literature uses the term *randomness deficiency* to refer to either the universal Martin-Löf test (F.5) or the universal Levin test (F.6). Keeping in mind conversions such as (F.1) between the two types of tests, we will develop our theory in terms of the latter.

Example 10 (Feature Selection through Universal Tests). Consider an infinite sequence of basis feature functions (f_i), where $\log P$ is expressed as a linear combination of finitely many features: $\log P := \sum_i \alpha_i f_i$. When an observation x exhibits atypical feature values under P, it would typically have a substantially higher likelihood under some modified linear combination $\log \tilde{P} := \sum_i \tilde{\alpha}_i f_i$.

If we assume the description of P is given in terms of the coefficient vector α, we can bound the Kolmogorov complexity:

$$K(x \mid P^*) \overset{+}{\leq} K(\tilde{\alpha} \mid \alpha) + \log \frac{1}{\tilde{P}(x)},$$

This leads to a lower bound on the Levin test:

$$\delta_P(x) = -\log P(x) - K(x \mid P^*) \overset{+}{\geq} \log \frac{\tilde{P}(x)}{P(x)} - K(\tilde{\alpha} \mid \alpha).$$

The bound becomes large when the improvement in likelihood (first term) substantially exceeds the complexity cost $K(\tilde{\alpha} \mid \alpha)$ required to modify the coefficients. This demonstrates how the universal tests naturally detect feature-based anomalies: when certain feature statistics of x are unusual under P, there exists an alternative distribution $\tilde{P}$ that better explains these feature values, leading to a high value of $\delta_P(x)$.

F.3 Decomposition of randomness deficiency

Lemma 16. *(Decomposition of randomness deficiency for a cause-effect pair) For any two random variables $X \to Y$ (i.e., X being the cause of Y), and for specific observations x and y, the following equality holds under Postulate 8.2.4 for the DAG in Fig. 8.1:*

$$\delta(x, y) \stackrel{+}{=} \delta(x) + \delta(y \mid x)$$

Proof. By Eqs. (8.5) and (8.6), we have:

$$\begin{aligned}
\delta(x, y) &\stackrel{+}{=} -\log P(x, y) - K(x, y \mid (P_{X,Y})^*) \\
&\stackrel{+}{=} -\log P(x, y) - K(x \mid (P_{X,Y})^*) - K(y \mid (x, P_{X,Y})^*), \\
\delta(x) + \delta(y \mid x) &\stackrel{+}{=} -\log P(x, y) - K(x \mid (P_X)^*) - K(y \mid (x, P_{Y|X})^*).
\end{aligned}$$

To complete the proof, it suffices to establish the following:

$$(1) \quad K(x \mid (P_{X,Y})^*) \stackrel{+}{=} K(x \mid (P_X)^*) \qquad \text{and} \qquad (2) \quad K(y \mid (x, P_{X,Y})^*) \stackrel{+}{=} K(y \mid (x, P_{Y|X})^*)$$

Proof of (1): Applying Lemma 15 to our bivariate case yields $x \perp\!\!\!\perp P_{Y|X} \,|P_X$. Hence,

$$K(x \mid (P_{X,Y})^*) \stackrel{+}{=} K(x \mid (P_{Y|X}, P_X)^*) \stackrel{+}{=} K(x \mid (P_X)^*).$$

Proof of (2): By Lemma 15, $y \perp\!\!\!\perp P_X \mid x, P_{Y|X}$, meaning that P_X becomes irrelevant when predicting y from a shortest program for x and $P_{Y|X}$. Hence,

$$K(y \mid (x, P_{X,Y})^*) \stackrel{+}{=} K(y \mid (x, P_{Y|X}, P_X)^*) \stackrel{+}{=} K(y \mid (x, P_{Y|X})^*).$$

□

Theorem 8.3.1 (Decomposition of multivariate joint randomness deficiency)**.** *Let the set of strings $x_1, x_2, \ldots, x_n$ be causally connected by a directed acyclic graph G, so that the causal Markov condition holds for G^m. Then the joint randomness deficiency of all strings $x_1, x_2, \ldots, x_n$ decomposes into the conditional randomness deficiencies of the mechanisms:*

$$\delta(x_1, \ldots, x_n) \stackrel{+}{=} \sum_{j=1}^{n} \delta(x_j \mid \mathrm{pa}_j),$$

where $\delta(x_j \mid \mathrm{pa}_j)$ denotes the randomness deficiency of x_j given its parents.

Proof. We prove this theorem by induction on n.

Base case: When $n = 1$, we have $\mathrm{pa}_1 = \emptyset$, so the claim is trivial.

Inductive Hypothesis: Assume that for any sequence of $n-1$ strings $x_1, x_2, \ldots, x_{n-1}$,

$$\delta(x_1, \ldots, x_{n-1}) \stackrel{+}{=} \sum_{j=1}^{n-1} \delta(x_j \mid \mathrm{pa}_j).$$

Inductive Step: Now we must prove the statement holds for n strings. The statistical Markov condition yields

$$P(x_n \mid x_1, \ldots, x_{n-1}) = P(x_n \mid \mathrm{pa}_n).$$

Meanwhile, the algorithmic Markov condition gives us Lemma 15, which implies

$$x_n \perp\!\!\!\perp x_1, \ldots, x_{n-1} \mid (\mathrm{pa}_n, P_{X_n \mid \mathrm{Pa}_n})^*.$$

Together with $P_{X_n|X_1,\ldots,X_{n-1}} = P_{X_n|\mathrm{Pa}_n}$, this yields

$$K(x_n \mid (x_1, \ldots, x_{n-1}, P_{X_n|X_1,\ldots,X_{n-1}})^*) \stackrel{+}{=} K(x_n \mid (\mathrm{pa}_n, P_{X_n|\mathrm{Pa}_n})^*).$$

Putting these results together,

$$\begin{aligned}
\delta(x_n \mid x_1, \ldots, x_{n-1}) &:= -\log P(x_n \mid x_1, \ldots, x_{n-1}) - K(x_n \mid (x_1, \ldots, x_{n-1}, P_{X_n|X_1,\ldots,X_{n-1}})^*) \\
&\stackrel{+}{=} -\log P(x_n \mid \mathrm{pa}_n) - K(x_n \mid (\mathrm{pa}_n, P_{X_n|\mathrm{Pa}_n})^*) \\
&= \delta(x_n \mid \mathrm{pa}_n).
\end{aligned}$$

Finally, we combine the inductive hypothesis with Lemma 16, with $x := (x_1, \ldots, x_{n-1})$ and $y := x_n$:

$$\begin{aligned}
\delta(x_1, \ldots, x_n) &\stackrel{+}{=} \delta(x_1, \ldots, x_{n-1}) + \delta(x_n \mid x_1, \ldots, x_{n-1}) \\
&\stackrel{+}{=} \sum_{j=1}^{n} \delta(x_j \mid \mathrm{pa}_j).
\end{aligned}$$

This completes the inductive step and the proof. □

F.4 Monotonicity of Randomness Deficiency

Theorem 8.4.1 (Weak anomalies do not cause stronger ones)**.** *If there is a unique root cause $j \in \{1, \ldots, n\}$, in the sense that*

$$\delta(x_i \mid \mathrm{pa}_i) \stackrel{+}{=} 0 \quad \text{for } i \neq j,$$

and the conditions of Theorem 8.3.1 are met,

$$\textit{then,} \qquad \delta(x_i) \stackrel{+}{\leq} \delta(x_j \mid \mathrm{pa}_j) \quad \forall i \in \{1, \ldots, n\}. \tag{8.7}$$

Proof. Theorem 8.3.1 states that the joint randomness deficiency of $x_1, \ldots, x_n$ decomposes into the sum of conditional randomness deficiencies of their mechanisms:

$$\delta(x_1, \ldots, x_n) \stackrel{+}{=} \sum_{j=1}^{n} \delta(x_j \mid \mathrm{pa}_j). \tag{F.7}$$

We are given that there exists a unique root cause $j \in \{1, \ldots, n\}$ such that $\delta(x_i \mid \mathrm{pa}_i) \stackrel{+}{=} 0$ for all $i \neq j$. Substituting this condition into Equation F.7, the sum simplifies:

$$\delta(x_1, \ldots, x_n) \stackrel{+}{=} \delta(x_j \mid \mathrm{pa}_j) + \sum_{k \neq j, k=1}^{n} \delta(x_k \mid \mathrm{pa}_k)$$

Since $\delta(x_k \mid \mathrm{pa}_k) \stackrel{+}{=} 0$ for $k \neq j$, this reduces to:

$$\delta(x_1, \ldots, x_n) \stackrel{+}{=} \delta(x_j \mid \mathrm{pa}_j). \tag{F.8}$$

Eq. (F.8) indicates that the total randomness deficiency of the joint observation $(x_1, \ldots, x_n)$ is approximately equal to the randomness deficiency of the unique root cause mechanism $\delta(x_j \mid \mathrm{pa}_j)$. Furthermore, according to Corollary 4.1.11 from [Gác21] which states non-increasingness of δ under marginalization:

$$\delta(x_i) \stackrel{+}{\leq} \delta(x_1, \ldots, x_n).$$

and thus:

$$\delta(x_i) \stackrel{+}{\leq} \delta(x_j \mid \mathrm{pa}_j) \quad \forall i \in \{1, \ldots, n\}.$$

□

F.5 Non-increasingness of Mahalanobis distance

Let $\mathbf{x} \in \mathbb{R}^n$, and let $Q\mathbf{x}$ denote the projection of $\mathbf{x}$ onto the subspace spanned by the variables $X_{i_1}, \ldots, X_{i_k}$, where $k < n$. We express the covariance matrix Σ_X in block matrix form as

$$\Sigma_X = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

with index 1 referring to the variables $i_1, \ldots, i_k$ and 2 to the remaining variables. Using a known formula for inversion of 2×2 block matrices (see Proposition 3.9.7 in [**bernstein2009**]), we obtain

$$\Sigma_X^{-1} = \begin{pmatrix} \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}A\Sigma_{21}\Sigma_{11}^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}A \\ -A\Sigma_{21}\Sigma_{11}^{-1} & A \end{pmatrix},$$

with $A := (\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}$. We now compute the difference between the squared Mahalanobis distances on $\mathbb{R}^n$ and $\mathbb{R}^k$:

$$\mathbf{x}^T\Sigma_X^{-1}\mathbf{x} - (Q\mathbf{x})^T\Sigma_{11}^{-1}Q\mathbf{x} = \mathbf{x}^T C\mathbf{x}, \tag{F.9}$$

with

$$C := \begin{pmatrix} \Sigma_{11}^{-1}\Sigma_{12}A\Sigma_{21}\Sigma_{11}^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}A \\ -A\Sigma_{21}\Sigma_{11}^{-1} & A \end{pmatrix} = \begin{pmatrix} \Sigma_{11}^{-1}\Sigma_{12} & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} A & A \\ A & A \end{pmatrix} \begin{pmatrix} \Sigma_{21}\Sigma_{11}^{-1} & 0 \\ 0 & -1 \end{pmatrix}.$$

Since A is positive semi-definite, and the rightmost matrix in the product is the transpose of the leftmost matrix, it follows that C is also positive semi-definite. Hence, Equation F.9 is non-negative.