Information-Theoretic Algorithms and Identifiability for Causal Graph Discovery

by

Spencer Compton

S.B. Computer Science and Engineering
Massachusetts Institute of Technology, 2022

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

Author . . . .
Department of Electrical Engineering and Computer Science
May 6, 2022

Certified by . . . .
Caroline Uhler
Associate Professor
Thesis Supervisor

Certified by . . . .
Kristjan Greenewald
Research Scientist, IBM Research
Thesis Supervisor

Accepted by . . . .
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Information-Theoretic Algorithms and Identifiability for Causal Graph Discovery

by

Spencer Compton

Submitted to the Department of Electrical Engineering and Computer Science
on May 6, 2022, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

It is a task of widespread interest to learn the underlying causal structure for systems of random variables. Entropic Causal Inference is a recent framework for learning the causal graph between two variables from observational data (i.e., without experiments) by finding the information-theoretically simplest structural explanation of the data. In this thesis, we develop theoretical techniques that enable us to show how Entropic Causal Inference permits learnability of causal graphs with particular information-theoretically simple structure. We show the first theoretical guarantee for finite-sample learnability with Entropic Causal Inference for pairs of random variables. Later, we extend this guarantee to show the first result for Entropic Causal Inference in systems with more than two variables: proving learnability of general directed acyclic graphs over many variables (under assumptions on the generative process). We implement and experimentally evaluate Entropic Causal Inference on synthetic and real-world causal systems. Moreover, we improve the best-known approximation guarantee for the Minimum Entropy Coupling problem. This information-theoretic algorithmic problem has direct relevance to Entropic Causal Inference and is also of independent interest. In totality, this thesis develops algorithmic and information-theoretic tools that shed light on how information-theoretic properties enable learning of causal graphs from both a practical and theoretical perspective.

Thesis Supervisor: Caroline Uhler
Title: Associate Professor

Thesis Supervisor: Kristjan Greenewald
Title: Research Scientist, IBM Research

Acknowledgments

I would like to extend my immense gratitude to all of my family, friends, and colleagues who have supported me throughout these years. It is this support and community that has made life at MIT so amazing.

I would like to thank my wonderful mentors and collaborators for this research: Kristjan Greenewald, Dmitriy Katz, and Murat Kocaoglu. I have cherished developing the research that constitutes the majority of this thesis together. From our collaboration, I have been greatly shaped by your patience and assurance to trust one's inner compass towards the mysteries that feel fundamental.

Furthermore, I would like to thank Ronitt Rubinfeld and Slobodan Mitrovic. I cannot give enough credit the positive impact you both have had on my mindset for how to treat research and people. I strive to emulate your curiosity and kindness.

Finally, I would also like to thank Caroline Uhler, IBM Research (particularly Luke Inglis), the MIT EECS department, and the MIT 6-A program for all their support throughout my time at MIT. Your help has made this thesis possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Very often, random variables in a system have relationships with each other (i.e., they are not independent). Such relationships can take many forms. For example, variables can be correlated. Correlation enables us to conclude statements such as *"Smoking is positively correlated with lung cancer."* However, correlation alone does not enable us to claim smoking *causes* cancer. Learning causal relationships enables counterfactual reasoning such as claiming *"If I make this person not smoke, they will be less likely to have lung cancer."* Here, knowing the causal relationships between variables enabled us to predict how interventions to a system would affect it.

More generally, learning causal relationships in a system enables higher quality decision-making. While we mentioned an example with just two random variables, it is of greater interest to learn causal structure in more complex systems with more random variables. In the study of graphical models, we can encode the causal structure of a system with a directed acyclic graph called the *causal graph.* In the causal graph, nodes represent random variables and edges represent relationships between said random variables. The task of identifying the true causal graph of a system is called *causal graph discovery.*

Of course, causal graph discovery relies on the assumptions we make on how the system is generated. Without any assumptions, a joint distribution of a system alone cannot completely inform us about the system's underlying causal mechanism. Yet, with some assumptions, we can begin to make inferences about systems' causal mechanisms. The gold-standard for learning causal graphs is to perform interventions (i.e., experiments). This is analogous to running randomized control experiments on smoking to identify its causal effect on lung cancer. However, in many real-world settings it is impossible or undesirable to perform such interventions (e.g., it would be unethical to perform an intervention to force an individual to smoke). In these cases, we must perform causal graph discovery with only observational data. In this body of work, we investigate information-theoretic structure where the true causal graph is *identifiable* from only observational data and providing provable algorithms to do so.

In Chapter 2, we study *Entropic Causal Inference* in the setting of pairs of variables. We consider causal relationships between two variables where one variable is a function of the other with low-entropy randomness (and more technical assumptions). In joint work (published as [12]) with Murat Kocaoglu, Kristjan Greenewald, and Dmitriy Katz, we show the first finite-sample identifiability result for Entropic Causal Inference and experimentally evaluate the approach.

In Chapter 3, we study Entropic Causal Inference in the setting of systems with many variables. We consider systems of causal relationships over many variables. In joint work with Murat Kocaoglu, Kristjan Greenewald, and Dmitriy Katz, we show the first result for Entropic Causal Inference beyond the pairwise setting. In particular, we show learnability of the underlying causal graph when variables are functions of their parents and low-entropy randomness (and more technical assumptions). Additionally, we experimentally evaluate this approach for learning causal graphs of real-world and synthetic systems.

In Chapter 4, we study *Minimum Entropy Coupling*, an algorithmic information-theory problem that is a key subroutine for Entropic Causal Inference, and is also of independent interest. We design a novel algorithmic analysis that improves the best-known polynomial-time additive-approximation guarantee to within $\log_2(e) \approx 1.44$ bits of the optimal.

At a high level, this body of work studies information-theoretically "simple" relationships between variables. The key results shown in Chapters 2 and 3 shed insight into how (under particular assumptions), simplicity in relationships between variables enables us to learn underlying causal structure that would otherwise be impossible to ascertain. Similar to how humans often use principles like Occam's razor to evaluate explanations, this work studies *"Under what conditions is the true generative model the most information-theoretically simplest way to produce a distribution?"* With a particular information-theoretic notion of simplicity in mind, the Minimum Entropy Coupling problem studied in Chapter 4 corresponds to solving the problem of fitting the simplest explanation to a causal graph. In totality, this body of work aims to build information-theoretic and algorithmic techniques, and to provide insight into learning causal structure with information-theoretic properties.

Chapter 2

Entropic Causal Inference: Identifiability and Finite Sample Results

2.1 Overview

In this chapter, we detail joint work with Murat Kocaoglu, Kristjan Greenewald, and Dmitriy Katz.

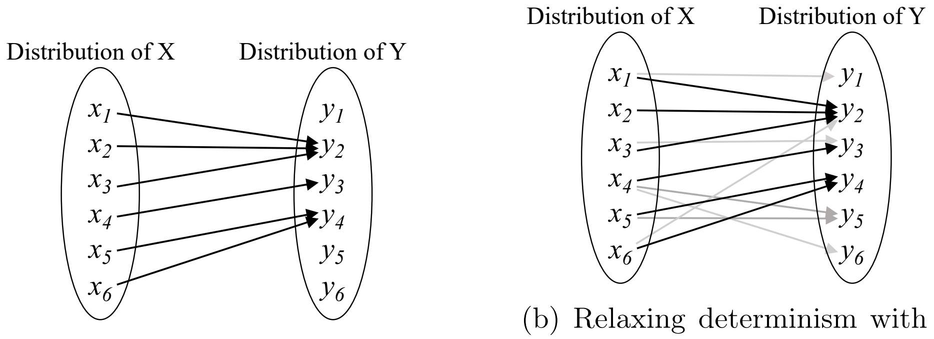
Entropic causal inference is a framework for inferring the causal direction between two categorical variables from observational data. The central assumption is that the amount of unobserved randomness in the system is not too large. This unobserved randomness is measured by the entropy of the exogenous variable in the underlying structural causal model, which governs the causal relation between the observed variables. [30] conjectured that the causal direction is identifiable when the entropy of the exogenous variable is not too large. In this paper, we prove a variant of their conjecture. Namely, we show that for almost all causal models where the exogenous variable has entropy that does not scale with the number of states of the observed variables, the causal direction is identifiable from observational data. We also consider the minimum entropy coupling-based algorithmic approach presented by [30], and for the first time demonstrate algorithmic identifiability guarantees using a finite number of samples. We conduct extensive experiments to evaluate the robustness of the method to relaxing some of the assumptions in our theory and demonstrate that both the constant-entropy exogenous variable and the no latent confounder assumptions can be relaxed in practice. We also empirically characterize the number of observational samples needed for causal identification. Finally, we apply the algorithm on Tübingen cause-effect pairs dataset.

2.2 Introduction

Understanding causal mechanisms is essential in many fields of science and engineering [56, 63]. Distinguishing causes from effects allows us to obtain a causal model of the environment, which is critical for informed policy decisions [48]. Causal inference has been recently utilized in several machine learning applications, e.g., to explain the decisions of a classifier [1], to design fair classifiers that mitigate dataset bias [28, 68] and to construct classifiers that generalize [62].

Consider a system that we observe through a set of random variables. For example, to monitor the state of a classroom, we might measure *temperature, humidity* and *atmospheric pressure* in the room. These measurements are random variables which come about due to the workings of the underlying system, the physical world. Changes in one are expected to cause changes in the other, e.g., decreasing the temperature might reduce the atmospheric pressure and increase humidity. As long as there are no feedback loops, we can represent the set of causal relations between these variables using a directed acyclic graph (DAG). This is called the *causal graph* of the system. Pearl and others showed that knowing the causal graph enables us to answer many causal questions such as, *"What will happen if I increase the temperature of the room?"* [48].

Therefore, for causal inference, knowing the underlying causal structure is crucial. Even though the causal structure can be learned from experimental data, in many tasks in machine learning, we only have access to a dataset and do not have the



(a) Deterministic relation.

(b) Relaxing determinism with noise.

Figure 2-1: Intuition behind the entropic causality framework. **(a)** Most deterministic maps would be non-deterministic in the opposite direction, requiring non-zero additional randomness. **(b)** Entropic causality relaxes the deterministic map assumption to a map that needs low-entropy, and demonstrates that, most of the time, the reverse direction needs more entropy than the true direction.

means to perform these experiments. In this case, observational data can be used for learning some causal relations. There are several algorithms in the literature for this task, which can be roughly divided into three classes: Constraint-based methods and score-based methods use conditional independence statements and likelihood function, respectively, to output (a member of) the equivalence class. An equivalence class of causal graphs are those that cannot be distinguished by the given data. The third class of algorithms impose additional assumptions about the underlying system or about the relations between the observed variables. Most of the literature focus on the special case of two observed variables X, Y and to understand whether X causes Y or Y causes X under different assumptions. Constraint or score-based methods cannot answer this question simply because observed data is not sufficient without further assumptions. In this work, we focus on the special case of two categorical variables. Even though the literature is more established in the ordinal setting, few results exist when the observed variables are categorical. The main reason is that, for categorical data, numerical values of variables do not carry any meaning; whereas in continuous data one can use assumptions such as smoothness or additivity [20].

We first start with a strong assumption. Suppose that the system is *deterministic*. This means that, even though observed variables contain randomness, the system has no additional randomness. When X causes Y, this assumption implies that $Y = f(X)$

for some deterministic map $f(.)$. Consider the example in Figure 2-1. Since there is no additional randomness, each value of X is mapped to a single value of Y. What happens if we did not know the causal direction and tried to fit a function in the wrong direction as $X = g(Y)$. Unlike f, g has to be one-to-many: $Y = 2$ is mapped to three different value of X. Therefore, it is impossible to find a deterministic function in the wrong causal direction for this system. In fact, it is easy to show that most of the functions have this property: If X, Y each has $n \geq 7$ states, all but 2^{-n} fraction of models can be identified.

Although there might be systems where determinism holds such as in traditional computer software, this assumption in general is too strict. Then *how much can we relax this assumption and still identify if X causes Y or Y causes X?* In general, we can represent a system as $Y = f(X, E)$ where E captures the additional randomness. To quantify this amount of relaxation, we use the entropy of the additional randomness in the structural equation, i.e., $H(E)$. For deterministic systems, $H(E) = 0$. This question was posed as a conjecture in [30], within the entropic causal inference framework.

We provide the first result in resolving this question. Specifically, we show that the causal direction is still identifiable for any E with constant entropy. Our usage of "constant" is relative to the support size n of the observed variables (note $0 \leq H(X) \leq \log(n)$). This establishes a version of Kocaoglu's conjecture.

A practical question is how much noise can the entropic causality framework handle: do we always need the additional randomness to not scale with n? Through experiments, we demonstrate that, in fact, we can relax this constraint much further. If $H(E) \approx \alpha \log(n)$, we show that in the wrong causal direction we need entropy of at least $\beta \log(n)$ for $\beta > \alpha$. This establishes that entropic causal inference is robust to the entropy of noise and for most models, reverse direction will require larger entropy. We finally demonstrate our claims on the benchmark Tübingen dataset.

We also provide the first finite-sample analysis and provide bounds on the number of samples needed in practice. This requires showing finite sample bounds for the minimum entropy coupling problem, which might be of independent interest. The following is a summary of our contributions.

- We prove the first identifiability result for the entropic causal inference framework using Shannon entropy and show that for most models, the causal direction between two variables is identifiable, if the amount of exogenous randomness does not scale with n, where n is the number of states of the observed variables.

- We obtain the first bounds on the number of samples needed to employ the entropic causal inference framework. For this, we provide the first sample-bounds for accurately solving the minimum entropy coupling problem in practice, which might be of independent interest.

- We show through synthetic experiments that our bounds are loose and entropic causal inference can be used even when the exogenous entropy scales with $\alpha \log(n)$ for $\alpha < 1$.

- We employ the framework on Tübingen data to establish its performance. We also conduct experiments to demonstrate robustness of the method to latent confounders, robustness to asymmetric support size, i.e., when X, Y have very different number of states, and finally establish the number of samples needed in practice.

Notation: We will assume, without loss of generality, that if a variable has n states, its domain is $[n] := \{1, 2, \ldots, n\}$. $p(x)$ is short for $p(X = x)$. $p(Y|x)$ is short for the distribution of Y given $X = x$. *Simplex* is short for probability simplex, which, in n dimensions is the polytope defined as $\Delta_n := \{(x_i)_{i\in[n]} : \sum_i x_i = 1, x_i \geq 0, \forall i \in [n]\}$. $\mathbb{1}_{\{\varepsilon\}}$ is the indicator variable for event ε. *SCM* is short for *structural causal model* and refers to the functional relations between variables. For two variables where X causes Y, the SCM is $Y = f(X, E), X \perp\!\!\!\perp E$ for some variable E and function f.

2.3 Related Work

There are a variety of assumptions and accompanying methods for inferring the causal relations between two observed variables [16,17,36,50,58]. For example, authors in [20] developed a framework to infer causal relations between two continuous variables if the exogenous variables affect the observed variable additively. This is called the *additive noise model (ANM)*. Under the assumption that the functional relation is non-linear they show identifiability results, i.e., for almost all models the causal direction between two observed variables can be identified. This is typically done by testing independence of the residual error terms from the regression variables. Interestingly in [34] authors show that independence of regression residuals leads the total entropy in the true direction to be smaller than the wrong direction, which can be used for identifiability thereby arriving at the same idea we use in our paper.

A challenging setting for causal inference is the setting with discrete and categorical variables, where the variable labels do not carry any specific meaning. For example, *Occupation* can be mapped to discrete values $\{0, 1, 2, \ldots\}$ as well as to one-hot encoded vectors. This renders methods which heavily rely on the variable values, such as ANMs, unusable. While extensions of ANMs to the discrete setting exist, they still utilize the variable values and are not robust to permuting the labels of the variables. One related approach proposed in [24] is motivated by Occam's razor and proposes to use the Kolmogorov complexity to capture the complexity of the causal model, and assume that the true direction is "simple". As Kolmogorov complexity is not computable, the authors resort to a proxy, based on minimum description length.

Another line of work uses the idea that causes are *independent* from the causal mechanisms, which is called the *independence of cause and mechanism* assumption. The notion of independence should be formalized since comparison is between a random variable and a functional relation. In [23, 25], authors propose using information geometry within this framework to infer the causal direction in deterministic systems. Specifically, they create a random variable using the functional relation based on uniform distribution and utilize the hypothesis that this variable should be independent from the cause distribution.

2.4 Identifiability with Entropic Causality

Consider the problem of identifying the causal graph between two observed categorical variables X, Y. We assume for simplicity that both have n states, although this is not necessary for the results. Similar to the most of the literature, we make the causal sufficiency assumption, i.e., there are no latent confounders and also assume there is no selection bias. Then without loss of generality, if X causes Y, there is a deterministic f and an exogenous (unmeasured) variable E that is independent from X such that $Y = f(X, E)$, where $X \sim p(X)$ for some marginal distribution $p(X)$. Causal direction tells us that, if we intervene on X and set $X = x$, we get $Y = f(x, E)$ whereas if we intervene on Y and set $Y = y$, we still get $X \sim p(X)$ since Y does not cause X.

Algorithms that identify causal direction from data introduce an assumption on the model and show that this assumption does not hold in the wrong causal direction in general. Hence, checking for this assumption enables them to identify the correct causal direction. Entropic causality [30] also follows this recipe. They assume that the entropy of the exogenous variable is bounded in the true causal direction. We first present their relevant conjecture, then modify and prove as a theorem.

Conjecture 1 ([30]). *Consider the structural causal model $Y = f(X, E), X \in [n], Y \in [n], E \in [m]$ where $p(X), f, p(E)$ are sampled as follows: Let $p(X)$ be sampled uniformly randomly from the probability simplex in n dimensions Δ_n, and $p(E)$ be sampled uniformly randomly from the set of points in Δ_m that satisfy $H(E) \leq \log(n) + \mathcal{O}(1)$. Let f be sampled uniformly randomly from all mappings $f : [n] \times [m] \rightarrow [n]$. Then with high probability, any $\tilde{E} \perp\!\!\!\perp Y$ that satisfies $X = g(Y, \tilde{E})$ for some mapping $g : [n] \times [m] \rightarrow [n]$ entails $H(X) + H(E) < H(Y) + H(\tilde{E})$.*

In words, the conjecture claims the following: Suppose X causes Y with the SCM $Y = f(X, E)$. Suppose the exogenous variable E has entropy that is within an additive constant of $\log(n)$. Then, for most of such causal models, any SCM that generates the same joint distribution in the wrong causal direction, i.e., Y causes X, requires a larger amount of randomness than the true model. The implication would be that if one can compute the smallest entropy SCM in both directions, then one can choose the direction that requires smaller entropy as the true causal direction.

We modify their conjecture in two primary ways. First, we assume that the exogenous variable has constant entropy, i.e., $H(E) = \mathcal{O}(1)$. Unlike the conjecture, our result holds for any such E. Second, rather than the total entropy, we were able to prove identifiability by only comparing the entropies of the simplest exogenous variables in both directions $H(E)$ and $H(\tilde{E})$.[1] In Section 2.6, we demonstrate that both criteria give similar performance in practice.

Our technical result requires the following assumption on $p(X)$, which, for constant ρ and d guarantees that a meaningful subset of the support of $p(X)$ is *sufficiently uniform*. We will later show that this condition holds with high probability, if $p(X)$ is sampled uniformly randomly from the simplex.

Assumption 1 ((ρ, d)-uniformity). *Let X be a discrete variable with support $[n]$. Then there exists a subset S of size $|S| \geq dn$, such that $p(X = x) \in [\frac{1}{\sqrt{\rho}n}, \frac{\sqrt{\rho}}{n}], \forall x \in S$.*

Our following theorem establishes that entropy in the wrong direction scales with n.

Theorem 1 (Entropic Identifiability). *Consider the SCM $Y = f(X, E), X \perp\!\!\!\perp E$, where $X \in [n], Y \in [n], E \in [m]$. Suppose E is any random variable with constant entropy, i.e., $H(E) = c = \mathcal{O}(1)$. Let $p(X)$ satisfy Assumption 1(ρ, d) for some constants $\rho \geq 1, d > 0$. Let f be sampled uniformly randomly from all mappings $f : [n] \times [m] \rightarrow [n]$. Then, with high probability, any $\tilde{E}$ that satisfies $X = g(Y, \tilde{E}), \tilde{E} \perp\!\!\!\perp Y$ for some g, entails $H(\tilde{E}) \geq (1 - o(1)) \log(\log(n))$. Specifically, for any $0 < r < q$, $H(\tilde{E}) \geq \left(1 - \frac{1+r}{1+q}\right) (0.5 \log(\log(n)) - \log(1 + r) - \mathcal{O}(1)), \forall n \geq \nu(r, q, \rho, c, d)$ for some ν.*

Theorem 1 shows that when $H(E)$ is a constant, under certain conditions on $p(X)$, with high probability, the entropy of any causal model in the reverse direction will be at least $\Omega(\log(\log(n)))$. Specifically, if a constant fraction of the support of $p(X)$ contains probabilities that are not too far from $\frac{1}{n}$, our result holds. Note that *with*

[1] Entropy of the exogenous variable, or in the case of Conjecture 1 the entropy of the system, can be seen as a way to model complexity and the method can be seen as an application of Occam's razor. In certain situations, especially for ordinal variables, it might be suitable to also consider the complexity of the functions.

high probability statement is induced by the uniform measure on f, and it is relative to n. In other words, Theorem 1 states that the fraction of non-identifiable causal models goes to 0 as the number of states of the observed variables goes to infinity. If a structure on the function is available in the form of a prior that is different from uniform, this can potentially be incorporated in the analysis although we expect calculations to become more tedious.

Through the parameters r, q we obtain a more explicit trade-off between the lower bound on entropy and how large n should be for the result. $\nu(r, q, \rho, c, d)$ is proportional to q and inversely proportional to r. The explicit form of ν is given in Proposition 1 in the supplement.

We next describe some settings where these conditions hold: We consider the cases when $p(X)$ has bounded element ratio, $p(X)$ is uniformly randomly sampled from the simplex, or $H(X)$ is large.

Corollary 1. *Consider the SCM in Theorem 1. Let $H(E) = c = \mathcal{O}(1)$ and f be sampled uniformly randomly. Let $p(x)$ be such that either (a) $\frac{\max_x p(x)}{\min_x p(x)} \leq \rho$, or (b) $p(x)$ is sampled uniformly randomly from the simplex Δ_n, or (c) $p(X)$ is such that $H(X) \geq \log(n) - a$ for some $a = \mathcal{O}(1)$.*

Then, with high probability, any $\tilde{E}$ that satisfies $X = g(Y, \tilde{E}), \tilde{E} \perp\!\!\!\perp Y$ for some deterministic function g entails $H(\tilde{E}) \geq 0.25 \log(\log(n)) - \mathcal{O}(1)$. Thus, there exists n_0 (a function of ρ, c) such that for all $n \geq n_0$, the causal direction is identifiable with high probability.

The proof is given in Section 2.9.1. Note that there is no restriction on the support size of the exogenous variable E.

Proof Sketch of Theorem 1. The full proof can be found in Appendix 2.9.1.

1. Bound $H(\tilde{E})$ via $H(\tilde{E}) \geq H(X|Y = y), \forall y \in [n]$.

2. Characterize the sampling model of f as a balls-and-bins game, where each realization of Y corresponds to a particular bin, each combination $(X = i, E = k)$ corresponds to a ball.

3. Identify a subset of “good" bins $\mathcal{U} \subseteq [m]$. Roughly, a bin is “good" if it does not contain a large mass from the balls other than the ones in $\{(i, 1) : i \in S\}$.

4. Show one of the bins in $\mathcal{U}$, say $y = 2$, has many balls from $\{(i, 1) : i \in S\}$.

5. Bound the contribution of the most-probable state of E to the distribution $p(X|Y = 2)$.

6. Characterize the effect of the other states of E and identify a support for X contained in S on which the conditional entropy can be bounded. Use this to lower bound for $H(X|Y = 2)$.

Conditional Entropy Criterion: From the proof of Proposition 1 in Appendix 2.9.1, we have $H(\tilde{E}) \geq \max_y H(X|Y = y) \geq (1 - o(1)) \log(\log(n))$. Further, we have $\max_x H(Y|X = x) \leq H(E) \leq c = \mathcal{O}(1)$. Hence not only is $H(\tilde{E}) > H(E)$ for large enough n, but $\max_y H(X|Y = y) > \max_x H(Y|X = x)$ as well. Therefore, under the assumptions of Theorem 1, $\max_y H(X|Y = y)$ and $\max_x H(Y|X = x)$ are sufficient to identify the causal direction:

Corollary 2. *Under the conditions of Theorem 1, we have that* $\max\limits_y H(X|Y{=}y) > \max\limits_x H(Y|X{=}x)$.

2.5 Entropic Causality with Finite Number of Samples

In the previous section, we provided identifiability results assuming that we have access to the joint probability distribution of the observed variables. In any practical problem, we can only access a set of samples from this joint distribution. If we assume we can get independent, identically distributed samples from $p(x, y)$, how many samples are sufficient for identifiability?

Given samples from N i.i.d. random variables $\{(X_i, Y_i)\}_{i\in[N]}$ where $(X_i, Y_i) \sim p(x, y)$, consider the plug-in estimators $\hat{p}(y) := \frac{1}{\{N\}} \sum_{i=1}^{N} \mathbb{1}_{\{Y_i=y\}}$ and $\hat{p}(x, y) :=$

$\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}_{\{X_i=x\}}\mathbb{1}_{\{Y_i=y\}}$ and define the estimator of the conditional $p(x|y)$ as $\hat{p}(x|y) \coloneqq \frac{\hat{p}(x,y)}{\hat{p}(y)}$. Define $\hat{p}(x)$ and $\hat{p}(y|x)$ similarly.

Definition 1. *The minimum entropy coupling of t random variables $U_1, U_2, \ldots, U_t$ is the joint distribution $p(u_1, \ldots, u_t)$ with minimum entropy that respects the marginal distributions of $U_i, \forall i$.*

The algorithmic approach of [30] relies on minimum entropy couplings. Specifically, they show the following equivalence: Given $p(x, y)$, let E be the minimum entropy exogenous variable such that $E \perp\!\!\!\perp X$, and there exists an f such that $Y = f(X, E), X \sim p(x)$ induces $p(x, y)$. Then the entropy of the minimum entropy coupling of the distributions $\{p(Y|x) : x \in [n]\}$ is equal to $H(E)$.

Therefore, understanding how having a finite number of samples affects the minimum entropy couplings allows us to understand how it affects the minimum entropy exogenous variable in either direction. Suppose $|\hat{p}(y|x) - p(y|x)| \le \delta, \forall x, y$ and $|\hat{p}(x|y) - p(x|y)| \le \delta, \forall x, y$. Given a coupling for distributions $p(Y|x)$, we construct a coupling for $\hat{p}(Y|x)$ whose entropy is not much larger. As far as we are aware, the minimum entropy coupling problem with sampling noise has not been studied.

Consider the minimum entropy coupling problem with n marginals $\mathbf{p}_k = [p_k(i)]_{i\in[n]}$, $k \in [n]$. Let $p(i_1, i_2, \ldots, i_n)$ be a valid coupling, i.e., $\sum_{j\neq k}\sum_{i_j=1}^{n} p(i_1, i_2, \ldots, i_n) = p_k(i_k)$, $\forall k, i_k$. Consider the marginals with sampling noise shown as $\hat{\mathbf{p}}_k = [\hat{p}_k(i)]_{i\in[n]}, k \in [n]$. Suppose $|\hat{p}_k(i) - p_k(i)| \le \delta, \forall i, k$. The following is shown in Section 2.9.1 of the supplement.

Theorem 2. *Let p be a valid coupling for distributions $\{\mathbf{p}_i\}_{i\in[n]}$, where $\mathbf{p}_i \in \Delta_n, \forall i \in [n]$. Suppose $\{\mathbf{q}_i\}_{i\in[n]}$ are distributions such that $|\mathbf{q}_i(j) - \mathbf{p}_i(j)| \le \delta, \forall i, j \in [n]$. If $\delta \le \frac{1}{n^2\log(n)}$, then there exists a valid coupling q for the marginals $\{\mathbf{q}_i\}_{i\in[n]}$ such that $H(q) \le H(p) + e^{-1}\log(e) + 2 + o(1)$.*

Theorem 2 shows that if the l_∞ norm between the conditional distributions and their empirical estimators are bounded by $\delta \le \frac{1}{n^2\log(n)}$, there exists a coupling that is within 3 bits of the optimal coupling on true conditionals. To guarantee this with the plug-in estimators, we have the following:

Lemma 1. *Let $X \in [n], Y \in [n]$ be two random variables with joint distribution $p(x, y)$. Let $\alpha = \min\{\min_i p(X = i), \min_j p(Y = j)\}$. Given N samples $\{(X_i, Y_i)\}_{i\in[N]}$ from independent identically distributed random variables $(X_i, Y_i) \sim p(x, y)$, let $\hat{p}(X|Y = y)$, $\hat{p}(Y|X = x)$ be the plug-in estimators of the conditional distributions. If $N = \Omega(n^4\alpha^{-2}\log^3(n))$, then $|\hat{p}(y|x)-p(y|x)| \leq \frac{1}{n^2\log(n)}$ and $|\hat{p}(x|y)-p(x|y)| \leq \frac{1}{n^2\log(n)}, \forall x, y$ with high probability.*

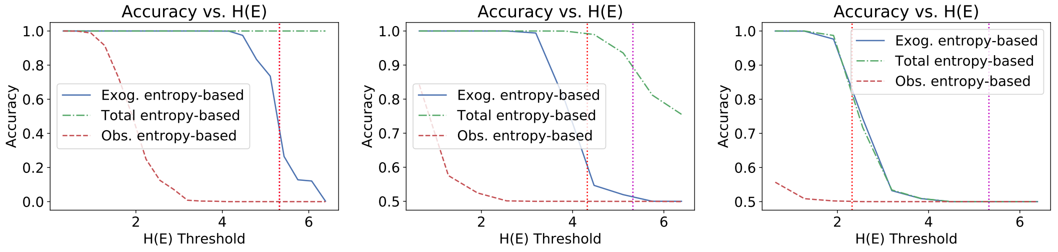
Next, we have our main identifiability result using finite number of samples:

Theorem 3 (Finite sample identifiability)**.** *Let $\mathcal{A}$ be an algorithm that outputs the entropy of the minimum entropy coupling. Consider the SCM in Theorem 1. Suppose E is any random variable with constant entropy, i.e., $H(E)=c=\mathcal{O}(1)$. Let $p(X)$ satisfy Assumption 1(ρ, d) for some constants $\rho \geq 1, d > 0$. Let f be sampled uniformly randomly from all mappings $f : [n] \times [m] \to [n]$. Let $\alpha = \min\{\min_i p(X = i), \min_j p(Y = j)\}$. Given $N = \Omega(n^4\alpha^{-2}\log^3(n))$ samples, let $\hat{p}(X|y), \hat{p}(Y|x)$ be the plug-in estimators for the conditional distributions. Then, for sufficiently large n, $\mathcal{A}(\{\hat{p}(X|y)\}_y) > \mathcal{A}(\{\hat{p}(Y|x)\}_x)$ with high probability.*

From the equivalence between minimum entropy couplings and minimum exogenous entropy, Theorem 3 shows identifiability of the causal direction using minimum-entropy exogenous variables. Similar to Corollary 1, the result holds when $p(X)$ is chosen uniformly randomly from the simplex:

Corollary 3. *Consider the SCM in Theorem 1, where $H(E)=c=\mathcal{O}(1)$, f is sampled uniformly randomly. Let $p(X)$ be sampled uniformly randomly from the simplex Δ_n. Given $N = \Omega(n^8\log^5(n))$ samples, let $\hat{p}(X|Y = y)$, $\hat{p}(Y|X = x)$ be the plug-in estimators for the conditional distributions. Then, for large enough n, $\mathcal{A}(\{\hat{p}(X|Y = y)\}_y) > \mathcal{A}(\{\hat{p}(Y|X = x)\}_x)$ with high probability.*

Conditional Entropy Criterion with Finite Samples: Note that the sample complexity in Theorem 3 scales with α^{-2} where $\alpha \coloneqq \min\{\min_i p(X=i), \min_j p(Y=j)\}$. If either of the marginal distributions are not strictly positive, this can make the bound of Theorem 3 vacuous. To address this, we use an internal result from the



(a) (b) (c)

Figure 2-2: m : number of states of X, n : number of states of Y in causal graph $X \rightarrow Y$. **(a)** $n = 40, m = 40$. Accuracy on simulated data: *Obs. entropy-based* declares $X \rightarrow Y$ if $H(X) > H(Y)$ and $Y \rightarrow X$ otherwise; *Exog. entropy-based* compares the exogenous entropies in both direction and declares $X \rightarrow Y$ if the exogenous entropy for this direction is smaller, and $Y \rightarrow X$ otherwise; *Total entropy-based* compares the total entropy of the model in both directions and declares the direction with smaller entropy as the true direction as proposed in [30]. **(b)** uses uniform mixture data from when $m = 40, n = 20$ and $m = 20, n = 40$. Similarly for **(c)** for $m = 40, n = 5$ and $m = 5, n = 40$. Magenta and red dashed vertical lines show $\log_2(\min\{m,n\})$ and $\log_2(\max\{m,n\})$, respectively.

proof of Theorem 1. In the proof we show that for some i, $p(Y=i) = \Omega(\frac{1}{n})$ and $H(X|Y=i) = \Omega(\log(\log(n)))$. Then, it is sufficient to obtain enough samples to accurately estimate $p(X|Y=i)$. Even though i is not known a priori, since $p(Y=i) = \Omega(\frac{1}{n})$, estimating conditional entropies $H(X|Y=j)$ where the number of samples $|\{(x, Y=j)\}_x|$ exceeds a certain threshold guarantees that $p(X|Y=i)$ is estimated accurately. We have the following result:

Theorem 4 (Finite sample identifiability via conditional entropy)**.** *Consider the SCM in Theorem 1, where $H(E) = c = \mathcal{O}(1)$, f is sampled uniformly randomly. Let $p(X)$ satisfy Assumption 1(ρ, d) for some constants $\rho \geq 1, d > 0$. Given $N = \Omega(n^2 \log(n))$ samples, let N_x be the number of samples where $X = x$ and similarly for N_y. Let $\hat{H}$ denote the entropy estimator of [64]. Then, for n large enough, $\max_{\{y:N_y \geq n\}} \hat{H}(X|Y=y) > \max_{\{x:N_x \geq n\}} \hat{H}(Y|X=x)$ with high probability.*

Theorem 4 shows that $\mathcal{O}(n^2 \log(n))$ samples are sufficient to estimate the large conditional entropies of the form $H(Y|x), H(X|y)$, which is sufficient for identifiability even for sparse $p(x,y)$.

2.6 Experiments

In this section, we conduct several experiments to evaluate the robustness of the framework. Complete details of each experiment are provided in the supplementary material. Unless otherwise stated, the greedy minimum entropy coupling algorithm of [30] is used to approximate $H(E)$ and $H(\tilde{E})$.

Implications of Low-Exogenous Entropy Assumption. We investigate the implications of this assumption. Specifically, one might ask if having low exogenous entropy implies $H(X)>H(Y)$. This would be unreasonable, since there is no reason for cause to always have the higher entropy.

In Figure 2-2, we evaluate the accuracy of the algorithm on synthetic data for different exogenous entropies $H(E)$. To understand the impact of the assumption on $H(X), H(Y)$, in addition to comparing exogenous entropies (*Exog. entropy-based*) and total entropies (*total entropy-based*) [30], we also show the performance of a simple baseline that compares $H(X)$ and $H(Y)$ (*obs. entropy-based*) and declares $X \rightarrow Y$ if $H(X) > H(Y)$ and vice versa.

We identify three different regimes, e.g., see Figure 2-2a: Regime 1: If $H(E) < 0.2\log(n)$, we get $H(X) > H(Y)$ most of the time. All methods perform very well in this regime which we can call *almost deterministic*. Regime 2: If $0.2\log(n) < H(E) < 0.6\log(n)$, accuracy of *obs. entropy-based* method goes to 0 since, on average, we transition from the regime where $H(X) > H(Y)$ to $H(X) < H(Y)$. Regime 3: $0.6\log(n) < H(E) < 0.8\log(n)$ where $H(X) < H(Y)$ most of the time. As can be seen, *total entropy-based* and *exog. entropy-based* methods both show (almost) perfect accuracy in Regime $1, 2, 3$ whereas *obs. entropy-based* performs well only in Regime 1.

We also evaluated the effect of the observed variables having different number of states on mixture data in Figure 2-2b, 2-2c. In this case, framework performs well up until about $0.8\log(\min\{m, n\})$.

Relaxing Constant Exogenous-Entropy Assumption. In Section 2.4, we demonstrated that the entropic causality framework can be used when the exogenous randomness is a constant, relative to the number of states n of the observed variables.

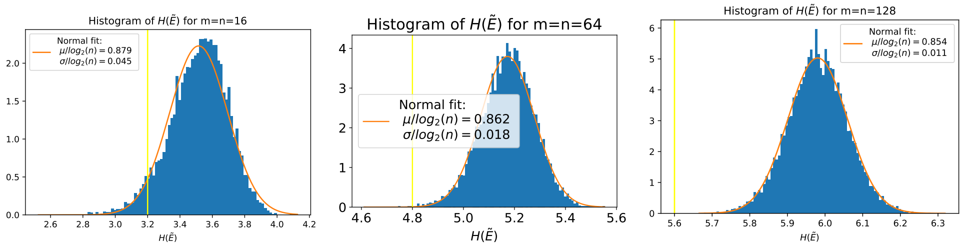
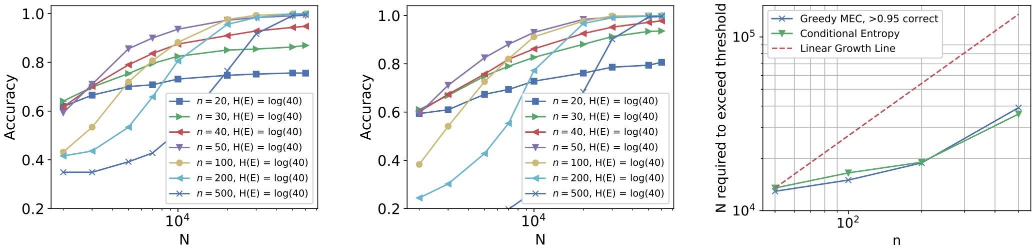


Figure 2-3: Histogram of $H(\tilde{E})$ when $H(E) \approx 0.8 \log_2(n)$. Yellow line shows $x = 0.8 \log_2(n)$



(a) Identification via conditional entropies ($H(E) \approx \log(40)$).

(b) Identification via MEC algorithm ($H(E) \approx \log(40)$).

(c) Number of samples vs. support size of observed variables.

Figure 2-4: (a) Probability of correctly discovering the causal direction $X \to Y$ as a function of n and number of samples N, using the conditional entropies as the test. (b) Probability of correctly discovering the causal direction $X \to Y$ using the greedy MEC algorithm. (c) Samples N required to reach 95% correct detection as a function of n, derived from the plots in Figure 2-4a and Figure 2-4b.

For very high dimensional variables, this might be a strict assumption. In this section, we conduct synthetic experiments to evaluate if entropic causality can be used when $H(E)$ scales with n. In particular, we test for various $\alpha < 1$ the following: *Is it true that the exogenous entropy in the wrong direction will always be larger, if the true exogenous entropy is* $\leq \alpha log(n)$? For $\alpha = \{0.2, 0.5, 0.8\}$, we sampled 10k $p(E)$ from Dirichlet distribution such that $H(E) \approx \alpha \log(n)$ and calculated exogenous entropy in the wrong direction $H(\tilde{E})$. Figure 2-3 shows the histograms of $H(\tilde{E})$ for $\alpha = 0.8$ and $n = \{16, 64, 128\}$. We observe that $H(\tilde{E})$ tightly concentrates around $\beta \log(n)$ for some $\beta > \alpha$. For reference, $\alpha \log(n)$ is shown by the vertical yellow line. Similar results are observed for other α values which are provided in the supplementary material.

Effect of Finite Number of Samples. In Section 2.5, we identified finite sample bounds for entropic causality framework, both using the exogenous entropies $H(E), H(\tilde{E})$ and using conditional entropies of the form $\max_y H(X|Y = y), \max_x H(Y|X = x)$. We now test if the bounds are tight.

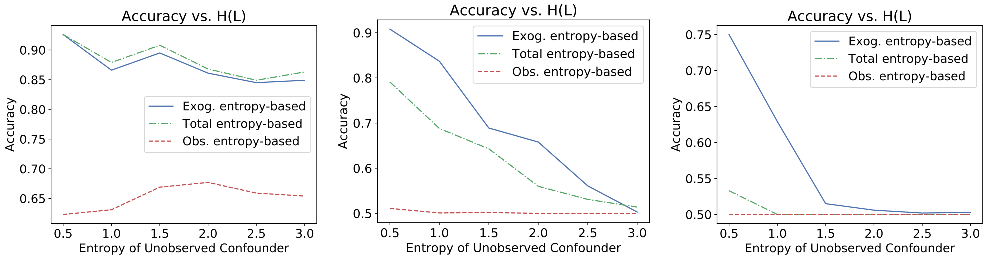
We observe two phases and a transition phenomenon in between. The first phase occurs for small values of n, for $n \in \{20, 30, 40\}$. Here, the fraction of identifiable causal models does not reach 1 as the number of samples is increased, but saturates at a smaller value. This is expected since exogenous noise is relatively high, i.e., $H(E) \geq \log(n)$. For $n > 40$, or equivalently, when $H(E) \leq \log(n)$, increasing number of samples increases accuracy to 1, as expected.

The greedy MEC criterion has slightly better performance (by $\approx 5\%$), indicating more robustness. This may be due to a gap between $H(\tilde{E})$ and $H(X|Y = y)$ since greedy-MEC output is not limited by $\log(n)$ unlike conditional entropy. In contrast to the $\tilde{O}(n^8)$ bound, the number of samples needed has a much better dependence on n. Figure 2-4c includes a dashed linear growth line for comparison.

Effect of Confounding The equivalence between finding the minimum entropy exogenous variable and finding the minimum entropy coupling relies on the assumption that there are no unobserved confounders in the system. Despite lack of theory, it is useful to experimentally understand if the method is robust to *light confounding*. One way to assess the effect of confounding is through its entropy: If a latent confounder L is a constant, i.e., it has zero entropy, it does not affect the observed variables. In this section, we simulate a system with light confounding by limiting the entropy of the latent confounder and observing how quickly this degrades the performance of the entropic causality approach.

The results are given in Figure 2-5. The setting is similar to that of Figure 2-2. We set $H(E) \approx 2$ and show accuracy of the method as entropy of the latent L is increased. Perhaps surprisingly, the effect of increasing the entropy of the confounder is very similar to the effect of increasing the entropy of the exogenous variable. This shows that the method is robust to light latent confounding.

Tübingen Cause-Effect Pairs In [30], authors employed the total entropy-based algorithm on Tübingen data [43] and showed that it performs similar to additive noise



(a) (b) (c)

Figure 2-5: Accuracy on simulated data with *light* confounding. Number of states and data are identical to those in Figure 2-2. We use exogenous entropy of 2 bits and add a confounder L. This can be interpreted as replacing some bits of the exogenous variable in Figure 2-2 with those of a latent confounder. Surprisingly, performance for $H(E)\!=\!2, H(L)\!=\!t$ is similar to the performance when $H(E)\!=\!2+t$ in Figure 2-2. This indicates that the proposed method is robust to latent confounders, as long as the total exogenous and confounder entropy is not very close to $\min\{\log(n), \log(m)\}$.

5-state quantization	Threshold (× log support)	0.7	0.8	0.85	0.9	1.0	1.2
	# of pairs	14	25	34	42	57	85
	Accuracy (%)	85.7	64.0	58.8	57.1	63.2	60.0
10-state quantization	Threshold (× log support)	0.7	0.8	0.85	0.9	1.0	1.2
	# of pairs	13	23	34	46	67	85
	Accuracy (%)	84.6	73.9	70.6	63.0	61.2	56.5
20-state quantization	Threshold (× log support)	0.7	0.8	0.85	0.9	1.0	1.2
	# of pairs	12	21	41	52	76	85
	Accuracy (%)	75.0	61.9	53.7	51.9	51.3	49.4

Table 2.1: Performance on Tübingen causal pairs with low exogenous entropy in at least one direction.

models with an accuracy of 64%. Next, we test if entropic causality can be used when we only compare exogenous entropies.

The challenge of applying entropic causality on Tübingen data is that most of the variables are continuous. Therefore, before applying the framework, one needs to quantize the data. The authors chose a uniform quantization, requiring both variables have the same number of states. We follow a similar approach. For $b \in \{5, 10, 20\}$, the value of n is chosen for both X, Y as the minimum of b, $N/10$, N_x^{uniq} and N_y^{uniq}, where N is the number of samples available for pair X, Y and N_x^{uniq}, N_y^{uniq} are the number of unique realizations of X, Y, respectively.

As a practical check for the validity of our key assumption, we make a decision based on the following: For a threshold t, algorithm makes a decision only for pairs for which either $H(E)\!\leq\! t\log(n)$ or $H(\tilde{E})\!\leq\! t\log(n)$. We report the accuracies in Table 2.1.

As we expect, for stricter thresholds, accuracy is improved, supporting the assumption that in real data, the direction with the smaller exogenous entropy is likely to be the true direction. Performance is most consistent with $b = 10$.

To check the stability of performance in regards to quantization, we conducted an experiment where we perturb the quantization intervals and take majority of 5 independent decisions. This achieves qualitatively similar (it is sometimes better, sometimes worse) performance shown in Table 2.3 in the appendix. Exploring best practices for how to quantize continuous data is an interesting avenue for future work.

We now compare performance with other leading methods on this dataset. The total-entropy approach for Entropic Causal Inference achieved 64.21% accuracy at 100% decision rate in [30]. ANM methods are evaluated on this data in [43], where they emphasize two ANM methods with consistent performance that achieve $63 \pm 10\%$ and $69 \pm 10\%$ accuracy. IGCI methods are also evaluated in [43] and were found to vary greatly with implementation and perturbations of data. No IGCI method had consistent performance. LiNGAM methods are evaluated in [21] and reported nonlinear approaches with 62% and 69% accuracy. Of these, only Entropic Causal Inference and IGCI can handle categorical data. As comparison with different approaches is difficult given limited data, we suggest assessing the MEC in both directions when deciding how to use our approach in combination with other methods.

2.7 Discussion

In this section we discuss several aspects of our method in relation with prior work. First, note that our identifiability result holds *with high probability* under the measure induced by our generative model. This means that, even under our assumptions, not all causal models will be identifiable. However, the non-identifiable fraction vanishes as n, i.e., the number of states of X, Y goes to infinity. In essence, this is similar to many of the existing identifiability statements that show identifiability except for an adversarial set of models [20]. Specifically in [30], the authors show that under the assumption that the exogenous variable has small support size, causal direction is identifiable with probability 1. This means that the set of non-identifiable models has Lebesgue measure zero. This is clearly a stronger identifiability statement. However, this is not surprising if we compare the assumptions: Bounding the support size of a variable bounds its entropy, but not vice verse. Therefore, our assumption can be seen as a relaxation of the assumption of [30]. Accordingly, a weaker identifiability result is expected.

Next, we emphasize that our key assumption, that in the true causal direction the exogenous variable has small entropy, is not universal, i.e., one can construct cause-effect pairs where the anti-causal direction requires less entropy. [22] provides an example scenario: Consider a ball traveling at a fixed and known velocity from the initial position X towards a wall that may appear or disappear at a known position with some probability. Let Y be the position of the ball after a fixed amount of time. Clearly we have $X \to Y$. If the wall appears, the ball ends up in a different position (y_0) from the one it would if the wall does not (y_1). Then the mapping $X \to Y$ requires an exogenous variable to describe the behavior of the wall. However, simply by looking at the final position, we can infer whether wall was active or not, and accordingly infer what the initial position was deterministically. This shows that our key assumption is not always valid and should be evaluated depending on the application in mind.

Finally note that the low-entropy assumption should not be enforced on the exogenous variable of the cause, since this would imply that X has small entropy. This brings about a conceptual issue to extend the idea to more than two variables: Which variables' exogenous noise should have small entropy? For that setting, we believe the original assumption of [30] may be more suitable: Assume that the total entropy of the system is small. In the case of more than two variables, this means total entropy of all the exogenous variables is small, without enforcing bounds on specific ones.

2.8 Conclusion

In this work, we showed the first identifiability result for learning the causal graph between two categorical variables using the entropic causal inference framework. We also provided the first finite-sample analysis. We conducted extensive experiments to conclude that the framework, in practice, is robust to some of the assumptions required by theory, such as the amount of exogenous entropy and causal sufficiency assumptions. We evaluated the performance of the method on Tübingen dataset.

2.9 Supplementary Material

2.9.1 Proofs

Proof of Theorem 1

Step 1. Bounding $H(\tilde{E})$ by $H(\tilde{E}) \geq H(X|Y=y), \forall y$: Consider any $\tilde{E} \perp\!\!\!\perp Y$ for which there exists a deterministic map g such that $X = g(\tilde{E}, Y)$. We have

$$\begin{aligned} p(X=x|Y=y) &= p(g(\tilde{E},Y)=x|Y=y) \\ &= p(g(\tilde{E},y)=x) = p(g_y(\tilde{E})=x), \end{aligned}$$

for $g_y(e) \coloneqq g(e,y), \forall e, y$, since $\tilde{E} \perp\!\!\!\perp Y$. Due to data processing inequality, it follows that $H(\tilde{E}) \geq H(X|Y=y)$.

In [30], this analysis is used to show that the minimum entropy exogenous variable $\tilde{E}$ can be obtained by solving the minimum entropy coupling problem on the conditional distributions $p(X|Y=y)$. Here, we use the conditional entropies to lower bound the entropy of the exogenous variable $\tilde{E}$. Therefore, in the rest of our analysis we attempt to show that under the given assumptions, with high probability, $H(X|Y=y)$ is large for some value of y.

Step 2. Generative process as a balls and bins game: In order to analyze the conditional distributions $p(X|Y=y)$ we relate the generative model to a balls and bins game:

Consider a deterministic map $f : [n] \times [m] \rightarrow [n]$. Let $p(X=i) = x_i$ and $p(E=k) = e_k$. Without loss of generality, assume that X and E are labeled in decreasing probability order. In other words, $e_k \geq e_l$ if $k < l$ and $x_i \geq x_j$ if $i < j$.[2] Let $\mathbf{M}$ be the matrix defined as $\mathbf{M}_{i,k} \coloneqq f(i,k)$. The probability distribution $p(Y|X)$ is determined by the causal mechanism, i.e., the structural equation $Y = f(X,E)$. The conditional distributions in the wrong causal direction, i.e., $p(X|Y)$ can then be

[2] This relabeling of X, E is without loss of generality since realization of f is symmetric across rows and columns.

calculated as follows:

$$p(X = i|Y = j) = \frac{1}{Z} x_i \sum_{k=1}^{m} \mathbb{1}_{\{\mathbf{M}_{i,k}=j\}} e_k.$$

$Z = \sum\limits_{i=1}^{n} x_i \sum\limits_{k=1}^{m} \mathbb{1}_{\{\mathbf{M}_{i,k}=j\}} e_k$ is the normalizing constant.

To sample f uniformly randomly from all the mappings is equivalent to filling the entries of $\mathbf{M}$ independently and uniformly randomly from $\mathcal{Y} = [n]$. A small example is given in Table 1, which shows a realization of f through matrix $\mathbf{M}$, and illustrates how this affects $p(X|Y = 1)$.

	$\mathcal{E}$	1	2	3	4	5
$\mathcal{X}$	PMF of E / PMF of X	e_1	e_2	e_3	e_4	e_5
1	x_1	2	3	2	1	1
2	x_2	3	2	3	3	1
3	x_3	3	1	2	3	2

	$\mathbb{P}(X = x\|Y = 1)$
$x = 1$	$\frac{x_1(e_4+e_5)}{Z}$
$x = 2$	$\frac{x_2 e_5}{Z}$
$x = 3$	$\frac{x_3 e_2}{Z}$

Table 2.2: Left: Balls and bins representation of function $f : \mathcal{X} \times \mathcal{E} \to \mathcal{Y}$, where $\mathcal{X} = \mathcal{Y} = [3]$ and $\mathcal{E} = [5]$. The function values for a given $X = i, E = k$ can be seen as realizations of a two dimensional balls and bins game. Right: Conditional probability values of X given $Y = 1$ for the given function. $Z = x_1(e_1 + e_3) + x_2(e_2) + x_3(e_5)$ is the normalization constant, which also gives $\mathbb{P}(Y = 1)$.

Any realization of f corresponds to a realization of matrix $\mathbf{M}$. The first column is of special interest to us because it corresponds to the value of E with the highest probability. The realization of $\mathbf{M}$ can be thought of as a balls and bins process, with the cells corresponding to balls and each entry $\mathbf{M}_{i,k}$ corresponding to which bin that cell's ball landed in.

Step 3. Identify a set of "good" bins: Each coordinate (i, k) is a ball and the value of $\mathbf{M}_{i,k}$ is the identity of the bin this ball is placed in. We utilize the existence of a set S as described in the theorem statement as follows: We focus on the set of balls corresponding to the cells $(i, 1)$ for $i \in S$. Our goal is to identify a bin which contains a large fraction of these balls. We also want this bin to not contain too much probability mass from balls outside of the set S in order to get a close bound in **Step**

6.

Recall that each bin y contains mass $x_i e_k$ when $\mathbf{M}_{i,k} = y$. To restrict our search of a good bin, we first discard all the bins that contain a large mass from entries of $\mathbf{M}$ that are either in rows corresponding to $x \notin S$ or columns other than the first column. Let $p(X, Y, E)$ represent the joint distribution between X, Y, E. Then we discard every value of y where $\sum_{x \notin S} \sum_{e=1}^{m} p(x, y, e) + \sum_{x \in S} \sum_{e=2}^{m} p(x, y, e)$ is large. We pick the threshold of $\frac{2}{n}$ and define the set $\mathcal{B}$ accordingly:

$$\mathcal{B} = \left\{ y : \sum_{x \notin S} p(x, y) + \sum_{x \in S} p(x, y, E > 1) > \frac{2}{n} \right\}.$$

We know that $|\mathcal{B}| \leq \frac{n}{2}$, since otherwise the total mass would exceed 1.[3] Let $\mathcal{U} := [n] \backslash \mathcal{B}$. Then $|\mathcal{U}| \geq n/2$. Note that $\mathcal{B}$ and $\mathcal{U}$ are determined in a manner not affected by the realized values of $\mathbf{M}_{x,1}$ for $x \in S$. We will next focus on only the values of $y \in \mathcal{U}$, and later quantify the following claim: A significant fraction of the probability mass that falls in any bin in $\mathcal{U}$ is due to entries from $\mathbf{M}_{x,1}$ for $x \in S$. Therefore, for one of these bins $y \in \mathcal{U}$, we can focus on obtaining a lower bound of $H(X|Y = y, X \in S, \mathbf{M}_{X,1} = y)$ to later show that $H(X|Y = y)$ cannot be much smaller.

Step 4. Show a bin from $\mathcal{U}$ has many balls from the first column of M and rows in S: We focus our attention to the balls in S and bins in $\mathcal{U}$. We want to show that $\exists y \in \mathcal{U}$ such that $\mathbf{M}_{x,1} = y$ for a large number of values of $x \in S$. Recall that since $|S| \geq dn$, we have at least dn balls falling into n bins. Moreover, since $|\mathcal{U}| \geq n/2$, at least $n/2$ of these bins are “good" for us. First, we show that, with high probability, at least $\frac{dn}{4}$ of the dn balls fall in the bins in $\mathcal{U}$.

Lemma 2. *Consider the process of uniformly randomly throwing $dn = \Theta(n)$ balls into n bins.*[4] *Let $\mathcal{U}$ be an arbitrary, fixed subset of bins with size $|\mathcal{U}| \geq \frac{n}{2}$. Then with high probability, at least $\frac{dn}{4}$ balls fall into the bins in $\mathcal{U}$. Moreover, these balls are also uniformly randomly thrown.*

[3] The probabilities we sum correspond to disjoint events, hence the total probability cannot exceed 1.

[4] Uniformity follows from uniformity of f.

The above lemma, proven in Appendix 2.9.1 is directly applicable to our setting, even though $\mathcal{U}$ is a random variable. This is because the realization of the entries of $\mathbf{M}$ outside the rows S or outside the first column, which determines the set $\mathcal{U}$ are independent from the entries in $\mathbf{M}$ in the rows S and in the first column. In other words, how balls are thrown into the bins in $\mathcal{U}$ is not affected by how $\mathcal{U}$ is chosen.

We want to use this to show that there is a bin $y \in \mathcal{U}$ such that the conditional distribution $p(X|Y=y)$ is due to many balls $x \in S$ where $\mathbf{M}_{x,1} = y$. We have shown that with high probability at least $\frac{dn}{4}$ balls land in bins corresponding to $y \in \mathcal{U}$. We apply a bound from Theorem 1 of [54], which implies that with high probability when there are b bins and ηb balls ($\eta = \Theta(1)$), the most loaded bin has at least $\frac{\ln(b)}{\ln(\ln(b))+\ln\left(\frac{1}{\eta}\right)}$ balls. We know that with high probability we have some number of balls in range $[\frac{nd}{4}, nd]$ in some number of good bins in range $[\frac{n}{2}, n]$. In terms of the established bound on the most loaded bin, this means $\eta \geq \frac{d}{4}$ and $b \in [\frac{n}{2}, n]$. If we substitute valid values of η and b that minimize the lower bound, we know that with high probability the heaviest loaded bin among $\mathcal{U}$ conditional distributions has at least $\frac{\ln(n)-\ln(2)}{\ln(\ln(n))+\ln(\frac{4}{d})}$ balls. Without loss of generality, suppose this bin has label 2. We show that $H(X|Y=2)$ is large using the above bound.

Step 5. Bounding $H(X|Y=2)$: Next, we obtain a lower bound for $H(X|Y=2)$. We utilize the following lemma, proved in Section 2.9.1 of the supplement:

Lemma 3. *Let X be a discrete random variable with distribution $[p_1, p_2, \ldots, p_n]$. Consider the random variable X' with distribution $[\frac{p_i}{\sum_{j \in S'} p_j}]_i$ for any $S' \subseteq [n]$. Then, $H(X) \geq \mu H(X')$, where $\mu = \sum_{i \in S'} p_i$.*

To use this lemma, we consider a specific distribution induced on the support of $X|Y=2$. First, let us define the following: For any subset $S' \subseteq [n], y \in [n]$, let $X_{S',y}$ be the discrete variable with the following distribution:

$$p(X_{S',y}{=}i){=}\frac{p(X=i|Y=y)}{\sum_{l \in S'} p(X=l|Y=y)}, \forall i \in S'. \tag{2.1}$$

We focus on $X_{S',2}$, where $S' = \{i : i \in S, \mathbf{M}_{i,1} = 2\}$. We first show that $H(X_{S',2})$ is large, and then show the total mass $\mu = \sum_{i \in S'} p(X=i|Y=2)$ that $X_{S',2}$ contributes to $(X|Y=2)$ is large, which allows us to use Lemma 3.

To show $H(X_{S',2})$ is large, we use the following lemma from [9]:

Lemma 4 (Theorem 2 of [9])**.** *Let X be a strictly positive discrete random variable on n states such that $\frac{\max_i p(X=i)}{\min_i p(X=i)} \leq \rho$. Then*

$$H(X) \geq \log(n) - \left(\frac{\rho \ln(\rho)}{\rho - 1} - 1 - \ln\left(\frac{\rho \ln(\rho)}{\rho - 1}\right)\right)\frac{1}{\ln(2)}.$$

To lower bound $H(X_{S',y})$ using the above lemma, we obtain an upper bound to $\rho' := \frac{\max_i p(X_{S',2}=i)}{\min_i p(X_{S',2}=i)}$ by utilizing our knowledge that $H(E) = c$. For each value $i \in S'$, we know that $\mathbf{M}_{i,1} = 2$. Thus, $p(X_{S',2} = i) \geq \frac{x_i e_1}{\mu}$. Also $p(X_{S',2} = i) \leq \frac{x_i \sum_{k=1}^{m} e_k}{\mu} = \frac{x_i}{\mu}$ and $\frac{\max_{i \in S} x_i}{\min_{i \in S} x_i} \leq \rho$. Therefore $\rho' \leq \frac{\max_i \frac{x_i}{\mu}}{\min_i \frac{x_i e_1}{\mu}} \leq \frac{\rho}{e_1}$.

In order to understand how small e_1 can be under the given constraints, we obtain a useful characterization for constant entropy distributions. The following lemma shows that the maximum probability value for any discrete distribution with constant entropy is a constant away from zero.

Lemma 5. *Let E be a discrete random variable with m states, with the probability distribution $[e_1, e_2, \ldots, e_m]$, where without loss of generality $e_i \geq e_j, \forall j > i$. If $H(E) \leq c$ then $e_1 \geq 2^{-c}$.*

The proof is given in Section 2.9.1 in the supplement.

Applying Lemmas 3-5, with some derivation we show in Section 2.9.1 of the supplement that:

Proposition 1 (**Step 6**)**.** *Under the conditions stated above,*

$$\begin{aligned} H(\tilde{E}) &\geq \max_y H(X|Y=y) \geq H(X|Y=2) \\ &\geq (1 - o(1))[\log(\log(n)) - \log(\log(\log(n))) - \mathcal{O}(1)]. \end{aligned}$$

Furthermore, to make the trade-off between the strength of the lower bound and *assumptions on n more explicit, when $n \geq \nu(r, q, \rho, c, d)$ with*

$$\nu(r, q, \rho, c, d) = \max\{4, e^{\left(\frac{4}{d}\right)^{1/r}}, 2e^{q^2 2^{2(c+1)} \rho}\},$$

we have

$$\begin{aligned} H(\tilde{E}) &\geq \max_y H(X|Y = y) \geq H(X|Y = 2) \\ &\geq \left(1 - \frac{1+r}{1+q}\right) (0.5 \log(\log(n)) - \log(1+r) - \mathcal{O}(1)) . \end{aligned}$$

This completes the proof of Theorem 1. □

Potential Improvements and Limitations: In our analysis, we use $\max_y H(X|Y = y)$ to bound $H(\tilde{E})$. One potential improvement might be obtained by considering the gap between $H(\tilde{E})$ and the collection $\{H(X|Y = y)\}_y$ for a given $p(x, y)$. [30] showed that the smallest $H(\tilde{E})$ is given by the minimum entropy coupling of the conditional distributions $\{p(X|Y = y)\}_y$. Follow-up works have developed minimum-entropy coupling algorithms [7,31,55] and obtained approximation guarantees. However there is currently no tight analysis characterizing this entropy gap.

Note that the original conjecture proposes that $H(E) \leq \log(n) + \mathcal{O}(1)$ is sufficient. This is a very strong statement and we believe, even if it is true, it requires a much deeper understanding on the minimum entropy couplings than is currently available in the literature. We do, however, provide evidence in Section 2.6 that $H(E) \leq \alpha \log(n)$ for $\alpha < 1$ seems sufficient for identifiability.

One point in our analysis that is related to this setting when $H(E)$ scales with n, is that we only considered the first column of the matrix $\mathbf{M}$, i.e., we have only taken into account the probability values of the form $x_i e_1$ contributing to the entropy of $H(X|Y = y)$. As long as the function f is sampled uniformly randomly in the considered generative model, this approach cannot give $H(\tilde{E}) \gg \log(\log(n))$ due to the support size of X being upper bounded by $\mathcal{O}(\log(n))$ with high probability from the balls and bins perspective. For when $H(E)$ is very small, we do expect this to be a reasonable approach as the remaining columns have very small probability values, hence very small impact. However, for going beyond the current analysis and for proving identifiability when $H(E)$ scales with n, we strongly believe that the effect of the remaining columns should be considered.

Proof of Lemma 2

Let ε be the event that less than $\frac{dn}{4}$ balls fall in the bins in $\mathcal{U}$. We provide an upper bound for the probability of this event $P(\varepsilon)$. Consider the indicator variables each corresponding to the event that a particular ball lands in $\mathcal{U}$. These indicator variables are independently and identically distributed, where each has probability $\frac{\mathcal{U}}{n} \geq \frac{1}{2}$ of being 1. We use Hoeffding's inequality to bound $P(\varepsilon)$. Let S_{dn} be the sum of the dn indicator variables (i.e., the number of the balls that land in bins corresponding to $\mathcal{U}$) and E_{dn} be the expected sum of the indicator variables ($E_{dn} = dn\left(\frac{\mathcal{U}}{n}\right)$).

$$P(\varepsilon) = P\left(S_{dn} < \frac{dn}{4}\right) \tag{2.2}$$

$$\leq P\left(|S_{dn} - E_{dn}| > \left|E_{dn} - \frac{dn}{4}\right|\right) \tag{2.3}$$

$$\leq P\left(|S_{dn} - E_{dn}| > \frac{dn}{2} - \frac{dn}{4}\right) \tag{2.4}$$

$$\leq P\left(|S_{dn} - E_{dn}| > \frac{dn}{4}\right) = 2e^{-\frac{dn}{8}} \tag{2.5}$$

(2.3) to (2.4) is due the fact that for all valid values of $\mathcal{U}$, it holds that $E_{dn} = dn(\frac{\mathcal{U}}{n}) \geq \frac{dn}{2}$. (2.5) is due to Hoeffding's inequality. As such, $P(\varepsilon) \leq 2e^{-\frac{dn}{8}}$. Thus, with high probability there are at least $\frac{dn}{4}$ balls that fall into bins corresponding to $\mathcal{U}$. Since balls are thrown independently and uniformly at random, conditioned on the balls that land in $\mathcal{U}$, they are thrown independently and uniformly at random. □

Proof of Lemma 3

Recall that $\mu = \sum_{i \in S'} p(X = i)$. We have

$$\begin{aligned}
H(X) &\geq \sum_{i \in S'} p(X = i) \log \left(\frac{1}{p(X = i)} \right) \\
&= \mu \left(\sum_{i \in S'} \frac{p(X = i)}{\mu} \log \left(\frac{1}{p(X = i)} \right) \right) \\
&\geq \mu \left(\sum_{i \in S'} \frac{p(X = i)}{\mu} \log \left(\frac{\mu}{p(X = i)} \right) \right) \\
&= \mu \left(\sum_{i \in S'} p(X' = i) \log \left(\frac{1}{p(X' = i)} \right) \right) \\
&= \mu H(X'). \qquad \square
\end{aligned}$$

Proof of Lemma 5

We show the contrapositive. Suppose that $p_1 \leq \varepsilon < 2^{-c}$. We have $p_i \leq p_1, \forall i \in [m]$. We consider all such distributions and find the one with smallest entropy:

$$\begin{aligned}
\min_{p_1 \geq p_2, \ldots \geq p_m} \quad & H([p_1, p_2, \ldots, p_m]) \\
\text{s.t.} \quad & \sum_i p_i = 1 \\
& \varepsilon \geq p_i \geq 0, \forall i \in [m]
\end{aligned} \tag{2.6}$$

For simplicity, suppose $\frac{1}{\varepsilon}$ is an integer. We show that the solution to the above optimization problem is strictly greater than c using majorization theory. For any given p, define the vector $u_p = [\sum_{j=1}^{i} p_j]_i$. Recall that a probability distribution p majorizes another distribution q if $u_p(i) \geq u_q(i), \forall i \in [m]$. Also if p majorizes q, we have $H(p) \leq H(q)$.

Consider all distributions in the feasible region of the above problem. For any p^*, consider the vector u_{p^*}. Clearly, $u_{p^*}(1) \geq \varepsilon$. Since $p_2 \leq p_1 < \varepsilon$, we have that $u_{p^*}(2) \leq 2\varepsilon$. Similarly, we have $u_{p^*}(i) \leq \varepsilon$. The uniform distribution achieves this upper bounding u vector, establishing that the uniform distribution majorizes every other distribution in the feasible set. Then for any distribution in the feasible region, we get that $H(p) \geq \log(\frac{1}{\varepsilon}) > c$.

Suppose $\frac{1}{\varepsilon}$ is not an integer. Let t be the largest integer such that $t\varepsilon \leq 1$. Then the above argument leads to the distribution with entropy

$$H = t\varepsilon \log\left(\frac{1}{\varepsilon}\right) + (1 - t\varepsilon)\log\left(\frac{1}{1 - t\varepsilon}\right). \tag{2.7}$$

Next, we show that if $\varepsilon < 2^{-c}$, above value is greater than c. We can rewrite

$$\begin{aligned} H &= t\varepsilon \log\left(\frac{1}{\varepsilon}\right) + (1 - t\varepsilon)\log\left(\frac{1}{1 - t\varepsilon}\right) && (2.8) \\ &\geq t\varepsilon \log\left(\frac{1}{\varepsilon}\right) + (1 - t\varepsilon)\log\left(\frac{1}{\varepsilon}\right) && (2.9) \\ &= \log\left(\frac{1}{\varepsilon}\right) > c && (2.10) \end{aligned}$$

since $1 - t\varepsilon \leq \varepsilon$. This concludes the proof. □

Proof of Proposition 1

By Lemma 5 we then know $\rho' \leq \frac{\rho}{e_1} \leq \rho 2^c$, and the size of the support of $X_{S',2}$ is the number of balls in the most loaded bin which is at least $\frac{\ln(n) - \ln(2)}{\ln(\ln(n)) + \ln(\frac{4}{d})}$. Using Lemma 4, we conclude $H(X_{S',2}) \geq \log(\frac{\ln(n) - \ln(2)}{\ln(\ln(n)) + \ln(\frac{4}{d})}) - (\frac{\rho 2^c \ln(\rho 2^c)}{\rho 2^c - 1} - 1 - \ln(\frac{\rho 2^c \ln(\rho 2^c)}{\rho 2^c - 1}))\frac{1}{\ln(2)}$.

Using our previous results, we know that $\min_{i \in S'} p(X = i, Y = 2) \geq \min_{i \in S'} e_1 x_i \geq \frac{e_1}{\sqrt{\rho n}} \geq \frac{2^{-c}}{\sqrt{\rho n}}$. Then, $p(X \in S', Y = 2) \geq (\frac{\ln(n) - \ln(2)}{\ln(\ln(n)) + \ln(\frac{4}{d})})(\frac{2^{-c}}{\sqrt{\rho n}}) = \frac{\ln(n) - \ln(2)}{(\ln(\ln(n)) + \ln(\frac{4}{d}))\sqrt{\rho n} 2^c}$. Additionally:

$$\begin{aligned} p(X \notin S', Y=2) = \sum_{i \in S^c} \sum_{j=1}^{m} p(X=i, Y=2, E=j) \\ + \sum_{i \in S, i \notin S'} \sum_{j=1}^{m} p(X=i, Y=2, E=j) \end{aligned} \tag{2.11}$$

$$\begin{aligned} &= \sum_{i \in S^c} \sum_{j=1}^{m} p(X=i, Y=2, E=j) \\ &+ \sum_{i \in S, i \notin S'} \sum_{j=2}^{m} p(X=i, Y=2, E=j) \end{aligned} \tag{2.12}$$

$$\begin{aligned} &\leq \sum_{i \in S^c} \sum_{j=1}^{m} p(X=i, Y=2, E=j) \\ &+ \sum_{i \in S} \sum_{j=2}^{m} p(X=i, Y=2, E=j) \leq \frac{2}{n}. \end{aligned} \tag{2.13}$$

We go from (2.11) to (2.12) by realizing that for any $i \in S$, $p(X = i, Y = 2, E = 1) > 0$ only if $\mathbf{M}_{x,1} = 2$ and thus $i \in S'$. We simplify (2.12) by definition of $\mathcal{U}$. As such, $p(X \in S'|Y = 2) = \frac{p(X \in S', Y=2)}{p(X \in S', Y=2) + p(X \notin S', Y=2)} \geq \frac{\frac{\ln(n) - \ln(2)}{(\ln(\ln(n)) + \ln(\frac{4}{d}))\sqrt{\rho} n 2^c}}{\frac{\ln(n) - \ln(2)}{(\ln(\ln(n)) + \ln(\frac{4}{d}))\sqrt{\rho} n 2^c} + \frac{2}{n}} = \frac{\ln(n) - \ln(2)}{\ln(n) - \ln(2) + (\ln(\ln(n)) + \ln(\frac{4}{d}))\sqrt{\rho} 2^{c+1}}$. Thus, we have shown that $H(X_{S',2}) \geq \log(\frac{\ln(n) - \ln(2)}{\ln(\ln(n)) + \ln(\frac{4}{d})}) - (\frac{\rho 2^c \ln(\rho 2^c)}{\rho 2^c - 1} - 1 - \ln(\frac{\rho 2^c \ln(\rho 2^c)}{\rho 2^c - 1}))\frac{1}{\ln(2)}$ and $P(X \in S', Y = 2) \geq \frac{\ln(n) - \ln(2)}{\ln(n) - \ln(2) + (\ln(\ln(n)) + \ln(\frac{4}{d}))\sqrt{\rho} 2^{c+1}}$.

Using Lemma 3 we have:

$$\begin{aligned}
H(\tilde{E}) &\geq H(X|Y=2) \geq P(X \in S', Y=2)(H(X_{S',2})) \\
&\geq \left(\frac{\ln(n) - \ln(2)}{\ln(n) - \ln(2) + (\ln(\ln(n)) + \ln(\frac{4}{d}))\sqrt{\rho}2^{c+1}}\right) \\
&\quad \left(\log(\frac{\ln(n) - \ln(2)}{\ln(\ln(n)) + \ln(\frac{4}{d})}) \right. \\
&\quad \left. - \left(\frac{\rho 2^c \ln(\rho 2^c)}{\rho 2^c - 1} - 1 - \ln(\frac{\rho 2^c \ln(\rho 2^c)}{\rho 2^c - 1})\right)\frac{1}{\ln(2)}\right) \\
&= \left(1 - \frac{(\ln(\ln(n)) + \ln(\frac{4}{d}))\sqrt{\rho}2^{c+1}}{\ln(n) - \ln(2) + (\ln(\ln(n)) + \ln(\frac{4}{d}))\sqrt{\rho}2^{c+1}}\right) \\
&\left(\log(\frac{\ln(n) - \ln(2)}{\ln(\ln(n)) + \ln(\frac{4}{d})}) \right. \\
&\quad \left. - \left(\frac{\rho 2^c \ln(\rho 2^c)}{\rho 2^c - 1} - 1 - \ln(\frac{\rho 2^c \ln(\rho 2^c)}{\rho 2^c - 1})\right)\frac{1}{\ln(2)}\right). \qquad (2.14)
\end{aligned}$$

Since $c = O(1)$ and $d = \Theta(1)$, this lower bound is asymptotically $H(\tilde{E}) \geq \max_y H(X|Y=y) \geq H(X|Y=2) \geq (1 - o(1))(\log(\log(n)) - \log(\log(\log(n))) - \mathcal{O}(1))$.

Now when $n \geq \nu(r, q, \rho, c, d)$, we can lower bound the $(1 - o(1))$ term as:

$$\begin{aligned}
&1 - \frac{(\ln(\ln(n)) + \ln(\frac{4}{d}))\sqrt{\rho}2^{c+1}}{\ln(n) - \ln(2) + (\ln(\ln(n)) + \ln(\frac{4}{d}))\sqrt{\rho}2^{c+1}} && (2.15) \\
&\geq 1 - \frac{(1+r)\ln(\ln(n))\sqrt{\rho}2^{c+1}}{\ln(n/2) + \ln(\ln(n))\sqrt{\rho}2^{c+1}} && (2.16) \\
&\geq 1 - \frac{(1+r)\sqrt{\ln(n/2)}\sqrt{\rho}2^{c+1}}{\ln(n/2) + \sqrt{\ln(n/2)}\sqrt{\rho}2^{c+1}} && (2.17) \\
&= 1 - \frac{1+r}{1 + \frac{\ln(n/2)}{\sqrt{\rho}2^{c+1}}} && (2.18) \\
&\geq 1 - \frac{1+r}{1+q} && (2.19)
\end{aligned}$$

We bound from (2.15) to (2.16) by using $n \geq e^{\left(\frac{4}{d}\right)^{1/r}}$ which implies $\ln(\ln(n)) + \ln(\frac{4}{d}) \leq (1+r)\ln(\ln(n))$. We go from (2.16) to (2.17) by using $\sqrt{\ln(n/2)} \geq \ln(\ln(n))$ when $n \geq 3$. We bound from (2.18) to (2.19) by using $n \geq 2e^{q^2 2^{2(c+1)}\rho}$. Next, we lower bound the term $\log\left(\frac{\ln(n)-\ln(2)}{\ln(\ln(n))+\ln(\frac{4}{d})}\right)$.

$$\begin{aligned}
&\log\left(\frac{\ln(n)-\ln(2)}{\ln(\ln(n))+\ln(\frac{4}{d})}\right) && (2.20)\\
&\geq \log\left(\frac{\ln(n/2)}{(1+r)\ln(\ln(n))}\right) && (2.21)\\
&\geq \log\left(\sqrt{\ln(n/2)}\right)-\log(1+r) && (2.22)\\
&\geq 0.5\log(0.5\log(n/2))-\log(1+r) && (2.23)\\
&\geq 0.5\log(\log(n))-\log(1+r)-1 && (2.24)
\end{aligned}$$

We bound from (2.20) to (2.21) by using $\ln(\ln(n))+\ln(\frac{4}{d})\leq(1+r)\ln(\ln(n))$. We bound from (2.21) to (2.22) using $\sqrt{\ln(n/2)}\geq\ln(\ln(n))$. We then substitute all of these bounds into our previous lower bound on $H(\tilde{E})$ (2.14) yielding:

$$\begin{aligned}
H(\tilde{E})\geq&\left(1-\frac{1+r}{1+q}\right)\Bigg(0.5\log(\log(n))-\log(1+r)\\
&-\mathcal{O}(1)-\frac{1}{\ln(2)}\bigg(\frac{\rho 2^c\ln(\rho 2^c)}{\rho 2^c-1}-1-\ln\left(\frac{\rho 2^c\ln(\rho 2^c)}{\rho 2^c-1}\right)\bigg)\Bigg)\\
=&\left(1-\frac{1+r}{1+q}\right)(0.5\log(\log(n))-\log(1+r)-\mathcal{O}(1)).
\end{aligned}$$

Proof of Corollary 1

Condition (a): Bounded Ratio. We know that $\frac{\max_x p(x)}{\min_x p(x)}\leq\rho$. Since $\sum_x p(x)=1$, $\min_x p(x)\leq\frac{1}{n}\leq\max_x p(x)$ and we have $\frac{\max_x p(x)}{1/n}\leq\rho\Rightarrow\max_x p(x)\leq\frac{\rho}{n}$ and similarly $\min_x p(x)\geq\frac{1}{\rho n}$. Then using Theorem 1, when $n\geq\nu(r=1,q=3,\rho^2,c,d=1)$, $H(\tilde{E})\geq\max_y H(X|Y=y)\geq 0.25\log(\log(n))-\mathcal{O}(1)$ with high probability (where the $\mathcal{O}(1)$ term is a function of only ρ, c). As such, there exists an n_0 (which is a function of only ρ, c) such that for all $n>n_0$, the causal direction is identifiable with high probability.

Condition (b): Sampled Uniformly on the Simplex. We first show that when the distribution of X is uniformly sampled from the simplex, there exist a set S that satisfies the assumptions of Theorem 1 with high probability.

Lemma 6. *When the x_i are sampled uniformly from the simplex, there exists a subset of the support with size at least $(e^{-\frac{1}{\sqrt{\rho}}} - \frac{1}{\sqrt{\rho}} - \delta)n$ for which all x_i are within a factor of $\sqrt{\rho}$ from $\frac{1}{n}$ and make up total probability mass $\geq \left(e^{-\frac{1}{\sqrt{\rho}}} - \frac{1}{\sqrt{\rho}} - \delta\right)\frac{1}{\sqrt{\rho}}$, with probability $> 1 - 2e^{-2\delta^2 n}$ for $\rho, n \geq 1$, $\delta > 0$.*

Proof. Let us call a probability "small" if $x_i \leq \frac{1}{\sqrt{\rho n}}$. We want to show that with high probability (at least $1 - 2e^{-2\delta^2 n}$), there are at most $(1 - e^{-\frac{1}{\sqrt{\rho}}} + \delta)n$ small x_i. Using Theorem 3 of [39], we know that for each x_i in a Dirichlet distribution with $\alpha = 1$ (i.e., the uniform distribution over the probability simplex) and support size n, $P(x_i > z) = (1 - z)^{n-1}$ (This is by setting $a_i = z$ and $a_j = 0, \forall j \neq i$ and using the fact that $P(x_i = 0) = 0, \forall i \in [n]$). As such, $P(x_i \leq z) = 1 - (1 - z)^{n-1}$. The probability that x_i is small is then equal to $P(x_i \leq \frac{1}{\sqrt{\rho n}}) = 1 - (1 - \frac{1}{\sqrt{\rho n}})^{n-1}$. This value is non-decreasing when $n \geq 1$, and approaches $1 - e^{-\frac{1}{\sqrt{\rho}}}$ as n approaches infinity. Hence when $n \geq 1$, the probability that any x_i is "small" is upper-bounded by $1 - e^{-\frac{1}{\sqrt{\rho}}}$. We want to show that the outcome that there are more than $(1 - e^{-\frac{1}{\sqrt{\rho}}} + \delta)n$ small x_i will not happen with high probability. To do this, we note that all x_i in a symmetric Dirichlet distribution are negatively associated (this follows from Lemma 9 in Section 2.9.1). This implies that the probability that there are at least $(1 - e^{-\frac{1}{\sqrt{\rho}}} + \delta)n$ small x_i is upper-bounded by the probability that there are at least that many x_i when we treat the x_i as if they are i.i.d. random variables. This allows us to use Hoeffding's inequality. Let S_n be the total number of small x_i and E_n be the expected number of small x_i. Since $E_n \leq (1 - e^{-\frac{1}{\sqrt{\rho}}})n$, then $P(S_n > (1 - e^{-\frac{1}{\sqrt{\rho}}} + \delta)n) \leq P(|S_n - E_n| > \delta n) < 2e^{-2\delta^2 n}$. As such, the probability that there are at most $(1 - e^{-\frac{1}{\sqrt{\rho}}} + \delta)n$ small x_i is at least $(1 - 2e^{-2\delta^2 n})$.

Let us call an x_i "big" if $x_i \geq \frac{\sqrt{\rho}}{n}$. There are at most $\frac{n}{\sqrt{\rho}}$ big x_i, since otherwise their total probability mass would exceed 1.

Next, consider the subset of x_i that are neither "big" nor "small". They are in the range $[\frac{1}{\sqrt{\rho n}}, \frac{\sqrt{\rho}}{n}]$. We know that with high probability $(1 - 2e^{-2\delta^2 n})$ there are at most $(1 - e^{-\frac{1}{\sqrt{\rho}}} + \delta)n$ small x_i and at most $\frac{n}{\sqrt{\rho}}$ big x_i. This means our desired subset has size at least $\left(n - (1 - e^{-\frac{1}{\sqrt{\rho}}} + \delta)n - \frac{n}{\sqrt{\rho}}\right) = \left(e^{-\frac{1}{\sqrt{\rho}}} - \frac{1}{\sqrt{\rho}} - \delta\right)n$ with probability at least $1 - 2e^{-2\delta^2 n}$. □

As such, if we set $\rho = 25$ and $\delta = 0.1$, there exists a subset of the support of size $\geq (e^{-\frac{1}{\sqrt{25}}} - \frac{1}{\sqrt{25}} - 0.1)n \geq 0.5n$ where all x_i are within a factor of $\sqrt{25} = 5$ from $\frac{1}{n}$ with probability $> 1 - 2e^{-2(0.1)^2 n} = 1 - 2e^{-0.02n}$. Using Theorem 1, we conclude that when $n \geq \nu(r = 1, q = 3, \rho = 25, c, d = 0.5)$, $H(\tilde{E}) \geq \max_y H(X|Y = y) \geq 0.25 \log(\log(n)) - \mathcal{O}(1)$ with high probability (where the $\mathcal{O}(1)$ term is a function of only c). As such, there exists an n_0 (which is a function of only c) such that for all $n > n_0$, the causal direction is identifiable with high probability.

Condition (c): High Entropy. We show that when X has entropy within an additive constant of $\log(n)$, there exists a set S that satisfies the assumptions of Theorem 1.

Lemma 7. *For any distribution X with support size n and entropy $\geq \log(n) - a$, there exists a subset S with all $x_i \in .[\frac{3}{40n}, \frac{2^{2b}}{n}]$ for $i \in S$, and support size $|S| \geq \frac{n}{2^{2b+3}}$, where $b = \max\{a, 2\}$.*

Proof. Let us call an x_i "large" if $x_i \geq \frac{2^{2b}}{n}$, and μ_{large} be the total probability mass contributed by large x_i. The upper bound for the sum of the terms in the formula for $H(X)$ corresponding to large x_i is $\mu_{\text{large}} \log(\frac{n}{2^{2b}})$. The upper bound for the sum of the terms in Shannon entropy corresponding to x_i that are not large is $(1 - \mu_{\text{large}}) \log(\frac{n}{1-\mu_{\text{large}}})$. Since entropy is greater than $\log(n) - a$ and $b = \max\{a, 2\}$, we have that entropy is greater than or equal to $\log(n) - b$ as well. Then, for the total entropy to be at least $\log(n) - b$ it must be true that $\mu_{\text{large}} \log(\frac{n}{2^{2b}}) + (1 - \mu_{\text{large}}) \log(\frac{n}{1-\mu_{\text{large}}}) \geq \log(n) - b$. It follows that $2b\mu_{\text{large}} + (1 - \mu_{\text{large}}) \log(1 - \mu_{\text{large}}) \leq b$. For $x \geq 0$, we have that $(1-x)\log(1-x) \geq -1.5x$. Then we have $2\mu_{\text{large}}(b - 0.75) \leq b$, or equivalently $\mu_{\text{large}} \leq \frac{b}{2(b-0.75)}$. Since $b \geq 2$, we have that $\mu_{\text{large}} \leq 0.8$.

Let us call an x_i "small" if it is $\leq \frac{0.075}{n}$, and let μ_{small} be the total probability mass in small x_i. Even if all x_i were small (although that would be impossible), $\mu_{\text{small}} \leq 0.075$. As such, $\mu_{\text{small}} + \mu_{\text{large}} \leq \frac{7}{8}$. This means at least $\frac{1}{8}$ total probability mass belongs to $x_i \in [\frac{0.075}{n}, \frac{2^{2b}}{n}]$. Our subset S of X will be all of these x_i. Since every element in X is upper-bounded by $\frac{2^{2b}}{n}$, S has a support size of at least $\frac{\frac{1}{8}}{\frac{2^{2b}}{n}} = \frac{n}{2^{2b+3}}$. □

We can therefore satisfy the conditions of Theorem 1 with $d = \frac{1}{2^{2\max\{a,2\}+3}}$ and

$\rho \leq (\frac{40}{3} 2^{2\max\{a,2\}})^2 \leq 2^{4\max\{a,2\}+8}$. Using Theorem 1, we conclude that when $n \geq \nu(r = 1, q = 3, \rho = 2^{4\max\{a,2\}+8}, c, d = \frac{1}{2^{2\max\{a,2\}+3}})$, $H(\tilde{E}) \geq \max_y H(X|Y = y) \geq 0.25 \log(\log(n)) - \mathcal{O}(1)$ with high probability (where the $\mathcal{O}(1)$ term is a function of only a, c). Hence there exists an n_0 (a function of only a, c) such that for all $n > n_0$, the causal direction is identifiable with high probability.

Proof of Theorem 2

Given the random variables $U_i, i \in [n]$ with marginal distributions $\mathbf{p_i}(u_i)$, let $p(u_1, u_2, \ldots, u_n)$ be a valid coupling. Then p satisfies $\mathbf{p_i}(u_i) = \sum_{k \neq i} \sum_{u_k \in [n]} p(u_1, u_2, \ldots, u_n)$ holds for all i, u_i. Therefore, for all i, u_i, we can define $S_{i,u_i} = \{(u_j)_{j \neq i} : p(u_1, u_2, \ldots u_n) > 0\}$. S_{i,u_i} contains the coordinates in the coupling that contribute non-zero mass to satisfy the i^{th} marginal distribution, specifically the probability that variable U_i takes the value u_i. Let us define the function $g_{i,u_i}((u_j)_{j \neq i}) \coloneqq p(u_1, \ldots, u_n)$. Then equivalently, we can write $\mathbf{p_i}(u_i) = \sum_{t \in S_{i,u_i}} g_{i,u_i}(t)$.

Consider a noisy version of the marginal distributions: Let $\hat{\mathbf{p}}_\mathbf{i}$ be the noisy marginals where $|\hat{\mathbf{p}}_\mathbf{i}(u_i) - \mathbf{p_i}(u_i)| \leq \delta$ for all i, u_i. Our strategy is to start with the coupling $p(u_1, \ldots, u_n)$ and convert it to a coupling for the noisy marginals. Let us define $T_i^+(p) \coloneqq \{u_i : \sum_{k \neq i} \sum_{u_k \in [n]} p(u_1, u_2, \ldots, u_n) < \hat{\mathbf{p}}_\mathbf{i}(u_i)\}, T_i^-(p) \coloneqq \{u_i : \sum_{k \neq i} \sum_{u_k \in [n]} p(u_1, u_2, \ldots, u_n) > \hat{\mathbf{p}}_\mathbf{i}(u_i)\}$. In words, $T_i^+(p)$ shows the coordinates of the i^{th} noisy marginal which has excess mass compared to the mass induced by coupling p. Similarly, $T_i^-(p)$ shows the coordinates of the i^{th} noisy marginal for which the coupling p has more mass than needed. We update p in two stages: First, we update p so that $T_i^-(p) = \emptyset$. In the second stage, we further update p so that $T_i^+(p) = \emptyset$ and $T_i^-(p) = \emptyset$, which shows that the updated p is a valid coupling for the noisy marginals $\hat{\mathbf{p}}_\mathbf{i}$. We finally bound the entropy of the new coupling relative to the initial coupling we started with.

First we observe the following: Consider any $u_i \in T_i^-$. Then there exists a function

Algorithm 1 Phase I

```
Input: Valid coupling $p_{\text{init}}$ for the marginals $\{\mathbf{p_i}\}_{i\in[n]}$. Noisy marginals $\{\hat{\mathbf{p}}_\mathbf{i}\}$
$p \leftarrow p_{\text{init}}$.
Construct $g_{i,u_i}, S_{i,u_i}, T_i^+, T_i^-$ from $p_{\text{init}}$ for all $i, u_i$.
while $\exists i \in [n]$ s.t. $T_i^- \neq \emptyset$ do
    Pick arbitrary $h_{i,u_i}$ for all $u_i$ such that
        $0 \leq h_{i,u_i}(t) \leq g_{i,u_i}(t), \forall t \in S_{i,u_i},$
        $\sum_{t\in S_{i,u_i}} h_{i,u_i}(t) = \hat{\mathbf{p}}_\mathbf{i}(u_i).$
    Update $p$ as follows:
        $p(u_1, u_2, \ldots, u_n) \leftarrow h_{i,u_i}((u_j)_{j\neq i}), \forall (u_j)_{j\neq i} \in S_{i,u_i}$    (2.27)
    Construct $g_{i,u_i}, S_{i,u_i}, T_i^+, T_i^-$ from $p$ for all $i, u_i$.
end while
return $p$
```

$h_{i,u_i}(t)$ such that

$$0 \leq h_{i,u_i}(t) \leq g_{i,u_i}(t), \forall t \in S_{i,u_i}, \tag{2.25}$$

$$\sum_{t\in S_{i,u_i}} h_{i,u_i}(t) = \hat{\mathbf{p}}_\mathbf{i}(u_i). \tag{2.26}$$

This is true since $\sum_{t\in S_{i,u_i}} g_{i,u_i}(t) = \mathbf{p_i}(u_i)$ and $\hat{\mathbf{p}}_\mathbf{i}(u_i) < \mathbf{p_i}(u_i), \forall u_i \in T_i$. We can describe the first phase as follows: For each $i \in [n]$ and $u_i \in T_i^-$, we pick an arbitrary h_{i,u_i} and update p to match the entries of h_{i,u_i}. Notice that each update of p changes the corresponding h, g functions. Our construction proceeds by updating these functions every time p is updated as given above. This procedure is summarized in Algorithm 1.

Note that the size of T_i^- after an update is at least one less than the size of T_i^- before the update. To see this, note that after the update in (2.27), $u_i \notin T_i^-$. Also by reducing elements of p, we can never add a new element to T_i^- for any i by definition of T_i^-. Therefore, after at most $\sum_i |T_i^-|$ applications of the above update for the initial sets T_i^-, we have $T_i^- = \emptyset, \forall i \in [n]$. Since there are at most n elements in T_i^- and n

such sets, the first phase terminates in at most n^2 steps.

Let p be the output of Algorithm 1 in the rest of the proof. In the second phase, we consider the updated T_i^+. Our strategy here is to distribute the remaining mass in each marginal as its own coupling and add this coupling to p that is the output of Algorithm 1. Let us represent the excess probability mass in coordinate u_i of marginal i relative to coupling p by r_{i,u_i}. Note that $r_{i,u_i}(p) \coloneqq \hat{\mathbf{p}}_\mathbf{i}(u_i) - \sum_{k\neq i}\sum_{u_k\in[n]} p(u_1, u_2, \ldots, u_n)$ may increase at each step of the first phase. The exact increase in this gap for each i, u_i depends on the choice of h_{i,u_i} function at each step. However, we can bound the total gap per marginal at the end of first phase as $\sum_{u_i\in[n]} r_{i,u_i}(p) \leq \delta n^2, \forall i$. Each step of Algorithm 1 can add a mass of at most δ to each marginal at each step (it terminates after at most $\sum_i |T_i^-|$ steps) and at the beginning of first phase, each coordinate of each marginal has at most δ excess mass (there are $\sum_i |T_i^+|$ coordinates with excess mass). As such, there is at most $\sum_i \delta|T_i^-| + \sum_i \delta|T_i^+| \leq \delta n^2$ total gap per marginal at the end of the first phase. Let $p(u_1, \ldots, u_n)$ be the output of Algorithm 1. [30] showed a greedy minimum entropy coupling algorithm that produces a coupling with support at most n^2. Let $q(u_1, u_2 \ldots, u_n)$ be the output of this greedy algorithm when given the excess marginal mass as its input. Then we have that $v \coloneqq p + q$ is a valid coupling for the noisy marginals. This is because, by feeding the greedy algorithm the excess marginal mass, we guarantee that the marginals of v are correct. Moreover, all cells in the coupling are in range $[0, 1]$ as no cell in p or q has negative value and their sum has the correct marginals.

Next, define the distribution $s : 2 \times [n]^n \to [0, 1]$ as follows:

$$s(0, u_1, u_2, \ldots, u_n) = p(u_1, u_2, \ldots, u_n), \tag{2.28}$$

$$s(1, u_1, u_2, \ldots, u_n) = q(u_1, u_2 \ldots, u_n). \tag{2.29}$$

From the argument above, it is easy to see that s is a valid probability distribution, i.e., it has non-negative entries and its entries sum to 1.

We compare entropy of the obtained coupling v with entropy of s and that with entropy of the initial coupling p_{init}. First, it is easy to see from concavity of entropy and Jensen's inequality that $H(v) \leq H(s)$. Let $\bar{H}$ be the extended entropy operator that admits vectors outside the simplex as input, for vectors whose entries are between 0 and 1: $\bar{H}(p(x)) = -\sum_x p(x)\log(p(x))$. We have the following lemma that allows us to compare $\bar{H}(p)$ with $H(p_{\text{init}})$:

Lemma 8. *Let $\mathbf{p} = [p_1, p_2, \ldots, p_n]$ be a discrete probability distribution. Let $\mathbf{q} = [q_1, q_2, \ldots, q_n]$ be a non-negative vector such that $q_i \leq p_i, \forall i \in [n]$. Then $\bar{H}(\mathbf{q}) \leq \bar{H}(\mathbf{p}) + \frac{\log(e)}{e}$.*

The proof is in Section 2.9.1 in the supplement.

From the lemma, we can conclude that $\bar{H}(p) \leq H(p_{\text{init}}) + \frac{\log(e)}{e}$. Finally, the maximum entropy contribution of q is when it induces uniform distribution over n^2 states. Since the total mass of q is δn^2, we have

$$\bar{H}(q) \leq n^2 \left(\frac{\delta n^2}{n^2} \log \left(\frac{n^2}{\delta n^2} \right) \right) \tag{2.30}$$

$$= \delta n^2 \log \left(\frac{1}{\delta} \right) \tag{2.31}$$

Suppose $\delta \leq \frac{1}{n^2 \log(n)}$. Then we can further bound $\bar{H}(q) \leq 2 + \frac{\log(\log(n))}{\log(n)} \leq 2 + o(1)$ since $\delta \log\left(\frac{1}{\delta}\right) \leq \frac{2\log(n)+\log(\log(n))}{n^2 \log(n)}$ if $\delta < \frac{1}{n^2 \log(n)}$.

Bringing it all together, we obtain the following chain of inequalities:

$$H(v) \leq H(s) = \bar{H}(p) + \bar{H}(q) \tag{2.32}$$

$$\leq H(p_{\text{init}}) + \frac{\log(e)}{e} + 2 + o(1). \tag{2.33}$$

This concludes the proof. □

Proof of Lemma 8

If $p_i < \frac{1}{\exp(1)}, \forall i$, due to monotonicity of $-p\log(p)$ in p, we have $\bar{H}(\mathbf{q}) \leq \bar{H}(\mathbf{p})$.

In general, no more than 2 states can satisfy $p_i > \frac{1}{\exp(1)}$. Therefore, $\bar{H}(q)$ can only be larger than $\bar{H}(p)$ due to two states. Let us call these two states p_1, p_2 without loss of generality. Reducing the probability of any other state only gives a looser bound.

We can obtain the largest entropy increase by solving the following optimization problem:

$$\begin{aligned}
\max_{p_1,p_2} \quad & \mathbb{1}_{\{p_1>1/e\}}\left(\frac{\log(e)}{e} - p_1\log\left(\frac{1}{p_1}\right)\right) \\
& + \mathbb{1}_{\{p_2>1/e\}}\left(\frac{\log(e)}{e} - p_2\log\left(\frac{1}{p_2}\right)\right) \qquad (2.34) \\
\text{subject to} \quad & p_1+p_2 \leq 1, \\
& p_1 \geq 0, p_2 \geq 0
\end{aligned}$$

Suppose $p_1 > 1/e$ and $p_2 < 1/e$. Then the solution is simply to set $p_1 = 1$ since this minimizes the entropy contribution of p_1. This gives a gap of $\frac{\log(e)}{e}$. Due to symmetry, we only need to investigate the case where $p_1 > 1/e$ and $p_2 > 1/e$. In this case, we have the following optimization problem:

$$\begin{aligned}
\min_{p_1,p_2} \quad & p_1\log\left(\frac{1}{p_1}\right) + p_2\log\left(\frac{1}{p_2}\right) \\
\text{subject to} \quad & p_1+p_2 \leq 1, \qquad (2.35) \\
& p_1 \geq 1/e, p_2 \geq 1/e
\end{aligned}$$

This is a concave minimization problem and the solution has to be at the boundary of the convex constraint region. If $p_1 = 1/e$, the maximum gap is obtained when p_2 is maximized to $p_2 = 1 - 1/e$ which gives a gap that is strictly less than $\frac{\log(e)}{e}$, hence we can discard this solution for the maximum entropy gap. $p_2 = 1/e$ gives the same solution from symmetry. When $p_1 + p_2 = 1$, the problem reduces to minimizing the binary entropy function, which again is minimized at the boundary. The boundary in this case is where either $p_1 = 1/e$ or $p_2 = 1/e$. Therefore, both probabilities being greater than $1/e$ cannot yield a better bound. □

Proof of Lemma 1

Joint Probabilities. First, we bound the estimates of the entries of the joint distribution between X and Y. Both X and Y have n states which we index as $i = 1, \ldots, n$ and $j = 1, \ldots, n$ respectively. Hence the joint distribution has n^2 states. Probability that $X = i$ and $Y = j$ is shown as p_{ij}. Suppose N samples from N

independent, identically distributed random variables are drawn as $\{(x_k, y_k)\}_{k\in[N]}$. This yields the empirical probability estimates (I is the indicator function)

$$\hat{p}_{ij} = \frac{1}{N}\sum_{k=1}^{N} I(x_k = i \,\&\, y_k = j).$$

Note that each of these estimates are averages of Bernoulli random variables with success probability p_{ij}. We also consider the marginal probability empirical estimates

$$\hat{p}_i^X = \frac{1}{N}\sum_{k=1}^{N} I(x_k = i)$$

and

$$\hat{p}_j^Y = \frac{1}{N}\sum_{k=1}^{N} I(y_k = j).$$

which are also averages of N Bernoulli random variables (with success probabilities p_i^X and p_j^Y respectively).

Since these estimates are clearly correlated with one another, our approach will be to use concentration results on individual entries of the joint distribution and then do a union bound over all $n^2 + 2n$ probabilities. Note that $I(x_k = i \,\&\, y_k = j) = 1$ with probability p_{ij} and 0 otherwise. Thus by Hoeffding's inequality [65],

$$\mathbb{P}\left\{|\hat{p}_{ij} - p_{ij}| \geq t\right\} \leq 2\exp\left(-2t^2N\right). \tag{2.36}$$

We can define an event $\mathcal{A}$ where all the probability estimates are within t of the truth:

$$\mathcal{A} = \left\{\max_{i,j\in 1,\ldots,n} |\hat{p}_{ij} - p_{ij}| \leq t\right\} \bigcap \left\{\max_{i\in 1,\ldots,n} |\hat{p}_i^X - p_i^X| \leq t\right\} \bigcap \left\{\max_{j\in 1,\ldots,n} |\hat{p}_j^Y - p_j^Y| \leq t\right\}.$$

Starting with (2.36) and taking the union bound over all $n^2 + 2n$ probabilities in the joint and marginal distribution, we obtain

$$\begin{aligned}\mathbb{P}(\mathcal{A}) &> 1 - 2(n^2 + 2n)\exp\left(-2t^2N\right) \qquad (2.37)\\ &> 1 - 4\exp(2\ln(n) - 2t^2N).\end{aligned}$$

Conditional Probabilities. Given the above bound on the estimates of the joint probabilities, we formulate bounds on the conditional probability estimates. Recall that

$$P(X = i|Y = j) = \frac{P(X = i, Y = j)}{P(Y = j)} = \frac{p_{ij}}{\sum_{i=1}^{n} p_{ij}}.$$

Using the plug-in approach, we have

$$\hat{p}_{i|j} = \frac{\hat{p}_{ij}}{\hat{p}_j^Y}.$$

Note that it is critical for $\hat{p}_j^Y$ to be bounded away from zero, otherwise a small error in $\hat{p}_{ij}$ may cause a large error in $\hat{p}_{i|j}$. In what follows, we set

$$\alpha = \frac{\min_{j=1,\dots,n} p_j^Y}{2}.$$

α will naturally appear in the number of samples, and notably must depend on n. Note that the case of $\sum_{i=1}^{n} p_{ij} = 0$ is allowable since if that is the case $Y = j$ will never occur and corresponding probability estimates will all be zero and the conditional probabilities will not be of interest.

Now consider any $t < \alpha$, assume that event $\mathcal{A}$ holds. We then have that all $\hat{p}_j^Y > p_j^Y - t > 2\alpha - t > \alpha$. Combined with the fact that under event $\mathcal{A}$, $|\hat{p}_{ij} - p_{ij}| < t$ and $t \geq 0$, it is easy to check that

$$\begin{aligned}
\hat{p}_{i|j} - p_{i|j} &= \frac{\hat{p}_{ij}}{\hat{p}_j^Y} - \frac{p_{ij}}{p_j^Y} \\
&< \frac{p_{ij} + t}{p_j^Y - t} - \frac{p_{ij}}{p_j^Y} \\
&= \frac{p_{ij}p_j^Y + tp_j^Y - p_{ij}p_j^Y + tp_{ij}}{p_j^Y(p_j^Y - t)} \\
&< \frac{tp_j^Y + tp_{ij}}{p_j^Y \alpha} \\
&< \frac{2t}{\alpha},
\end{aligned}$$

where the last inequality follows since $p_{ij} < p_j^Y$ by definition. Similarly,

$$\begin{aligned} p_{i|j} - \hat{p}_{i|j} &= \frac{p_{ij}}{p_j^Y} - \frac{\hat{p}_{ij}}{\hat{p}_j^Y} \\ &< \frac{p_{ij}}{p_j^Y} - \frac{p_{ij} - t}{p_j^Y + t} \\ &= \frac{p_{ij}p_j^Y + tp_{ij} - p_{ij}p_j^Y + tp_j^Y}{p_j^Y(p_j^Y + t)} \\ &< \frac{tp_j^Y + tp_{ij}}{p_j^Y 2\alpha} \\ &< \frac{t}{\alpha}, \end{aligned}$$

hence

$$|\hat{p}_{i|j} - p_{i|j}| < \frac{2t}{\alpha}.$$

Since by (2.37) the event $\mathcal{A}$ holds with probability at least $1-4\exp(2\log(n)-2t^2N)$, we have

$$\mathbb{P}\left(\max_{i,j\in 1,\ldots,n} |\hat{p}_{i|j} - p_{i|j}| \geq \frac{2t}{\alpha}\right) \leq 4\exp(2\ln(n) - 2t^2N). \tag{2.38}$$

The derivation of the bound for the conditional probability estimates in the other direction is similar and relies on the same event $\mathcal{A}$ holding. Hence the probability the bounds hold in both directions simultaneously remains $\mathbb{P}(\mathcal{A})$.

Achieving error of $\delta = 1/(n^2\ln(n))$. Let $\alpha = \frac{\min\{\min_x p(x), \min_y p(y)\}}{2}$. Suppose we want $2t/\alpha = 1/(n^2\ln(n))$. Then we need $t = 1/(2n^2\alpha^{-1}\ln(n))$. Note that $t < \alpha$ as required above. Suppose further that we want this to hold with probability at least $1 - 4/n$. By the above, we require

$$\begin{aligned} 2\ln(n) - 2t^2N &< -\ln(n) \\ 3\ln(n) &< \frac{2N}{4n^4\alpha^{-2}\ln^2(n)} \\ 6n^4\alpha^{-2}\ln^3(n) &< N \end{aligned}$$

Hence N needs to be $\Omega(n^4\alpha^{-2}\ln^3(n))$. □

Proof of Theorem 3

From the equivalence between the minimum entropy coupling problem and the problem of finding the exogenous variable with minimum entropy, the output of $\mathcal{A}(\{\hat{p}(Y|X = x)\}_x)$ is the smallest entropy of any exogenous variable for the causal model $X \to Y$. Similarly, this claim holds for $\mathcal{A}(\{\hat{p}(X|Y = y)\}_y)$ as well. From Theorem 1, entropy in the direction $Y \to X$ scales with n using $p(X|Y = y)$. From Theorem 6 of [18], it can be seen that the given sampling error can induce an entropy difference of at most $o(1)$ in the conditional entropies. Hence, even with noisy conditionals, $\max_y \hat{H}(X|Y = y)$ scales with n, implying that $\mathcal{A}(\{\hat{p}(X|Y = y)\}_y)$ scales with n. In the forward direction, the true exogenous variable provides a valid coupling under the true joint distribution without sampling noise. From Lemma 2, given N samples, there exists a valid coupling in the forward direction that is constant entropy away from the true exogenous variable. Hence $\mathcal{A}(\{p(Y|X = x)\}_x)$ is constant. Since $\mathcal{A}(\{p(X|Y = y)\}_y)$ scales with n, the result follows.

Proof of Theorem 4

We first show that the $H(X|Y = 2)$ conditional entropy will have enough samples to be included in the criterion listed in Theorem 4. As $N = \Omega(n^2 \log(n))$, we have at least $c_1 n^2 \log(n)$ samples for $c_1 = \Theta(1)$. As shown in the proof of Theorem 1, $p(Y = 2) = \Omega(\frac{1}{n}) \geq \frac{c_4}{n}$ where $c_4 = \Theta(1)$. Following a rejection sampling approach, we use Hoeffding's inequality to show that if $c_1 n^2 \log n$ samples are drawn from the joint distribution, then with probability $1 - o(1)$ we will successfully draw $\Omega(n \log(n))$ independent samples from the distribution $p(X|Y = 2)$. Specifically, let S_n denote the number of samples (out of $c_1 n^2 \log(n)$ total samples from the joint distribution) for which $Y = 2$, and $E_n = \mathbb{E}[S_n]$ denote the expected number of such samples. We have $E_n \geq (c_1 n^2 \log(n))(\frac{c_4}{n}) = c_1 c_4 n \log(n)$. Hence using Hoeffding's inequality, $P\left(S_n < \frac{c_1 c_4 n \log(n)}{2}\right) \leq P\left(|S_n - E_n| > \frac{c_1 c_4 n \log(n)}{2}\right) < 2e^{-\frac{2(c_1 c_4 n \log(n))^2}{c_1 n^2 \log(n)}} = 2e^{-2c_1 c_4^2 \log(n)} = o(1)$. Hence $S_n \geq \frac{c_1 c_4 n \log(n)}{2} \gg n$ with probability $1 - o(1)$. Thus the $\hat{H}(X|Y = 2)$, which we use for identifiability, will have sufficient number of samples to be included in the criterion in Theorem 4.

We now show that each conditional entropy in the criterion in Theorem 4 will have error bounded by a constant with high probability. Immediately following from Corollary 1.12 of [64], for a distribution D with support size n, $|H(D) - \hat{H}(D)| \leq 1$ with probability $1 - e^{-n^{c_2}}$ given a sample of size at least $\frac{c_3 n}{\log(n)}$ where $c_2, c_3 = \Theta(1)$. Since we only calculate conditional entropy estimates with $\geq n$ samples, the number of samples $n \gg \frac{c_3 n}{\log(n)}$ for all considered conditional entropies. Hence the total probability of any computed conditional entropy estimate being off by more than 1 is $\leq ne^{-n^{c_2}} = o(1)$ by the union bound. Since by the proof of Theorem 1 we know $\max_x H(X|Y = y) \leq c \ll \Omega(\log(\log(n))) \leq H(X|Y = 2)$, it immediately follows that $\max_{x, \hat{p}(X=x)N \geq n} \hat{H}(Y|X = x) \leq c + 1 \ll \Omega(\log(\log(n))) - 1 \leq \max_{y, \hat{p}(Y=y)N \geq n} \hat{H}(X|Y = y)$. □

Proof of Corollary 3

This generative model satisfies the assumptions of Theorem 3 following from the proof of Corollary 1. As such, under this generative model for sufficiently large n and $N = \Omega(n^4 \alpha^{-2} \log^3(n))$ samples, $\mathcal{A}(\{\hat{p}(X|Y = y)\}_y) > \mathcal{A}(\{\hat{p}(Y|X = x)\}_x)$ with high probability.

We show a lower bound on α with high probability, under this generative model. As mentioned in the proof of Corollary 1, under this generative model, for any i, $P(x_i \leq z) = 1 - (1 - z)^{n-1}$. We aim to show that with high probability, $x_i \geq \frac{1}{n^2 \log(n)}, \forall i \in [n]$ when n is sufficiently large.

We lower bound the probability of this not happening as $(1 - (1 - \frac{1}{n^2 \log(n)})^{n-1})n$ by the union bound. Note that $\lim_{n \to \infty} \frac{(1-(1-\frac{1}{n^2 \log(n)})^{n-1})n}{1/\log(n)} = 1$.

Hence for sufficiently large n the probability that there exists an $x_i < \frac{1}{n^2 \log(n)}$ is upper bounded by $\frac{2}{\log(n)}$. Thus, we have a high probability lower bound for α. We substitute this for α in our lower bound for the number of required samples in the previous paragraph. This yields that under this generative model for sufficiently large n and $N = \Omega(n^8 \log^5(n))$ samples, $\mathcal{A}(\{\hat{p}(X|Y = y)\}_y) > \mathcal{A}(\{\hat{p}(Y|X = x)\}_x)$ with high probability.

Proof of Negative Association

Lemma 9. *Let $[x_i]_{i\in[n]}$ be a vector, uniformly randomly sampled from the probability simplex in n dimensions. Then $[x_i]_{i\in[n]}$ is negatively associated.*

Proof. Let $x_i = \frac{z_i}{\sum_j z_j}$, where each z_i is independent and identically distributed exponential random variable with mean 1, i.e. distributed as Exp(1). Then $[x_i]_i$ is a discrete probability distribution uniformly randomly chosen from the simplex in n dimensions. We will show that x_i are negatively associated. The following argument is provided by [53] as an answer on the online forum `https://mathoverflow.net/`, which we reproduce here for completeness.

Consider the following theorem:

Theorem 5. *[27] Let $z_1, z_2, \dots, z_n$ be n random variables with log-concave probability densities. Then $(z_1, z_2, \dots, z_n)$ conditioned on $\sum_{i\in[n]} z_i$ are negatively associated.*

Note that exponential distribution is log-concave. Hence the theorem is applicable in our setting. Furthermore, the distribution induced on $(\frac{z_i}{\sum_{j\in[n]} z_j})_{i\in[n]}$ is identical to the distribution induced on $(z_1, z_2, \dots, z_n)$ conditioned on $\sum_{i\in[n]} z_i = 1$. This concludes the proof. □

2.9.2 Additional Experiments and Experimental Details

Experimental Details

In this section, we provide the complete details of every experiment given in the main text, as well as provide additional results that we were not able to present in the main text due to space constraints.

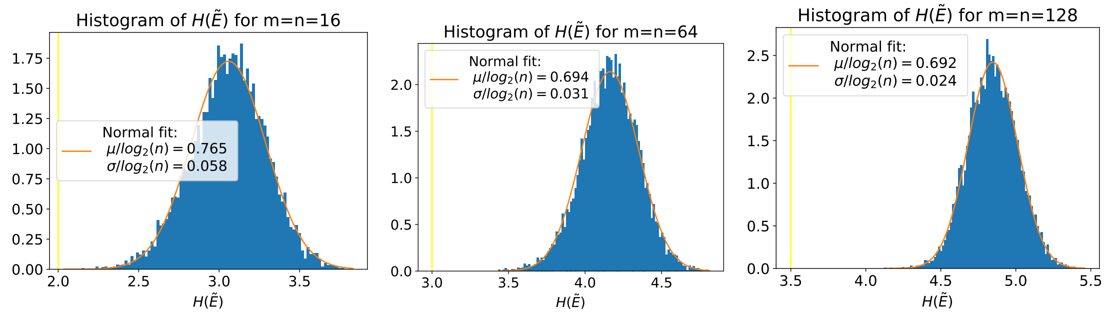
Sampling low-entropy exogenous variables: We use Dirichlet distribution to sample the distribution for the exogenous variable from the probability simplex. Dirichlet has the parameter α which affects the entropy of the distribution obtained by sampling the corresponding Dirichlet distribution: Smaller α values lead to sampling distributions with smaller entropy. Suppose we want to sample distributions for E such that $H(E) \leq \theta$. Since a good α value for this θ is not known a priori, we use the following adaptive sampling scheme: Suppose we want to sample N distributions for E such that $H(E) \leq \theta$. We initialize with $\alpha^{(0)} = 1$ and obtain $10N$ samples from Dirichlet with parameters $\alpha^{(0)}$. If there are at least N samples out of $10N$ which has entropy less than θ, we are done. If not, we set $\alpha^{(1)} = 0.5\alpha^{(0)}$ and iterate until for a particular $\alpha^{(i)}$ such that at least N out of $10N$ samples satisfy the entropy condition.

Details about Figure 2-2: We set E to have mn number of states where m, n are the number of states of X and Y, respectively. It can be shown that this many number of states is sufficient to obtain any joint distribution. We uniformly randomly sample the function f in the structural equation $Y = f(X, E)$. We also independently and uniformly randomly sample $p(X)$ from the simplex, i.e., we obtain samples from Dirichlet distribution with parameter $\alpha = 1$. For $m = n = 40$, we choose 20 values of θ, i.e., entropy thresholds for the exogenous variable E, uniformly spaced in the range $[0, \log(m)]$. For $m \neq n$, we choose 10 θ values in the range $[0, \log(\max\{m, n\})]$.

When $m \neq n$, we use a mixture data as follows: We obtain 10000 samples from the graph $X \rightarrow Y$ and we obtain 10000 samples from $X \leftarrow Y$. We operate on this mixed data. This is done to reflect the fact that, there is no reason for the cause or the effect variable to have less or more number of states. Accuracy shown in the figures reflect the fraction of times each algorithm correctly identifies the true causal direction. Total entropy-based compares $H(X) + H(E)$ and $H(Y) + H(\tilde{E})$ where E and $\tilde{E}$ are the outputs of the greedy minimum entropy coupling algorithm in the direction $X \rightarrow Y$ and $X \leftarrow Y$, respectively.

Details about Figure 2-3: We sample exogenous variable using the above adaptive sampling method so that, for each value of n, we have $H(E) \leq 0.8\log(n)$. The other details are identical (e.g., 10000 samples for each configuration.) Due to the sampling method, we observe that most of the samples are very close to $H(E) \approx 0.8\log(n)$. We then obtain the histogram plots for $H(\tilde{E})$, where $\tilde{E}$ is the output of the greedy minimum entropy coupling algorithm in the wrong direction. As observed, data fits well to a Gaussian and is highly concentrated around $0.854\log(n)$.

Details about Figure 2-5: In this section, we introduce a latent confounder L. First, distribution of L and distribution of E are sampled independently. Then the



(a) (b) (c)

Figure 2-6: Histogram of $H(\tilde{E})$ when $H(E) \approx 0.5 \log_2(n)$. Yellow line shows $x = 0.5 \log_2(n)$

distributions $p(X|l), p(Y|x, l, e)$ are sampled uniformly randomly from the simplex for every configuration of x, l, e. We use the adaptive sampling described above to sample E such that $H(E) \leq 2$. Using the same sampling method, we sweep through different entropy thresholds for the latent confounder L and sample such that $H(L) \leq \phi$ for $\phi \in \{0.5, 1, 1.5, 2, 2.5, 3\}$. The settings for m, n and how data is mixed is identical to the procedure used to obtain Figure 2-2: When $m \neq n$, we use uniformly mixed data from $X \rightarrow Y$ and $X \leftarrow Y$. For each configuration, we obtain 1000 total number of samples and report the accuracy of the method to identify the true causal direction.

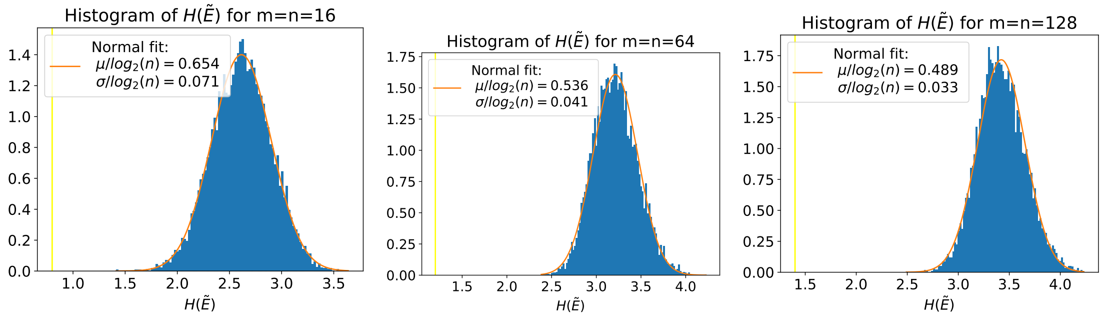
Relaxing constant exogenous entropy assumption

As indicated in Section 2.6, we provide additional experiments for $\alpha = 0.2$ and 0.5 in Figure 2-7 and Figure 2-6, respectively. As can be seen, for both α values, i.e., when $H(E) \leq \alpha \log(n)$, $H(\tilde{E})$ highly concentrates around $\beta \log(n)$ for some $\beta > \alpha$.

Additional results on the finite sample regime

Figure 2-8 shows results on finite sample identifiability for the setting considered in the figure in the main text, except with smaller $H(E) \leq \ln(4)$.

Results for $p(X)$ drawn from Dir(1) are shown Figure 2-9, as described in the main text. We find that the greedy MEC performance degrades to a level that is similar to the conditional entropy criterion. This might be explained by the fact that

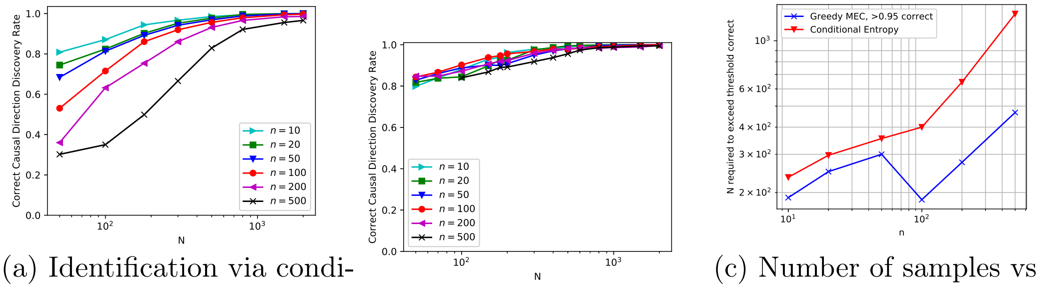


(a)

(b)

(c)

Figure 2-7: Histogram of $H(\tilde{E})$ when $H(E) \approx 0.2\log_2(n)$. Yellow line shows $x = 0.2\log_2(n)$



(a) Identification via conditional entropies ($H(E) \leq \ln(4)$).

(b) Identification via MEC algorithm ($H(E) \leq \ln(4)$).

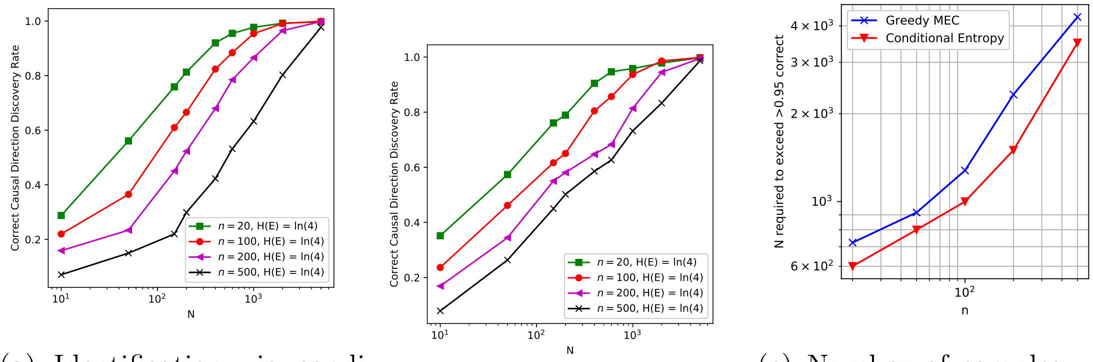
(c) Number of samples vs. support size of observed variables.

Figure 2-8: Finite sample identifiability of the causal direction via entropic causality. (a) Probability of correctly discovering the causal direction $X \to Y$ as a function of n and number of samples N, using the conditional entropies as the test. (b) Probability of correctly discovering the causal direction $X \to Y$ using the greedy MEC algorithm to test the direction. (c) Samples N required to reach 95% correct detection as a function of n, derived from the plots in Figure 2-8a and Figure 2-8b.

if $p(X|Y = y)$ are close to uniform, then the gap between $H(\tilde{E})$ and $H(X|Y = y)$ vanishes.

Additional Tuebingen Experiments

In this section, we perform additional experiments to evaluate the stability of the method to choice of quantization on the Tuebingen dataset. Specifically, to quantize $[a, b]$ into n intervals, we perturb each quantization point $\{a + \frac{(b-a)i}{n}\}_i$ with a uniform noise in $[-\frac{(b-a)}{8n}, \frac{(b-a)}{8n}]$. For every pair, this is done 5 times independently and the majority decision is taken. The results, which show similar performance to Table 2.1



(a) Identification via conditional entropies ($H(E) = \ln(4)$).

(b) Identification via MEC algorithm ($H(E) = \ln(4)$).

(c) Number of samples vs. support size of observed variables ($H(E) = \ln(4)$).

Figure 2-9: Finite sample identifiability of the causal direction via entropic causality, where $p(x) \sim \text{Dir}(1)$ (uniform on the simplex). (a) Probability of correctly discovering the causal direction $X \to Y$ as a function of n and number of samples N, using the conditional entropies as the test. (b) Probability of correctly discovering the causal direction $X \to Y$ as a function of n and number of samples N, using the greedy MEC algorithm to test the direction. (c) Samples N required to reach 98% correct detection as a function of n, derived from the plots in Figure 2-9a and Figure 2-9b.

	Threshold ($\times$ log support)	0.6	0.7	0.8	0.85	0.9	1.0	1.2
5-state quantization	# of pairs	10	13	32	42	53	69	85
	Accuracy (%)	90.0	61.5	53.1	54.8	56.5	58.5	57.6
	Threshold ($\times$ log support)	0.6	0.7	0.8	0.85	0.9	1.0	1.2
10-state quantization	# of pairs	8	12	23	39	49	71	85
	Accuracy (%)	87.5	66.7	60.9	53.8	51.0	52.1	57.6
	Threshold ($\times$ log support)	0.6	0.7	0.8	0.85	0.9	1.0	1.2
20-state quantization	# of pairs	5	10	15	31	54	78	85
	Accuracy (%)	60.0	70.0	73.3	54.8	48.1	48.7	55.3

Table 2.3: Performance on Tübingen causal pairs with low exogenous entropy in at least one direction. Chosen based on majority voting on 5 random quantizations.

are shown in Table 2.3, demonstrating a degree of stability to choice of quantization. We observe that perturbed quantization demonstrates better performance for 20−state quantization, whereas it shows somewhat worse performance for the 5 and 10−state quantizations. This indicates that more research is needed to determine the optimal quantization for a given dataset.

Chapter 3

Entropic Causal Inference: Graph Identifiability

3.1 Overview

In this chapter, we detail joint work with Kristjan Greenewald, Dmitriy Katz, and Murat Kocaoglu.

Entropic causal inference is a recent framework for learning the causal graph between two variables from observational data by finding the information-theoretically simplest structural explanation of the data, i.e., the model with smallest entropy. In our work, we first extend the causal graph identifiability result in the two-variable setting under relaxed assumptions. We then show the first identifiability result using the entropic approach for learning causal graphs with more than two nodes. Our approach utilizes the property that ancestrality between a source node and its descendants can be determined using the bivariate entropic tests. We provide a sound sequential peeling algorithm for general graphs that relies on this property. We also propose a heuristic algorithm for small graphs that shows strong empirical performance. We rigorously evaluate the performance of our algorithms on synthetic data generated from a variety of models, observing improvement over prior work. Finally we test our algorithms on real-world datasets.

3.2 Introduction

Causal reasoning is essential for high-quality decision-making, as, for instance, it improves interpretability and enables counterfactual reasoning [2, 44, 45]. By learning the relationships between causes and effects, we can predict how various interventions would affect a system. Advances in causality enable us to better answer questions such as *"Why does this phenomenon occur in the system?"* or *"What could happen if the system were perturbed in this particular way?"* Moreover, causal inference methods are being utilized to tackle key challenges for reliability of ML systems, such as domain adaptation [38, 69] and generalization (e.g. via causal transportability or imputation) [3, 49, 61].

Structural causal models (SCMs) represent relationships in a system of random variables [47]. In particular, each variable is modeled with a structural equation that characterizes how the variable is realized. Causal graphs are directed acyclic graphs (DAGs) that are used to represent such systems, where nodes and edges correspond to variables and the causal relations between these variables, respectively. A variable's structural equation is a function of the variable's corresponding node's parents in the graph.

Learning such causal graphs can be done through a series of interventions. However, in many settings it is not possible to perform such interventions. A large amount of literature has focused on learning the causal graph from observational data with additional "faithfulness assumptions" [60], though in general it is impossible to fully learn the causal graph without stronger assumptions on the generative model. A variety of stronger assumptions and corresponding methodologies exist in the literature [19, 37, 51, 59]. Most of these methods, however, are limited to continuous variables and thus cannot handle categorical data, especially in the multivariate setting.

A recent framework explicitly designed to handle categorical data is *entropic causal inference* [12, 29]. At a high level, the underlying assumption of this approach is that true causal mechanisms in nature are often "simple," taking inspiration from the

Occam's razor principle. The authors adopt an information-theoretic realization of this principle by using "entropy" to measure the complexity of a causal model. As we further explore in this work, entropic causal inference provides a means to measure the amount of randomness a generative model would require to produce an observed distribution. As Occam's razor prefers simpler explanations, entropic causal inference prefers generative models with small randomness. We do not expect following this preference to always lead to the discovery of true causal relationships (just as one does not expect a simpler explanation to be always be the correct one), but view this as a guiding intuition that mirrors nature and experimental observations. Our experiments on semi-synthetic data demonstrates that the low-entropy assumption indeed holds in certain settings.

Previously, the framework was applied to discovering the causal direction between two random variables given that the amount of randomness in the true causal relationship is small. We focus on extending this framework to learn larger causal graphs instead of just cause-effect pairs. Suppose the observed variables have n states. Our contributions follow:

1. We show pairwise identifiability with strictly relaxed assumptions compared to the previously known results. We enable learning the causal graph $X \rightarrow Y$ from observational data even when (i) the cause variable X has low entropy of $o(\log(n))$ and (ii) the exogenous noise has non-constant entropy, i.e., $\mathcal{O}(1) \ll H(E) = o(\log\log(n))$.

2. We show the first identifiability result for causal graphs with more than two observed variables, with a new peeling algorithm for general graphs.

3. We propose a heuristic algorithm that searches over all DAGs and outputs the one that requires the minimum entropy to fit to the observed distribution.

4. We experimentally evaluate our algorithms and show that entropic approaches outperform the discrete additive noise models in synthetic data. We also apply our algorithms on semi-synthetic data using the *bnlearn*[1] repository and

[1] `https://www.bnlearn.com/bnrepository/`

demonstrate the applicability of low-entropy assumptions and the proposed method.

3.3 Related Work

Learning causal graphs from observational data has been studied extensively in the case of continuous variables. [37] proposes an algorithm for learning linear structural causal models when the error variance is known. Similarly, [51] show that linear models with Gaussian noise become identifiable if the noise variance is the same for all variables. A more general modeling assumption is the additive noise model (ANM). In [59], the authors show that for almost all linear causal models, the causal graph is identifiable if the additive exogenous noise is non-Gaussian.

In the case of discrete and/or categorical variables, causal discovery literature is much more sparse. [5] introduces a method for categorical cause-effect pairs when there exists a hidden intermediate representation that is compact. In [14, 25], the authors propose using an information-geometric approach called IGCI that is based on independence of cause and the causal mechanism. However IGCI can provably recover the causal direction only in the case of deterministic relations. An extension of additive noise models to discrete data is done in [52] where identifiability is shown between two variables. The authors also propose using the regression-based algorithm of [42] (which made continuous domain ANM applicable to arbitrary graphs) for the discrete setting as well. Without specific assumptions on the graph and the generative mechanisms, this is a heuristic algorithm, i.e., identifiability in polynomial time is not guaranteed by discrete ANM on graphs with more than two nodes.

One related idea is to use Kolmogorov complexity to determine the simplest causal model [24]. Minimum-description length has been used as a substitute for Kolmogorov complexity (which is not computable) in a series of follow-up papers [4, 41]. Our extension of entropic causal inference to graphs can be seen as an information-theoretic realization of this promise, where the complexity of the causal model is captured by its entropy. Other information-theoretic concepts such as interaction information [17] and directed information [15] have also been studied in the context of causality.

3.4 Background and Notation

Causal Graphs and Learning: Consider a causal system where each variable is generated as a function of a subset of the rest of the observed variables and some additional randomness. Such systems are modeled by structural equations and are called structural causal models (SCMs). Let $X_1, X_2, \ldots, X_{|V|}$ be the set of observed variables. Accordingly, there exists functions f_i and exogenous noise terms E_i such that $X_i = f_i(\mathrm{Pa}_i, E_i)$. This equation in a causal system should be understood as an assignment operator since changing Pa_i affects X_i whereas changing X_i does not affect Pa_i. We say the set of variables Pa_i *cause* X_i. A directed acyclic graph (DAG) can be used to summarize these causal relations, which is called the *causal graph*. We denote the causal graph by $G = (V, \mathcal{E})$ where V is the set of observed nodes and $\mathcal{E}$ is the set of directed edges. There are $|V|$ nodes, $X_1, X_2, \ldots, X_{|V|}$, where each X_i corresponds to an observable random variable. Edges are constructed by adding a directed arrow from every node in the set Pa_i to X_i for all i. Pa_i then becomes the set of parents of X_i in G. We assume causal sufficiency, i.e., that there are no unobserved confounders, and that there is no selection bias. Under these assumptions, $\mathrm{Pa}_i \perp\!\!\!\perp E_i$. Additionally, for simplicity of presentation we denote the number of states of all variables as n (i.e. $|X_i| = n$ for all i). Note that our proofs do not require each observed variable to strictly have the same number of states; we merely need them scale together, i.e. if $X_1 \in [n_1]$ and $X_2 \in [n_2]$ then $\frac{n_1}{n_2} = \Theta(1)$. All big-o notation in the paper is relative to n. Our goal is to infer the directed causal graph from the observed joint distribution $p(X_1, X_2, \ldots, X_{|V|})$ using the assumptions of the entropic causality framework as needed.

Even without making any parametric assumptions, we can learn some properties of the graph from purely observational data. Algorithms relying on conditional independence tests (such as the PC or IC algorithms [47, 60]) can identify the *Markov equivalence class* (MEC) of G, i.e. the set of graphs that produce distributions with the exact same set of conditional independence relations. Moreover, a graph's Markov equivalence class uniquely determines its *skeleton* (the set of edges, ignoring orientation) and *unshielded colliders* (the induced subgraphs of the form $X \rightarrow Z \leftarrow Y$). A Markov equivalence class is summarized by a mixed graph called the *essential graph*, which has the same skeleton and contains a directed edge if all graphs in the equivalence class orient the edge in the same direction. All other edges are undirected. The problem of determining the true causal graph from observational data thus reduces to orienting these remaining undirected edges, given enough samples to perform conditional independence tests reliably.

Entropic Causality Framework: Without interventional data, one needs additional assumptions to refine the graph structure further than the equivalence class. The key assumption of the entropic causality framework is that, in nature, true causal models are often "simple." In information-theoretic terms, this is formalized as the entropy of exogenous variables often being small. Previous work has shown guarantees for identifying the direction between a causal pair X, Y where $Y = f(X, E), X \perp\!\!\!\perp E$ for some exogenous variable E from observational data. The work of [29] showed that when the support size of the exogenous variable (i.e. the Renyi-0 entropy $H_0(E) = |E|$) is small, with probability 1 it is impossible to factor the model in the reverse direction (as $X = g(Y, \tilde{E})$) with an exogenous variable with small support size (i.e. $|\tilde{E}|$ must be large). Thus, one can identify the causal pair direction by fitting the smallest cardinality exogenous variable in both directions and checking which direction enables the smaller cardinality. [29] conjectured that this approach also would work well for Shannon entropy. [12] resolved this conjecture, showing identifiability for causal pairs under particular generative assumptions.

Definition 2 ((α, β)-support)**.** *A discrete random variable X is said to have (α, β)-support if at least α states of X have probability of at least β.*

[12] assumes that the cause variable X has $(\Omega(n), \Omega(\frac{1}{n}))$-support and that the Shannon entropy of the exogenous variable (i.e. $H(E) = H_1(E)$) is small. Specifically, they showed that when $H(E) = O(1)$, $H(\tilde{E}) = \Omega(\log(\log(n)))$ with high probability.

The high probability statement is with respect to the selection of the function f, i.e., for all but a vanishing (in n) fraction of functions f, identifiability holds. Moreover, they showed that this approach was robust to only having a polynomial number of samples, whereas the result of [29] that assumed small $|E|$ required knowing the exact joint distribution, e.g. from an oracle or infinite samples.

Algorithmically, one can provably orient causal pairs under the assumptions of [12] by comparing the minimum entropy exogenous variable needed to factor the pair in both directions (i.e. comparing the minimum $H(E)$ for which there exists a function f and $E \perp\!\!\!\perp X$ such that $Y = f(X, E)$, and the analogous quantity minimizing $H(\tilde{E})$). Finding this minimum entropy exogenous variable is an optimization problem equivalent to the *minimum-entropy coupling problem* for the conditionals, specifically, the minimum $H(E)$ in the direction $X \to Y$ is the same as the minimum-entropy coupling for $[(Y|X = i)], \forall i \in [n]$ [8, 29, 46]. Accordingly, we denote the entropy of the minimum-entropy coupling for a variable X conditioned on a set S as $\mathrm{MEC}(X|S)$. [12] showed $\mathrm{MEC}(Y|X) < \mathrm{MEC}(X|Y)$ with high probability.

3.5 Tightening the Entropic Identifiability Result for Cause-Effect Pairs

In this work, we leverage results for the bivariate entropic causality setting to learn general graphs. Theorem 1 of [12] provides identifiability guarantees in the bivariate setting. However, the assumptions of their theorem are not general enough to imply an identifiability result on graphs with more than 2 nodes. Specifically, a fundamental challenge in applying bivariate causality to discover each edge in a larger graph is confounding due to the other variables, i.e., when one considers a pair of variables, the remaining variables act as confounders. These confounders cannot be controlled for since we do not know the causal graph and conditioning on other variables unknowingly creates additional dependencies. One natural approach to handle confounding is to recursively discover source nodes by conditioning on the common causes that are discovered so far in the graph. This idea will form the basis for our peeling algorithm to be proposed in Section 3.6.1. We are interested in learning graphs where the exogenous variable for every node has small entropy (in particular, $H(E_i) = o(\log(\log(n))))$. When conditioning on the source nodes, some nodes X (e.g. the children of the source nodes) will thus have conditional entropies of order $H(X|\text{sources}) = o(\log(\log(n)))$ since for the children of source nodes, the only remaining randomness on X will be due to the low-entropy exogenous variable. This creates problems when attempting to orient edges connected to these variables conditioned on the source nodes. Specifically, Theorem 1 of [12] requires the cause variable X to have $(\Omega(n), \Omega(\frac{1}{n}))$-support which enforces $H(X) = \Omega(\log(n))$ – and this is not satisfied for the above nodes with $o(\log(\log(n)))$ entropy.

In the following bivariate result, we instead only require $(\Omega(n), \Omega(\frac{1}{n\log(n)}))$-support, and simultaneously relax the exogenous variable constraint from $H(E) = O(1)$ to $H(E) = o(\log(\log(n)))$. This condition can be satisfied for X with $H(X) = O(1)$ as needed.

Theorem 6. *Consider the SCM $Y = f(X, E), X \perp\!\!\!\perp E$, where $X, Y \in [n], E \in [m]$. Suppose E is any random variable with entropy $H(E) = o(\log(\log(n)))$. Let X have $(\Omega(n), \Omega(\frac{1}{n\log(n)}))$-support. Let f be sampled uniformly randomly from all mappings $f : [n] \times [m] \to [n]$. Suppose n is sufficiently large. Then, with high probability, any $\tilde{E}$ that satisfies $X = g(Y, \tilde{E}), \tilde{E} \perp\!\!\!\perp Y$ for some g, entails $H(\tilde{E}) \geq \Omega(\log(\log(n)))$.*

While interesting in its own right, we apply this tightened bivariate identifiability result to the general graph case in Section 3.6. Note that the assumption of a uniformly random f (also used in [12], [29]) is not meant as a description of how nature generates causal functions, but as the least-restrictive option for putting a measure on the space of possible functions so that high-probability statements can be made rigorously. Theorem 6 can be immediately adapted to any alternative distribution on the space of f that does not assign any individual value of f probability mass more than $n^{c'}$ times the probability mass assigned by the uniform distribution, for some constant c'.

Proof overview for Theorem 6. Here we provide the intuition behind the proof strategy, the full proof is given in Appendix 3.9.1. It is simple to show that the minimum entropy required to fit the function in the incorrect direction, $H(\tilde{E})$, is lower-bounded as $H(\tilde{E}) \geq \max_y H(X|Y = y)$. The overarching goal of our proof method is then to show that there exists a state y' of Y such that $H(X|Y = y') = \Omega(\log(\log(n)))$.

To accomplish this, we start by showing that the $(\Omega(n), \Omega(\frac{1}{n\log(n)}))$-support of X implies existence of a subset S of $\Omega(n)$ states of X that each have probability $\Omega(\frac{1}{n\log(n)})$ and are all relatively close in probability to each other. We call this subset S, the *plateau* states. If one envisions them as adjacent in the PMF of X, these states would have similar heights and thus look like a plateau.

Now, we conceptualize the realization of f as a balls-and-bins game, where each element of $X \times E$ (a ball) is mapped i.i.d. uniformly randomly to a state of Y (a bin). Using balls-and-bins arguments, it is our hope to show that there is a bin that receives $\Omega(\frac{\log(n)}{\log(\log(n))})$ *plateau balls* of the form $(X \in S, E = e_1)$, where e_1 is the most probable state of E, and that this will cause the corresponding conditional distribution to have large entropy. The primary intuition is that a bin receiving many plateau balls would cause the corresponding conditional distribution to have many plateau states that all have near-uniform probabilities, and this near-uniform subset of the conditional distribution would contribute a significant fraction of the probability mass to guarantee that its entropy is large. With the stronger assumptions on (α, β)-support by [12], this proof method suffices. However, as we are assuming a weaker notion of (α, β)-support, it is not clear that the plateau balls would make up a significant fraction of the conditional's mass to guarantee large entropy.

In a sense, the plateau balls are probability masses that are "helping" us make some conditional entropy large. The proof of [12] takes the perspective that all remaining mass from non-plateau states are "hurting" our effort to make a conditional distribution with large entropy. To accommodate our relaxed assumptions, we take a more nuanced perspective on *helpful* and *hurtful* mass. Consider a non-plateau state x of X that contributes a small amount of mass towards the conditional distribution corresponding to a state y of Y. With the perspective of [12], this would be viewed as hurtful mass because it is from a non-plateau state of X. But intuitively, in terms of its contribution to $H(X|Y = y)$, it does not matter whether x is a plateau state or not. Through careful analysis, we can show that if $P(X = x|Y = y)$ is small then they are not "too hurtful." We follow this intuition to make a new definition of the good mass, where we set a threshold $\mathcal{T}$, define the first $\mathcal{T}$ mass we receive from a non-plateau state of X as *helpful* mass for the state of Y, and the surplus beyond $\mathcal{T}$ from the non-plateau state of X as *hurtful* mass for the state of Y. As before, all mass from plateau states will be helpful. With this new perspective and a careful analysis, we show that there is a state y' that receives many plateau balls, and has much more helpful mass than hurtful mass. This then enables us to show that $H(X|Y = y')$ is large, proving the theorem.

3.6 Learning Graphs via Entropic Causality

Now, we focus on how to leverage the capability of correctly orienting causal pairs to learn causal graphs *exactly*. In comparison, traditional structure learning methods only learn the Markov equivalence class of graphs from observational data. For example, given the line graph $X \to Y \to Z$, such methods would deduce the true graph is either $X \to Y \to Z$ or $X \leftarrow Y \leftarrow Z$, but not that it is exactly $X \to Y \to Z$.

As was discussed in Section 3.4, learning the entire graph can be reduced to correctly orienting each edge in the skeleton. However, we cannot naively use a pairwise algorithm, as the rest of the observed variables can act as confounders. We examine how different pairwise oracles can enable us to characterize the value of using minimum entropy couplings to learn causal graphs. One example of a natural-feeling oracle is one that can correctly orient any edges that have no active confounding. Such an oracle enables learning of directed trees and complete graphs. However, it cannot be used to learn all general graphs (see Section 3.9.1 for an example). We propose an alternative oracle, that can distinguish between a source node and any node it can reach:

Definition 3 (Source-pathwise oracle)**.** *A source-pathwise oracle for a DAG G always* *orients $A \rightarrow B$ if A is a source and there exists a directed path from A to B in G.*

Let us formalize our entropic method for causal pairs as the following oracle:

Definition 4 (MEC oracle)**.** *A minimum entropy coupling (MEC) oracle returns $X \rightarrow Y$ if MEC($Y|X$) $<$ MEC($X|Y$) and $X \leftarrow Y$ otherwise, given the joint distribution $p(X,Y)$.*

We aim to show that our MEC oracle is a source-pathwise oracle for graphs with the following assumptions:

Assumption 2 (Low-entropy assumption)**.** *Consider an SCM where $X_i = f_i(Pa_i, E_i)$, $Pa_i \perp\!\!\!\perp E_i, \forall i$, where $X_i \in [n], E_i \in [m]$. Suppose $|V| = O(1)$, $H(E_i) = o(\log(\log(n)))$ and E_i has $(\Omega(n), \Omega(\frac{1}{n\log(n)}))$-support for all i, and f_i are sampled uniformly randomly from all mappings $f_i : [n] \times [n]^{|Pa_i|} \rightarrow [n]$.*

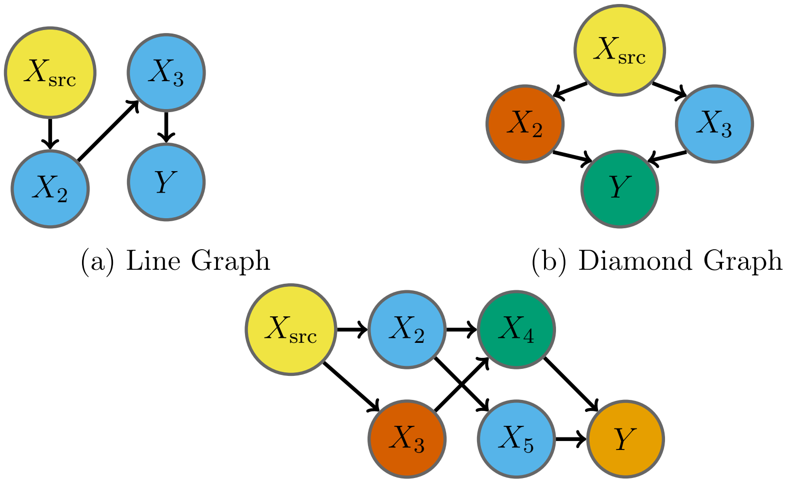
We are now ready to show the main result of our paper. We show that, under certain generative model assumptions, applying entropic causality on pairs of observed variables acts as a source-pathwise oracle for DAGs:

Theorem 7. *For any SCM under Assumption 2, the MEC oracle is a source-pathwise oracle for the causal graph with high probability for sufficiently large n.*

Characterizing entropic causality as a source-pathwise oracle enables us to identify the true causal graph for general graphs. We outline the key intuitions of our proof:

Proof overview for Theorem 7. Suppose X_{src} is a source and Y is a node such that there is a path from X_{src} to Y. To show the MEC oracle is a source-pathwise oracle, we show that $\text{MEC}(X_{\text{src}}|Y) > \text{MEC}(Y|X_{\text{src}})$. As in Theorem 6, we will accomplish this by showing there is a state y' of Y such that $H(X_{\text{src}}|Y = y') = \Omega(\log(\log(n)))$.

We begin by conceptualizing the realization of all f_i as a balls-and-bins game. Every node X_i is a uniformly random function f_i of $\text{Pa}(X_i) \cup E_i$. Let us define each ball as the concatenation of X and all E_i other than E_{src}. More formally, we denote each ball as $(X = x, E_1 = e_1, \ldots, E_{\text{src}-1} = e_{\text{src}-1}, E_{\text{src}+1} = e_{\text{src}+1}, \ldots, E_{|V|} = e_{|V|})$,



(a) Line Graph

(b) Diamond Graph

(c) Hall Graph

Figure 3-1: Graphs colored according to the Random Function Graph Decomposition (Definition 5) used in the proof of Theorem 7.

and each ball has a corresponding probability mass of $P(X = x) \times \Pi_{i \neq \text{src}} P(E_i = e_i)$. To view the realization of f_i as a balls-and-bins game, we consider the nodes in an arbitrary topological order for the graph. When we process a node X_i, we group balls according to their configuration of $(\text{Pa}(X_i) \cup E_i)$. This is because balls with the same configuration correspond to the same cell of the function f_i. For each group of balls that all share the same configuration, we uniformly randomly sample a state of X_i to assign all the balls in the group. This is essentially realizing one cell of f_i. Groups are assigned independently of other groups. In this sense, each realization of f_i is a balls-and-bins game where we group balls by their configuration, and throw them together into states of X_i (bins).

Let us define plateau balls as those who have a plateau state of X_{src} and have the most probable state of E_i for every $i \neq \text{src}$. As was done in Theorem 6, our goal is to show that there will be a state y' of Y such that y' receives many plateau balls and much more helpful mass than hurtful mass. However, it is not immediately clear how to show this in the graph setting. For intuition, we explore two special cases.

Consider the case of a line graph (Figure 3-1a). For simplicity of this proof overview, assume all X_i other than X_{src} are deterministic functions of their parents (i.e., $H(E_i) = 0$). Using techniques similar to Theorem 6, we can show there are many bins of X_2 that receive many plateau balls and much more helpful mass than hurtful mass. Moreover, we can then use similar techniques to show a non-negligible proportion of those bins will have their corresponding balls mapped together to a bin of X_3 where it does not encounter much hurtful mass. We can repeat this argument again to show some of these desirable bins "survive" from X_3 to Y. While only a scalingly small fraction of these desirable bins "survive" each level, this still ensures the survival of at least one bin if the number of vertices is constant. This will accomplish our goal of having a state y' of Y with many plateau balls and much more helpful mass than hurtful mass.

On the other hand, consider the case of a diamond graph in Figure 3-1b. Again, assume for simplicity that all X_i other than X_{src} are deterministic functions of their parents. Note that when we realize f_Y, two balls are always mapped independently unless they share the same configuration of $\text{Pa}(Y) = \{X_2, X_3\}$. We observe that X_2 and X_3 are both independent deterministic functions of X_{src}. Accordingly, the probability of two particular states $x, x' \in X_{\text{src}}$ satisfying $f_2(x) = f_2(x')$ and $f_3(x) = f_3(x')$ is equal to $\frac{1}{n^2}$. Therefore, almost all pairs of balls are mapped to Y from X_{src} independently. Since everything is almost-independently mapped to Y, we can treat it like a bivariate problem and use techniques similar to Theorem 6.

We are able to prove correctness for both of these graphs, but we do so in ways that are essentially opposite. For the line graph, we utilize strong dependence as bins with desired properties "survive" throughout the graph. For the diamond graph, we utilize strong independence as balls are all mapped to Y essentially independently. To combine the intuitions of these two cases into a more general proof, we introduce the Random Function Graph Decomposition:

Definition 5 (Random Function Graph Decomposition). *Given a DAG and a pair of nodes (X_{src}, Y), Random Function Graph Decomposition colors the nodes iteratively following any topological order of the nodes as follows:*

1. *Color the node with a new color if X_{src} is a parent of the node or if the node has parents of different colors.*

Algorithm 2 Learning general graphs with oracle

```
$\mathcal{R} \leftarrow \{1, \dots, |V|\}$ {set of remaining nodes}
$\mathcal{I} \leftarrow \emptyset$ {set of pairs found to be conditionally independent}
$\mathcal{T} \leftarrow [\ ]$ {list of nodes in topological order}
while $|\mathcal{R}| > 0$ do
    $\mathcal{N} \leftarrow \emptyset$ {set of nodes discovered as non-sources}
    $\mathcal{C} \leftarrow \{1, \dots, |V|\} \backslash \mathcal{R}$ {condition on previous sources}
    for all $(X_i, X_j) \in \{\mathcal{R} \times \mathcal{R}\}$ do
        if $X_i \notin \mathcal{N}$ and $X_j \notin \mathcal{N}$ and $(X_i, X_j) \notin \mathcal{I}$ then
            if $\mathrm{CI}(X_i, X_j | \mathcal{C})$ then
                $\mathcal{I} \leftarrow \mathcal{I} \cup (X_i, X_j)$
            else if Oracle$(X_i, X_j | \mathcal{C})$ orients $X_i \to X_j$ then
                $\mathcal{N} \leftarrow \mathcal{N} \cup \{X_j\}$ {$X_j$ is not a source}
            else
                $\mathcal{N} \leftarrow \mathcal{N} \cup \{X_i\}$ {$X_i$ is not a source}
            end if
        end if
    end for
    $\mathcal{S} \leftarrow \mathcal{R} \backslash \mathcal{N}$ {the remaining nodes that are a source}
    $\mathcal{R} \leftarrow \mathcal{R} \backslash \mathcal{S}$ {remove sources from remaining nodes}
    for all $X_i \in \mathcal{S}$ do
        append $X_i$ to $\mathcal{T}$
    end for
end while{Now, $\mathcal{T}$ is a valid topological ordering}
for all $(i, j) \in \{1, \dots, |V|\}^2$ where $i < j$ do
    if $\mathrm{CI}(\mathcal{T}(i), \mathcal{T}(j) | \{\mathcal{T}(1), \dots, \mathcal{T}(j-1)\} \backslash \mathcal{T}(i))$ then
        no edge between $\mathcal{T}(i)$ and $\mathcal{T}(j)$
    else
        orient $\mathcal{T}(i) \to \mathcal{T}(j)$
    end if
end for
```

2. *Color the node with the color of its parents if all of the node's parents have the same color.*

Using the Random Function Graph Decomposition, we claim that when a node is assigned a new color as in step 1, we utilize independence as in the diamond graph (Figure 3-1b), and when a node inherits its color as in step 2 we utilize dependence as in the line graph in Figure 3-1a. We illustrate Figure 3-1c as an example. With a careful analysis, we utilize these intuitions to prove the MEC oracle is a source-pathwise oracle with high probability.

3.6.1 Peeling Algorithm for Learning Graphs

In the previous section, we have shown how entropic causality can be used as a source-pathwise oracle. Next, we show how to learn general graphs with a source-pathwise oracle. Our algorithm will iteratively determine the graph's sources, condition on the discovered sources, determine the graph's sources after conditioning, and so on. Doing this will enable us to find a valid lexicographical ordering of the graph. Given a lexicographical ordering, we can learn the skeleton with $O(n^2)$ conditional independence tests.

Now, we outline how we iteratively find the sources. In each stage, we consider all the remaining nodes as candidate sources. It is our goal to remove all non-sources from our set of candidates. To do this, we iterate over all pairs of candidates and do a conditional independence test *conditioned on the sources that are found so far*. If the pair is conditionally independent, we do nothing. We note that this will never happen for a pair where one node is a true source and the other node is reachable from the source through a directed path: Conditioning on previously found sources cannot d-separate such paths. Otherwise, the pair is conditionally dependent. We then use the source-pathwise oracle to orient between the two nodes, and eliminate the sink node of the orientation as a candidate (i.e., if we orient $A \to B$, we eliminate B as a candidate source).

Suppose two nodes are dependent conditioned on the past sources. Then either the pair contains a source node and a descendant of the source node, or it contains two non-source nodes. In the former case when the pair contains a source node the source-pathwise oracle will always orient correctly and the non-source node will be eliminated. In the latter case when the pair are two non-sources we can safely eliminate either as a source candidate and accordingly oracle output is irrelevant. By the end of this elimination process, we can show that only true sources will remain as candidates in each step, which enables us to obtain a valid lexicographical ordering, and thus learn the causal graph. We summarize this procedure as Algorithm 2. The following theorem shows the correctness of Algorithm 2 given a source-pathwise oracle:

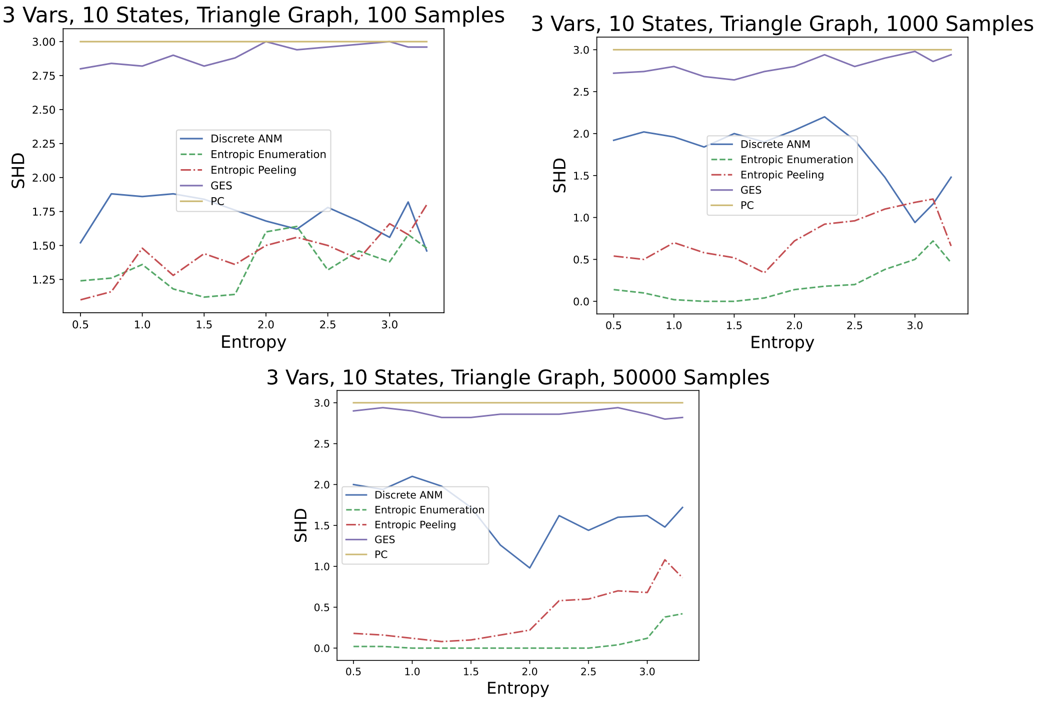


Figure 3-2: Performance of methods in the unconstrained setting in the triangle graph $X \rightarrow Y \rightarrow Z, X \rightarrow Z$: 50 datasets are sampled for each configuration from the unconstrained model $X = f(\text{Pa}_X, E_X)$. The $x-$axis shows entropy of the exogenous noise. The exogenous noise of the first variable is fixed to be large (≈ 3.3 bits), hence it is a high entropy source (HES). Entropic methods consistently outperform the ANM algorithm in almost all regimes.

Theorem 8. *Algorithm 2 learns any causal graph $D = (V, E)$ with $\mathcal{O}(|V|^2)$ calls to a source-pathwise oracle and $\mathcal{O}(|V|^2)$ conditional independence tests.*

Finally, we show that we can use entropic causality together with Algorithm 2 for learning general causal graphs:

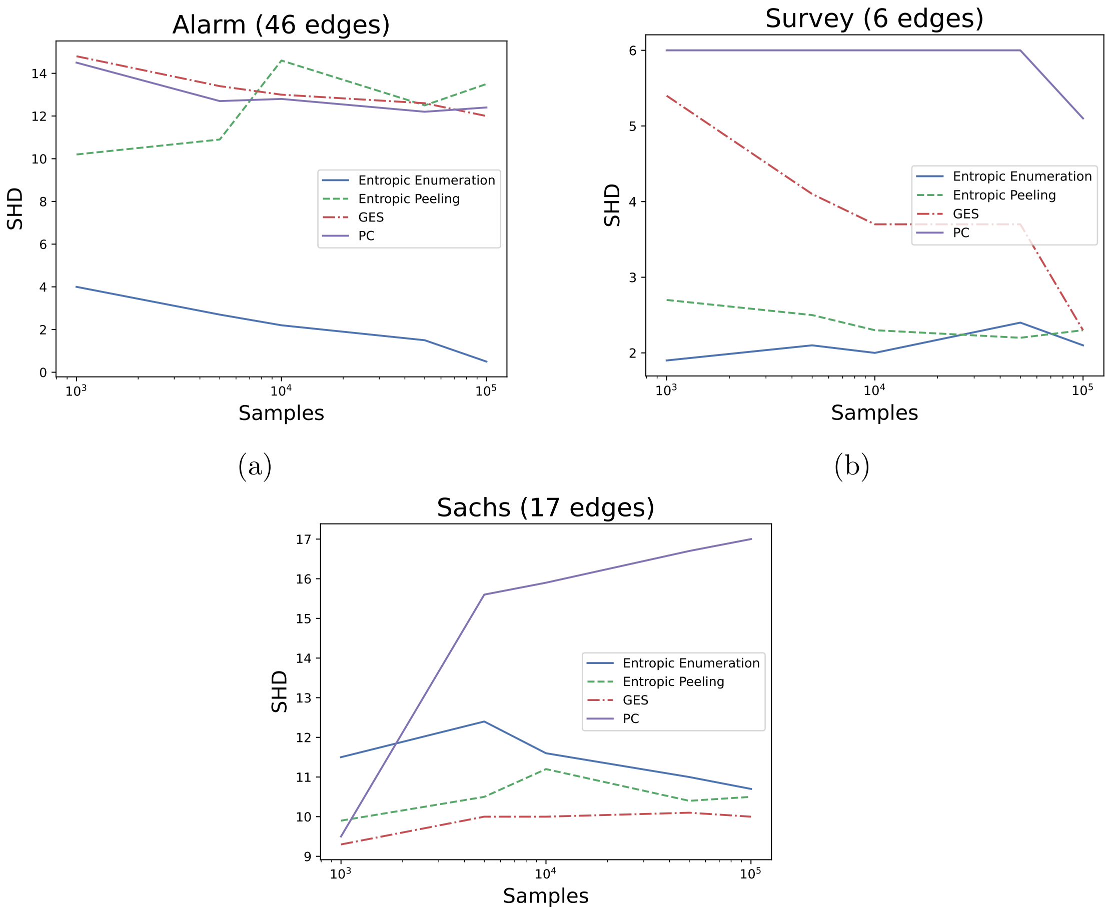
Corollary 4. *For any SCM under Assumption 2, using entropic causality for pairwise comparisons in Algorithm 2 learns, with high probability, the causal graph that is implied by the SCM.*

3.7 Experiments

We first introduce a heuristic that we call the *entropic enumeration* algorithm. In this algorithm, we enumerate over all possible causal graphs consistent with the skeleton and calculate the minimum entropy needed to generate the observed distribution from the graph with independent noise at each node. The minimum entropy needed to generate the joint distribution with some graph D is $\sum_{X_i} \mathrm{MEC}(X_i|\mathrm{Pa}_D(X_i))$ where $\mathrm{Pa}_D(X_i)$ denotes the parents of X_i in D. The graph requiring the least randomness is then selected.

We are not aware of any provably correct method for causal discovery between categorical variables that are non-deterministically related. For discrete variables, the only such method other than entropic causality is the discrete additive noise model [52]. We compare entropic causality to discrete ANM for learning causal graphs, using the graph extension of ANM proposed by [42]. To isolate the role of our algorithms in identifying causal graphs beyond the equivalence class, we support every algorithm in our comparisons with the skeleton of the true graph (obtainable from conditional independence tests given enough data). We evaluate performance via the structural Hamming distance (SHD) from the estimated graph to the true causal graph. See the Appendix for implementation details.

Performance on Synthetic Data. Figure 3-2 compares the performance of entropic peeling, entropic enumeration and discrete ANM algorithms for the triangle graph, i.e., the graph with edges $X \to Y$, $Y \to Z$, and $X \to Z$. Every datapoint is obtained by averaging the SHD to the graph for 50 instances of structural models. To ensure that the entropy of the exogenous nodes are close to the value on the x-axis, their distributions are sampled from a Dirichlet distribution with a parameter that is obtained through a binary search. We observe that the entropic methods consistently outperform the ANM approach. Importantly, we observe how entropic methods are able to near-perfectly learn the *exact* triangle graph in almost all regimes, even though all triangle graphs are in the same Markov equivalence class and thus traditional structure learning algorithms like PC or GES cannot learn anything. With enough



(a)

(b)

(c)

Figure 3-3: Performance of methods on networks from the *bnlearn* repository with varying samples: 10 datasets are sampled for each configuration from the *bnlearn* network. E.g., entropic enumeration exactly recovers Alarm, no algorithm correctly learns half of Sachs.

samples, entropic enumeration learns the graph near-perfectly until the exogenous noise nears $\log(n)$, exceeding our theoretical guarantee of $o(\log(\log(n)))$. In Figure 3-2, we fix the source node to have high entropy. Our motivation is that if all nodes have essentially zero randomness, then we expect the performance of any method to degrade as there is no randomness in samples to observe causality or faithfulness. In Figure 3-9 in Appendix, we do not fix a high-entropy source and still observe that entropic methods outperform ANM in almost all regimes. Experiments with different and larger graphs can be seen in the Appendix.

Performance in Discrete Additive Noise Regime. In this section, we compare the performance of the entropic algorithms and discrete ANM *when the true SCM is a discrete additive noise model.* Using the discrete ANM generative model, we observe that entropic enumeration out-performs the discrete ANM method with few samples and matches its performance with many samples. This demonstrates that even though entropic methods are designed for the general unconstrained SCM class, they perform similarly to ANM which was designed specifically for this setting. Please see Figure 3-8 in Appendix for the results.

Effect of Finite Samples. We observe that entropic methods, particularly enumeration, work well even in regimes with low samples. Experiments focusing on the impact of finite samples can be found in the Appendix.

Performance on Real-World Data. Due to the computational cost of discrete ANM, we compared entropic causality against GES and PC algorithms to evaluate how well it learned real-world causal graphs from the *bnlearn* repository beyond their equivalence class. Figure 3-3 shows performance on three of the six networks we evaluated (see Appendix for remaining networks). Of particular interest is Figure 3-3a, where entropic enumeration almost perfectly identifies a graph with 46 edges from its skeleton and finite samples. Again, we do not claim that the assumptions of entropic causality are universally true in nature, but instead that there are real settings such as Figure 3-3a where the framework enables us to learn causal graphs. Our experiments, exceeding our best theoretical guarantees, show that even when the number of nodes is the same as the number of states, entropic causality can be used for learning the causal graph with a moderate number of samples.

3.8 Conclusion

In this work, we have extended the entropic causality framework to graphs. An identifiability result was proven, and two algorithms were presented and experimentally evaluated — a theoretically-motivated sequential peeling algorithm and a heuristic entropic enumeration algorithm that performs better on small graphs. Overall, we observed strong experimental results in settings much more general than the assumptions used in our theory, indicating that a much stronger theoretical analysis might be possible. We note however that the quantity $H(E_i) = \Theta(\log(\log(n)))$ appears to be approximately a phase transition for the balls-and-bins setting, and posit that the development of novel tools may be required for such an extension of the theory.

We suggest such an advancement may involve an increased focus on a total entropy criterion (i.e., an extension of comparing $H(X) + H(E)$ to $H(Y) + H(\tilde{E})$ in the bivariate case), as in our proposed algorithm of entropic enumeration. Experiments indicate that this performs well, and one might argue that it appears to be more conceptually justified. For one, it mirrors Occam's razor in that it prefers the causal graph with minimal total randomness required. While we do not claim that this methodology will always discover the true generative model (as Occam's razor does not require the simplest explanation to *always* be true), we believe these intuitions mirror nature more often than not, as confirmed by our experimental results. Moreover, such an approach appears to fare better with counter-examples for exogenous-based criterion such as the traveling ball scenario of [22] discussed in [12]. Showing theoretical guarantees for this approach's performance is of interest in future work, and can be framed more generally as, *"Under what conditions is the true generative model the most information-theoretically efficient way to produce a distribution?"*

3.9 Supplementary Material

3.9.1 Proofs

Proof of Theorem 6

Proof Outline.

Following the approach described in the proof overview in the main text (with the descriptions of helpful and hurful mass), we first introduce the *surplus* of a state of Y to characterize the amount of hurtful mass it receives:

Recall that S is the set of “plateau states" of X, i.e., those whose probabilities are close to one another.

Definition 6 (Surplus)**.** *We define the surplus of a state y of Y as* $z_y = \sum_{j \notin S} \max(0, P(X = j, Y = i) - \mathcal{T})$.

Intuitively, only values from states of X outside the plateau states which exceed the threshold will significantly “hurt" conditional entropy $H(X|Y = y)$. We will show there is a state y' of Y where $z_{y'}$ is small and y' receives $\Omega(\frac{\log(n)}{\log(\log(n))})$ plateau balls. To bound $z_{y'}$, we will characterize it as the sum of contributions from three types of balls from $(X \backslash S) \times E$.[2]

Definition 7 (Ball characterizations)**.** *We characterize three types of balls:*

1. *Dense balls. Consider a set L of states of X, where a state of X is in L if $P(X = x) \geq \frac{1}{\log^3(n)}$. Dense balls are all balls of the form $(x \in L, e \in E)$. We call these dense balls, because the low-entropy of E will prevent the collective mass of these balls from “expanding” well.*

2. *Large balls. For all balls of the form $(x \in X \backslash (S \cup L), e \in E)$ where the ball has mass $\geq \frac{\mathcal{T}}{2}$.*

3. *Small balls. For all balls of the form $(x \in X \backslash (S \cup L), e \in E)$ where the ball has mass $< \frac{\mathcal{T}}{2}$.*

[2] In the proofs, with a slight abuse of notation, we use X, E both for the observed and exogenous variables, respectively and their supports.

We will show there are a non-negligible fraction of bins such that z_y is small. To do so, we will bound the contribution from dense balls by showing that the small entropy of E prevents "spread" in a sense, as there cannot be many states of Y that receive much contribution towards z_y from these dense balls. We will bound contribution from large balls by bounding the number of large balls, and showing that a non-negligible number of bins receive no large balls. Finally, we will bound contribution from small balls by showing how they often are mapped to states of Y that have yet to receive $\frac{\mathcal{T}}{2}$ mass from the corresponding state of X, meaning they often don't immediately increase z_y.

Finally, we will show (with high probability) the existence of a bin y' with small $z_{y'}$ that will receive many plateau balls, and how this will imply $H(X|Y = y') = \Omega(\log(\log(n)))$.

Complete Proof.

Bounding $H(\tilde{E})$ via $H(X|Y = y)$. Because $\tilde{E} \perp\!\!\!\perp Y$, it must be true that $H(\tilde{E}) \geq \max_y H(X|Y = y)$. This is simple to prove by data processing inequality and is shown in Step 1 of the proof of Theorem 1 by [12]. We aim to show there exists a y' such that $H(X|Y = y') = \Omega(\log(\log(n)))$.

Showing existence of a near-uniform plateau. First, we aim to find a subset of the support of X whose probabilities are multiplicatively close to one another. Here, we have a looser requirement for closeness than [12]. Instead of requiring these probabilities to be within a constant factor of each other, we allow them to be up to a factor of $\log^{c_{\text{close}}}(n)$ apart where c_{close} is a constant such that $0 < c_{\text{close}} < 1$. While there are multiple values of c_{close} that would be suitable for our analysis, for simplicity of presentation we choose $c_{\text{close}} = \frac{1}{4}$ throughout. This set of states of X that are multiplicatively close to one another will be called the *plateau* of X. We begin by showing how the $(\Omega(n), \Omega(\frac{1}{n \log(n)}))$-support assumption implies a plateau of states of X:

Lemma 10 (Plateau existence)**.** *Suppose X has $(c_{support}n, \frac{1}{c_{lb} n \log(n)})$-support for constants $0 < c_{support} \leq 1$ and $c_{lb} \geq 1$. Additionally, assume n is sufficiently large such that $\frac{\log(2c_{lb}/c_{support})}{\log(\log(n))} \leq 1$. Then, there exists a subset $S \subseteq [n]$ of the support of X, such* *that the following three statements hold:*

1. $\frac{\max_{i \in S} P(X=i)}{\min_{i \in S} P(X=i)} \leq \log^{c_{close}}(n)$

2. $\min_{i \in S} P(X = i) \geq \frac{1}{c_{lb} n \log(n)}$

3. $|S| \geq \frac{c_{close} c_{support} n}{6}$*, for any* $0 < c_{close} < 1$.

Proof. By definition of $(c_{\text{support}} n, \frac{1}{c_{\text{lb}} n \log(n)})$ support, there are at least $c_{\text{support}} n$ states of X with probability in range $[\frac{1}{c_{\text{lb}} n \log(n)}, 1]$. Moreover, at most $\frac{c_{\text{support}} n}{2}$ states will have probabilities in range $[\frac{2}{c_{\text{support}} n}, 1]$. Otherwise, they would have total probability mass > 1 which is impossible. Therefore, there are at least $\frac{c_{\text{support}} n}{2}$ states with probabilities in range $[\frac{1}{c_{\text{lb}} n \log(n)}, \frac{2}{c_{\text{support}} n}]$.

Now, we aim to divide the range $[\frac{1}{c_{\text{lb}} n \log(n)}, \frac{2}{c_{\text{support}} n}]$ into a number of contiguous segments such that all values in any segment are multiplicatively within $\log^{c_{\text{close}}}(n)$ of each other. To do so, we can create segments $[\frac{1}{c_{\text{lb}} n \log(n)} \times (\log^{c_{\text{close}}}(n))^i, \frac{1}{c_{\text{lb}} n \log(n)} \times (\log^{c_{\text{close}}}(n))^{i+1}]$ from $i = 0$ until the smallest i that satisfies $\frac{1}{c_{\text{lb}} n \log(n)} \times (\log^{c_{\text{close}}}(n))^{i+1} \geq \frac{2}{c_{\text{support}} n}$. Accordingly, we need $\lceil \frac{\log((2/(c_{\text{support}} n))/(1/(c_{\text{lb}} n \log(n))))}{\log(\log^{c_{\text{close}}}(n))} \rceil \leq 1 + \frac{1}{c_{\text{close}}} + \frac{\log(2c_{\text{lb}}/c_{\text{support}})}{c_{\text{close}} \log(\log(n))} \leq \frac{3}{c_{\text{close}}}$ groups. Hence one group must have at least $\frac{c_{\text{support}} n/2}{3/c_{\text{close}}} = \frac{c_{\text{close}} c_{\text{support}} n}{6}$ states of X that are multiplicatively within $\log^{c_{\text{close}}}(n)$ and have probability at least $\frac{1}{c_{\text{lb}} n \log(n)}$. □

Lower-bounding the most probable state of E. Our proof method focuses on a balls-and-bins game where states of $X \times E$ are balls and states of Y are bins. We focus first on *plateau balls*, which are balls corresponding to states of S (the set of plateau states of X) and the highest probability state of E. In particular, they are balls of the form $(X \in S, E = e_1)$ where e_1 is the most probable state of E. To show that these plateau balls have enough probability mass to be helpful, we first show that $P(E = e_1)$ is relatively large:

Lemma 11. *If* $H(E) \leq c_{close} \log(\log(n))$ *then* $P(E = e_1) \geq \frac{1}{\log^{c_{close}}(n)}$

Proof. For any distribution with entropy H, its state with the highest probability has at least probability 2^{-H} (see Lemma 5 of [12]). Thus if $H(E) \leq c_{\text{close}} \log(\log(n))$ then $P(E = e_1) \geq 2^{-c_{\text{close}} \log(\log(n))} = \frac{1}{\log^{c_{\text{close}}}(n)}$. □

Introducing surplus. We now begin proving how there exists a bin that receives a large amount of mass that helps the bin have large conditional entropy (such helpful mass includes the plateau balls), and not much mass that hurts the conditional entropy making it small. To formalize this hurtful mass, recall the *surplus* quantity described in Definition 6. This surplus is a way of quantifying the probability mass received by a state of Y that is hurtful towards making the conditional entropy large. We define surplus, with the threshold of $\mathcal{T}$ specified as $\frac{12}{n\log(n)}$ as follows:

Definition 8 (Surplus, $\mathcal{T} = \frac{12}{n\log(n)}$)**.** *We define the surplus of a state i of Y as $z_i = \sum_{j \notin S} \max(0, P(X = j, Y = i) - \frac{12}{n\log(n)})$, where S is the set of plateau states of X.*

Characterizing balls-and-bins. Now we will show that there are a non-negligible number of states of Y where the surplus is small. Recall from our proof outline that we view the process of realizing the random function f as a balls-and-bins game. In particular, each element of $X \times E$ (a ball) is i.i.d. uniformly randomly assigned to a state of Y (a bin). Only balls of the form $(x \in X \backslash S, e \in E)$ affect a bin's surplus. To bound surplus for bins, we characterize it as the sum of contributions from three types of balls from $(X \backslash S) \times E$, and restate this characterization from the proof outline:

Definition 7 (Ball characterizations)**.** *We characterize three types of balls:*

1. *Dense balls. Consider a set L of states of X, where a state of X is in L if $P(X = x) \geq \frac{1}{\log^3(n)}$. Dense balls are all balls of the form $(x \in L, e \in E)$. We call these dense balls, because the low-entropy of E will prevent the collective mass of these balls from "expanding" well.*

2. *Large balls. For all balls of the form $(x \in X \backslash (S \cup L), e \in E)$ where the ball has mass $\geq \frac{\mathcal{T}}{2}$.*

3. *Small balls. For all balls of the form $(x \in X \backslash (S \cup L), e \in E)$ where the ball has mass $< \frac{\mathcal{T}}{2}$.*

Bounding the harmful effects of dense balls. Recall that $\mathcal{T} = \frac{12}{n\log(n)}$. Now, we show how to bound the contribution of dense balls towards surplus. By our assumptions, $Y = f(X,E)$, and $H(E)$ is small, meaning there is not much randomness in our function. We defined L as states of X with probability at least $\frac{1}{\log^3(n)}$, so $|L| \leq \log^3(n)$. We would like to show that there are not too many bins where the dense balls contribute a significant amount to surplus. If $H(E) = 0$, this would be easy to show as then there would only be $|L| \leq \log^3(n)$ dense balls and thus they could only affect the surplus of $\log^3(n)$ bins. However, we aim to show this claim in the more general setting where $H(E) = o(\log(\log(n)))$. To accomplish this, we follow the same intuition to show that the limited entropy of E prevents this small number of states of X from greatly "spreading" to significantly affect a large number of states of Y. In particular, we show:

Lemma 12 (Limited expansion)**.** *Suppose Y can be written as a function $f(X,E)$ and $X \perp\!\!\!\perp E$. Consider any subset R of the support of X. For any subset T of the support of Y that satisfies $\forall t \in T : P(X \in R, Y = t) > \delta$, the cardinality of T is upper bounded as $|T| \leq \frac{H(E)+\log(|R|)+2}{\delta\log(\frac{1}{\delta})}$.*

Proof. Consider a variable X', whose distribution is obtained from the distribution of X by keeping only the states in R, and then normalized. More formally, for any $i \in R$, $P(X' = i) = \frac{P(X=i)}{P(X \in R)}$, and for any $i \notin R$, $P(X' = i) = 0$.

Recall $Y = f(X,E)$. Using the same f, E, we define $Y' = f(X', E)$. Note that $P(X \in R, Y = i) \leq P(Y' = i)$. If $P(X \in R, Y = i) \geq \delta$, then it must be true that $P(Y' = i) \geq \delta$. Moreover, this implies that if there exists such a subset T then $H(Y') \geq |T|\delta\log(\frac{1}{\delta}) - 2$ (note the negative two is from the fact that modifying a distribution by adding non-negative numbers to probabilities can decrease entropy by at most 2). Moreover, by data-processing inequality note that $H(Y') \leq H(X') + H(E|X') \leq H(X') + H(E) \leq \log(|R|) + H(E)$, where previous inequality is due to the fact that conditioning reduces entropy. This implies the desired inequality for the cardinality of set T. □

To more directly use this for our goal, we present:

Corollary 5. *There exist no subset $|T| = n/4$ such that $\forall t \in T : P(X \in L, Y = t) \geq \frac{1}{n \log(\log(n)) \log^{2c_{close}}(n)}$*

Proof. Note that $|L| \leq \log^3(n)$. By Lemma 12, any such T must satisfy:

$$\begin{aligned}
&|T| && (3.1)\\
&\leq \frac{H(E) + \log(|L|) + 2}{1/(n \cdot \log(\log(n)) \cdot \log^{2c_{\text{close}}}(n)) \cdot \log(n)} && (3.2)\\
&\leq \frac{5\log^2(\log(n)) \cdot n \cdot \log^{2c_{\text{close}}}(n)}{\log(n)} && (3.3)\\
&\leq \frac{5\log^2(\log(n)) \cdot n \cdot \log^{1/2}(n)}{\log(n)} && (3.4)\\
&\leq \frac{n}{4} && (3.5)
\end{aligned}$$

We obtain Step 3.4 by previously setting $c_{\text{close}} = \frac{1}{4}$. We obtain Step 3.5 when n is sufficiently large such that $\frac{5\log^2(\log(n))}{\log^{1/2}(n)} \leq \frac{1}{4}$. It can be shown that $n \geq 5$ is sufficient. □

As a result, dense balls cannot significantly affect the surplus of many bins.

Bounding the harmful effects of large balls. We now show how large balls cannot significantly affect the surplus of too many bins, by showing there is a non-negligible number of bins that receive no large balls.

Lemma 13 (Avoided big)**.** *Given a balls-and-bins game with $c \cdot n\ln(n)$ balls mapped uniformly randomly to n bins, at least $\frac{n^{1-c}}{2}$ bins will receive no balls with high probability if c is a constant such that $0 < c \leq \frac{1}{3}$.*

Proof. This follows directly from [67]. By [67], with high probability the number of empty bins will be $n^{1-c} \pm O(\sqrt{n\log(n)})$. For sufficiently large n, $O(\sqrt{n\log(n)}) \leq \frac{n^{2/3}}{2} \leq \frac{n^{1-c}}{2}$ and thus the number of empty bins is at least $\frac{n^{1-c}}{2}$ with high probability. □

Note how this relates to the coupon collector's problem, where it is well-known that $\Theta(n\log(n))$ trials are necessary and sufficient to receive at least one copy of all coupons with high probability. This is analogous to the number of balls needed such that every bin has at least one ball. The result of Lemma 13 is intuitive from the coupon collector's problem, because the number of trials needed concentrates very well. Meaning, with a constant-factor less number of trials than the expectation required, there are many coupons that have not yet been collected with high probability.

Corollary 6. *As there are at most $\frac{1}{\mathcal{T}/2} \leq \frac{n\log(n)}{6} \leq \frac{1}{4} \cdot n\ln(n)$ large balls, with high probability there are at least $\frac{n^{3/4}}{2}$ bins that receive no large balls.*

Bounding the harmful effects of small balls. For the small balls, we will also show that they cannot contribute too much surplus to too many states of Y. We will notably use that all small balls correspond to a state of X where $P(X = x) \leq \frac{1}{\log^3(n)}$. We will utilize this to show that most small balls are assigned to a state of Y that has not yet received $> \frac{\mathcal{T}}{2}$ mass from its corresponding state of X, and accordingly would not increase the surplus. To accomplish this, we define a surplus quantity that only takes into account small balls:

Definition 9 (Small ball surplus)**.** *We define the small ball surplus of a state y of Y as*

$$z_y^{small} = \sum_{x \notin (S \cup L)} \max\left(0, \left(\sum_{\substack{e: \\ P(X=x,E=e) \\ < \frac{\mathcal{T}}{2}}} P(X = x, E = e, Y = y)\right) - \mathcal{T}\right).$$

With this notion of surplus constrained to small balls, we show the following:

Lemma 14 (Small ball limited surplus)**.** *With high probability, there are at most $\frac{n}{4}$ values of i, i.e., number of bins, where $z_i^{small} \geq \frac{1}{n\log(\log(n))\log^{2c_{close}}(n)}$.*

Proof. We will consider all small balls in an arbitrary order. Let $x(t)$ be the corresponding state of X for the t-th small ball, $e(t)$ the corresponding state of E, and $w_{\text{ball}}(t)$ be the ball's probability mass (i.e., $P(X = x(t), E = e(t))$). Recall that for all small balls it must hold that $x(t) \notin L$ and thus $P(X = x(t)) < \frac{1}{\log^3(n)}$. We define the total small ball surplus as $Z^{\text{small}} = \sum_{y \in Y} z_y^{\text{small}}$. Now, we will consider all small balls in an arbitrary order and realize their corresponding entry of f to map them to a state of Y. Initially, we have not realized the entry of f for any balls and thus all $z_y^{\text{small}} = 0$ and $Z^{\text{small}} = 0$. As we map small balls to states of Y, we define $\Delta(t)$ as the increase of Z^{small} after mapping the t-th ball to a state of Y. By definition, $\sum_t \Delta(t)$ is equal to Z^{small} after all values of f have been completely realized.

Our primary intuition is that we will show for many small balls it holds that $\Delta(t) = 0$. As a result, we expect Z^{small} to not be very large.

As a result, we expect Z^{small} to not be very large. Let $y(t)$ be equal to $f(x(t), e(t))$, the state of Y that the t-th ball is mapped to. As f is realized for each configuration, let $w_Y^t(y, x)$ denote the total mass of balls assigned to state y of Y so far from state x of X, i.e., $w_Y^t(y', x') := \sum\limits_{x', e : f(x', e) = y'} w_{\text{ball}}(t')$.

We upper-bound the expectation of $\Delta(t)$:

Claim 1. Regardless of the realizations of all $\Delta(t')$ for $t' < t$, it holds that $\Delta(t)$ is a random variable with values in range $[0, w_{\text{ball}}(t)]$ and $E[\Delta(t)] \leq \frac{w_{\text{ball}}(t)}{\log^2(n)}$.

Proof. The only conditions under which $\Delta(t)$ takes a positive value (which is upper-bounded by $w_{\text{ball}}(t)$), is when $w_Y^t(y(t), x(t)) > \frac{\mathcal{T}}{2}$ before the t-th ball is realized. Recall that $P(x(t)) \leq \frac{1}{\log^3(n)}$. Accordingly, the number of states y' of Y where $w_Y^t(y', x(t)) > \frac{\mathcal{T}}{2}$ is upper-bounded by $\frac{P(X = x(t))}{\mathcal{T}/2} \leq \frac{1/\log^3(n)}{6/(n \log(n))} = \frac{n \log(n)}{6 \log^3(n)} \leq \frac{n}{\log^2(n)}$. This is due to the fact that balls partition the total mass of $P(X = x(t))$ since we have $P(X = x(t)) = \sum_e P(X = x(t), E = e)$. This implies that the probability that the t-th ball will be mapped to a state y' of Y such that $w_Y^t(y', x(t))$ already exceeds the threshold of $\mathcal{T}/2$ (in other words where we might have $\Delta(t) > 0$) is upper-bounded by $\frac{n/\log^2(n)}{n} = \frac{1}{\log^2(n)}$ due to the fact that the function f is realized independently and uniformly randomly for each pair of (x, e), i.e., for every distinct ball. Accordingly, $E[\Delta(t)] \leq \frac{w(t)}{\log^2(n)}$. □

This enables us to upper-bound the sum of $\Delta(t)$:

Claim 2. $\sum_t \Delta(t) \leq \frac{1}{4 \log(n)}$ with high probability.

Proof. We will transform $\Delta(t)$ into a martingale. In particular, we define $\Delta'(t) = \Delta'(t-1) + \Delta(t) - E[\Delta(t)|\Delta(1), \ldots, \Delta(t-1)]$. We define $\Delta'(0) = 0$, and note that $\Delta'(c)$ is a martingale. By Azuma's inequality, we show $|\sum_t \Delta'(t)| \leq \frac{1}{8\log(n)}$ with high probability:

$$
\begin{aligned}
&P[|\Delta(t)| > \varepsilon] < 2e^{-\frac{\varepsilon^2}{2\sum c_i^2}} \\
&\leq 2e^{-\frac{\left(\frac{1}{8\log(n)}\right)^2}{2(\max_i c_i)\cdot\sum c_i}} \\
&\leq 2e^{-\frac{\left(\frac{1}{8\log(n)}\right)^2}{2\times\mathcal{T}/2\cdot 1}} \\
&= 2e^{\frac{-n\log(n)}{12\times 8\times|V|\times\log(n)}}
\end{aligned}
$$

Accordingly, by definition of $\Delta'(t)$ this implies $|(\sum_t \Delta(t)) - \sum_c E[\Delta(t)|\Delta(1), \ldots, \Delta(c-1)]| \leq \frac{1}{8\log(n)}$. By Claim 16 we know all $E[\Delta(t)|\Delta(1), \ldots, \Delta(c-1)] \leq \frac{w_{\text{config}(c)}}{\log^2(n)}$ and accordingly, $\sum_t E[\Delta(t)|\Delta(1), \ldots, \Delta(c-1)] \leq \frac{1}{\log^2(n)}$. Together, these imply $\sum_t \Delta(t) \leq \frac{1}{8\log(n)} + \frac{1}{\log^2(n)}$ with high probability, and for sufficiently large n it holds that $\frac{1}{\log^2(n)} \leq \frac{1}{8\log(n)}$. Thus, our high-probability on $|\Delta'(t)|$ implies that $\sum_t \Delta(t) \leq \frac{1}{4\log(n)}$ with high probability. □

Finally, we conclude that our upper-bound on $\sum_t \Delta(t)$ implies an upper-bound on the number of states of Y with non-negligible small ball support:

Claim 3. If $\sum_t \Delta(t) \leq \frac{1}{4\log(n)}$, then there are at most $\frac{n}{4}$ bins where
$z_i^{\text{small}} \geq \frac{1}{n\log(\log(n))\log^{2c_{\text{close}}}(n)}$.

Proof. $Z^{\text{small}} = \sum_t \Delta(t) \leq \frac{1}{4\log(n)}$. Given this upper-bound for total small ball surplus, we can immediately upper-bound the number of states of Y with small ball surplus greater than $\frac{1}{n\log(\log(n))\log^{2c_{\text{close}}}(n)}$ by the quantity $\frac{1/(4\log(n))}{1/(n\cdot\log(\log(n))\cdot\log^{2c_{\text{close}}}(n))} \leq \frac{n\cdot\log(\log(n))\cdot\log^{1/2}(n)}{4\log(n)} \leq \frac{n}{4}$. We obtain this by using $c_{\text{close}} = \frac{1}{4}$ and for sufficiently large n such that $\log(\log(n)) \leq \log^{1/2}(n)$.

□

□

Combining the three ball types: many bins with small surplus. Now, we combine all these intuitions to show there are many bins that have a small amount of surplus. We have shown that, with high probability, the are at most $n/4$ bins with non-negligible mass from dense balls by Corollary 5, and at most $n/4$ bins with non-negligible mass from small balls Lemma 20. Combining these sets, there are at most $n/2$ bins with non-negligible mass from dense balls or small balls. By Corollary 11, with high probability at least $\frac{n^{3/4}}{2}$ bins will receive no large balls. Our goal is to show the intersection of the sets is large, so there are many bins that have small surplus.

Lemma 15. *Let there be two sets $A, B \subseteq [n]$, where $|A| \geq \frac{n}{2}$ and A and B are both independently uniformly random subsets of size $|A|$ and $|B|$, respectively. It holds that $P(|A \cap B| \geq \frac{|B|}{4}) \geq 1 - 2e^{\frac{-|B|}{8}}$.*

Proof. To accomplish this, we will heavily utilize properties of negative association (NA). Lemma 8 of [67] shows that permutation distributions are NA. Lemma 9 of [67] shows closure properties of NA random variables. In particular, they show that concordant monotone functions defined on disjoint subsets of a set of NA variables are also NA. Accordingly, consider concordant monotone functions where each bin i has a random variable $\mathcal{A}_i$ that takes value 1 if it is the first $|A|$ values of a permutation distribution and value 0 otherwise. These random variables are thus NA. Suppose we first realize the set B, independently of the realization of A. Then, a bin $y \in B$ would be in $A \cap B$ if $A_y = 1$. It is clear this formulation of the random process has a bijective mapping with the true random process, so $P(|A \cap B| \geq \frac{|B|}{4}) = P(\sum_{y \in B} \mathcal{A}_y \geq \frac{|B|}{4})$. By Theorem 5 of [67], we can use Hoeffding's upper tail bound to show $P(\sum_{y \in B} \mathcal{A}_y < \frac{|B|}{4}) \leq P(|\sum_{y \in B} \mathcal{A}_y - E[\sum_{y \in B} \mathcal{A}_y]| > \frac{|B|}{4}) \leq 2e^{\frac{-|B|}{8}}$. □

Corollary 7. *With high probability, there are at least $\frac{n^{3/4}}{8}$ bins with surplus $z_y \leq \frac{2}{n \log(\log(n)) \log^{2c_{close}}(n)}$.*

Proof. We have defined three types of balls, and have proven results that show how there are many bins with negligible bad contribution for each type of ball. Now, we combine these with Lemma 15 to show there are many bins where there is not much bad contribution in total. By Corollary 5 there are at most $n/4$ bins with more than $\frac{1}{n\log(\log(n))\log^{2c_{\text{close}}(n)}(n)}$ mass from dense balls. By Lemma 20, there are at most $n/4$ bins with small ball surplus more than $\frac{1}{n\log(\log(n))\log^{2c_{\text{close}}(n)}(n)}$. Let A be the set of bins with at most $\frac{1}{n\log(\log(n))\log^{2c_{\text{close}}(n)}(n)}$ mass from dense balls and at most $\frac{1}{n\log(\log(n))\log^{2c_{\text{close}}(n)}(n)}$ small ball surplus. By combining Corollary 5 and Lemma 20 we know $|A| \geq \frac{n}{2}$ with high probability. Let B be the set of bins that receive no big balls. By Corollary 11, it holds that $|B| \geq \frac{n^{3/4}}{2}$ with high probability. By Lemma 15, it holds that $|A \cap B| \geq \frac{n^{3/4}}{8}$ with failure probability at most $2e^{\frac{-2n^{3/4}}{16}}$. Moreover, all such bins will have total surplus at most $\frac{2}{n\log(\log(n))\log^{2c_{\text{close}}}(n)}$, because they receive no large balls and total surplus is then upper-bounded by the sum of small ball surplus and total mass from dense balls. □

Existence of a small surplus bin with many plateau balls. Recall plateau balls, which are balls of $X \times E$ that take the form $(x \in S, E = e_1)$, where e_1 is the most probable state of E. We show that at least one of the bins with small surplus will receive many plateau balls with high probability:

Lemma 16. *There exists a bin with surplus at most* $\frac{2}{n\log(\log(n))\log^{2c_{close}}(n)}$ *and at least* $\frac{\log(n)}{2\log(\log(n))}$ *plateau balls.*

Proof. Note that total surplus is independent of how plateau balls are mapped. Accordingly, we have determined a set of $\frac{n^{3/4}}{8}$ bins with small enough surplus. We aim to show that one of these bins receives a large number of plateau balls with high probability. We will rely on negative association (NA) in the balls-and-bins process to prove our result.

Claim 4. Indicator variables for if a bin receives some threshold of balls in a i.i.d. uniformly random balls-and-bins game are NA.

Proof. This follows immediately by using results of [67]. By Theorem 10 of [67], the random variables of the number of balls assigned to each bin are NA. By Lemma 9 of [67], concordant monotone functions define on disjoint subsets of a set of NA random variables are NA. Accordingly, if we have an indicator variable for whether a bin receives at least some number of balls, these indicator variables are NA. □

Now, we lower-bound the expectation of these indicator variables:

Claim 5. Suppose cn balls ($c \leq 1$) are thrown i.i.d. uniformly randomly into n bins. The probability that a particular bin receives at least $k = \frac{d\log(n)}{\log(\log(n))}$ balls is at least $\frac{1}{en^d}$ given that $\frac{d}{c} \leq \log(\log(n))$.

Proof. We use the method outlined by [13]. We lower-bound the probability of a bin receiving at least k balls as follows:

$$\begin{aligned}\binom{cn}{k}\cdot(\frac{1}{n})^k\cdot(1-\frac{1}{n})^{cn-k} &\geq (\frac{cn}{k})^k\cdot\frac{1}{n^k}\cdot\frac{1}{e}\\ &\geq \frac{1}{e}\cdot(\frac{c}{k})^k\\ &= \frac{1}{e}\cdot(\frac{c\log(\log(n))}{d\log(n)})^{\log_{\log(n)}(n^d)}\\ &\geq \frac{1}{e}\cdot(\frac{1}{\log(n)})^{\log_{\log(n)}(n^d)} \qquad (3.6)\\ &= \frac{1}{en^d}\end{aligned}$$

We obtain Step 3.6 by using $\frac{d}{c} \leq \log(\log(n))$. □

By Lemma 10 there are at least $\frac{c_{\text{close}}c_{\text{support}}}{6}\cdot n = \frac{c_{\text{support}}}{24}\cdot n$ plateau balls. Now, consider NA indicator variables $\mathcal{B}_i$ for whether or not a particular bin receives at least $\frac{\log(n)}{2\log(\log(n))}$ plateau balls. By Claim 4, these indicator variables are NA. By Claim 5, it holds that $E[\mathcal{B}_i] \geq \frac{1}{en^{0.5}}$ for sufficiently large n where $\frac{1/2}{c_{\text{support}}/24} = \frac{12}{c_{\text{support}}} \leq \log(\log(n))$. Finally, we can upper-bound the probability that $\mathcal{B}_i = 0$ for all bins with small enough surplus, of which there are at least $\frac{n^{3/4}}{8}$. Using marginal probability bounds for NA variables shown in Corollary 3 of [67], all such $\mathcal{B}_i = 0$ with probability at most $(\frac{1}{en^{0.5}})^{\frac{n^{3/4}}{8}}$. □

Proving large conditional entropy. Finally, we show how the existence of a bin with small surplus and many plateau balls implies that the bin has large conditional entropy:

Lemma 17 (High-entropy conditional)**.** *Given a bin y' that has $z_{y'} \leq \frac{2}{n\cdot\log(\log(n))\cdot\log^{2c_{close}}(n)}$, and receives $\frac{\log(n)}{2\log(\log(n))}$ plateau balls, then $H(X|Y = y') =$*

$\Omega(\log(\log(n)))$.

Proof. To show $H(X|Y=y')$ is large, we first define the vector v such that $v(x) = P(X=x, Y=y')$. Similarly, we define $\overline{v}(x) = \frac{v}{P(Y=y')}$, meaning $\overline{v}(x) = P(X=x|Y=y')$ and $|\overline{v}|_1 = 1$. Our underlying goal is to show $H(\overline{v})$ is large. To accomplish this, we will split the probability mass of v into three different vectors $v_{\text{initial}}, v_{\text{plateau}}, v_{\text{surplus}}$ such that $v = v_{\text{initial}} + v_{\text{plateau}} + v_{\text{surplus}}$. The entries of v_{plateau} will correspond to mass from plateau states of X, v_{initial} will correspond to the first $\mathcal{T}$ mass from non-plateau states of X, and v_{surplus} will correspond to mass that contributes to the surplus $z_{y'}$. We more formally define the three vectors as follows:

- v_{plateau}. The vector of probability mass from plateau states of X. $v_{\text{plateau}}(x)$ is 0 if $x \notin S$ and $v_{\text{plateau}}(x) = P(X=x, Y=y')$ if $x \in S$.

- v_{initial}. For non-plateau states of X, their first $\mathcal{T}$ probability mass belongs to v_{initial}. $v_{\text{initial}}(x) = \min(P(X=x, Y=y'), \mathcal{T})$ if $x \notin S$ and $v_{\text{initial}}(x) = 0$ otherwise.

- v_{surplus}. For non-plateau states of X, their probability mass beyond the first $\mathcal{T}$ mass belongs to v_{surplus}. This corresponds to the surplus quantity. $v_{\text{surplus}}(x) = \max(0, P(X=x, Y=y') - \mathcal{T})$ if $x \notin S$ and $v_{\text{surplus}}(x) = 0$ otherwise. By this definition, $z_{y'} = |v_{\text{surplus}}|_1$.

To show $H(X|Y=y') = H(\overline{v})$ is large, we divide our approach into two steps:

1. Show there is substantial helpful mass: $|v_{\text{initial}} + v_{\text{plateau}}|_1$
$= \Omega\left(\frac{1}{n \cdot \log(\log(n)) \cdot \log^{2c_{\text{close}}}(n)}\right)$

2. Show the distribution of helpful mass has high entropy: $H\left(\frac{v_{\text{initial}}+v_{\text{plateau}}}{|v_{\text{initial}}+v_{\text{plateau}}|_1}\right) = \Omega(\log(\log(n)))$.

3. Show that, even after adding the hurtful mass, the conditional entropy is large: $H(X|Y=y') = H(\overline{v}) \geq H\left(\frac{v_{\text{initial}}+v_{\text{plateau}}}{|v_{\text{initial}}+v_{\text{plateau}}|_1}\right) - O(1) = \Omega(\log(\log(n)))$

In the first step, we are showing that the distribution when focusing on just the helpful mass of $v_{\text{initial}}, v_{\text{plateau}}$ has high a substantial amount of probability mass. In the second step, we prove how this distribution of helpful mass has high entropy. In the third step, we show that the hurtful mass of v_{surplus} does not decrease entropy more than a constant.

First, we show that there is a substantial amount of helpful mass:

Claim 6. $|v_{\text{initial}} + v_{\text{plateau}}|_1 = \frac{1}{2c_{\text{lb}}n \cdot \log(\log(n)) \cdot \log^{2c_{\text{close}}}(n)}$

Proof. Recall that the bin y' received $\frac{\log(n)}{2\log(\log(n))}$ plateau balls. As defined in Lemma 10, the set S of plateau states is defined such that $\frac{\max_{x \in S} P(X=x)}{\min_{x \in S} P(X=x)} \leq \log^{c_{\text{close}}}(n)$ and $\min_{x \in S} P(X = x) \geq \frac{1}{c_{\text{lb}} n \log(n)}$. Also recall that by Lemma 11 the most probably state of E has large probability. In particular, $P(E = e_1) \geq \frac{1}{\log^{c_{\text{close}}}(n)}$. Let the subset $S' \subseteq S$ be the subset of plateau states of X such that their plateau ball is mapped to y'. In particular, for every $x \in S'$ it holds that $f(x, e_1) = y'$. Accordingly, $P(X = x, Y = y') \geq P(X = x) \cdot P(E = e_1)$ for $x \in S'$. Thus, the total weight from plateau states of X is at least $|S'| \cdot \min_{x \in S'} P(X = x) \cdot P(E = e_1) \geq |S'| \cdot \frac{\max_{x \in S'} P(X=x)}{\log^{c_{\text{close}}}(n)} \cdot P(E = e_1) \geq \frac{1}{2c_{\text{lb}} n \log(\log(n)) \log^{2c_{\text{close}}}(n)}$. □

Next, we show the distribution of helpful mass has high entropy:

Claim 7. $H\left(\frac{v_{\text{initial}}+v_{\text{plateau}}}{|v_{\text{initial}}+v_{\text{plateau}}|_1}\right) \geq \frac{\log(\log(n))}{4}$

Proof. Let us define $\overline{v}_{\text{helpful}} = \frac{v_{\text{initial}}+v_{\text{plateau}}}{|v_{\text{initial}}+v_{\text{plateau}}|_1}$ to be the vector of helpful mass, and we will show $H(\overline{v}_{\text{helpful}})$ is large by upper-bounding $\max_x \overline{v}_{\text{helpful}}(x)$.

For non-plateau states of X, it follows from Claim 20 that $\max_{x \notin S} \overline{v}_{\text{helpful}}(x) \leq \frac{\mathcal{T}}{|v_{\text{initial}}+v_{\text{plateau}}|_1} \leq \frac{\mathcal{T}}{\frac{1}{2c_{\text{lb}} n \cdot \log(\log(n)) \cdot \log^{2c_{\text{close}}}(n)}} = \frac{24c_{\text{lb}} \log(\log(n)) \cdot \log^{2c_{\text{close}}}(n)}{\log(n)}$.

For plateau states of X, in Claim 20 we also developed the lower-bound of $|v_{\text{initial}} + v_{\text{plateau}}|_1 \geq |S'| \cdot \frac{\max_{x \in S'} P(X=x)}{\log^{c_{\text{close}}}(n)} \cdot P(E = e_1) \geq \frac{\log(n) \cdot \max_{x \in S'} P(X=x)}{2\log^{2c_{\text{close}}}(n) \log(\log(n))}$. Accordingly, we can upper-bound $\max_{x \in S'} \overline{v}_{\text{helpful}}(x) \leq \frac{\max_{x \in S'} P(X=x)}{|v_{\text{initial}}+v_{\text{plateau}}|_1} \leq \frac{2\log(\log(n)) \log^{2c_{\text{close}}}(n)}{\log(n)}$.

Accordingly, we can lower-bound the entropy of $H(\overline{v}_{\text{helpful}}) = \sum_x \overline{v}_{\text{helpful}}(x) \cdot \log(\frac{1}{\overline{v}_{\text{helpful}}(x)}) \geq \sum_x \overline{v}_{\text{helpful}}(x) \cdot \log(\frac{1}{\max_{x'} \overline{v}_{\text{helpful}}(x')}) = \log(\frac{1}{\max_{x'} \overline{v}_{\text{helpful}}(x')}) \geq \log(\frac{24c_{\text{lb}} \log(n)}{\log^{2c_{\text{close}}}(n) \log(\log(n))}) = (1 - 2c_{\text{close}}) \log(\log(n)) - \log(\log(\log(n))) - \log(24c_{\text{lb}}) =$

$\frac{\log(\log(n))}{2} - \log(\log(\log(n))) - \log(24c_{\text{lb}}) \geq \frac{\log(\log(n))}{4}$ for sufficiently large n where $\frac{\log(\log(n))}{2} \geq \log(\log(\log(n))) + \log(24c_{\text{lb}})$. □

Finally, we show the hurtful mass does not decrease entropy much, and thus our conditional distribution has high entropy:

Claim 8. $H(X|Y=y') = H(\overline{v}) \geq \Omega(1) \cdot H\left(\frac{v_{\text{initial}}+v_{\text{plateau}}}{|\overline{v}_{\text{initial}}+\overline{v}_{\text{plateau}}|_1}\right) - O(1) = \Omega(\log(\log(n)))$

Proof. We lower-bound $H(\overline{v})$ with the main intuitions that $H\left(\frac{v_{\text{initial}}+v_{\text{plateau}}}{|v_{\text{initial}}+v_{\text{plateau}}|_1}\right) = \Omega(\log(\log(n)))$ and $\frac{|v_{\text{initial}}+v_{\text{plateau}}|_1}{|v_{\text{initial}}+v_{\text{plateau}}+v_{\text{surplus}}|_1} = \Omega(1)$. We more precisely obtain this lower-bound for $H(\overline{v})$ as follows:

$$\begin{aligned}
H(\overline{v}) &= H\left(\frac{v_{\text{initial}}+v_{\text{plateau}}+v_{\text{surplus}}}{|v_{\text{initial}}+v_{\text{plateau}}+v_{\text{surplus}}|_1}\right) \\
&= \sum_x \frac{v_{\text{initial}}(x)+v_{\text{plateau}}(x)+v_{\text{surplus}}(x)}{|v_{\text{initial}}+v_{\text{plateau}}+v_{\text{surplus}}|_1} \times \\
&\qquad \log \frac{|v_{\text{initial}}+v_{\text{plateau}}+v_{\text{surplus}}|_1}{v_{\text{initial}}(x)+v_{\text{plateau}}(x)+v_{\text{surplus}}(x)} \\
&\geq \sum_x \frac{v_{\text{initial}}(x)+v_{\text{plateau}}(x)}{|v_{\text{initial}}+v_{\text{plateau}}+v_{\text{surplus}}|_1} \times \\
&\qquad \log \frac{|v_{\text{initial}}+v_{\text{plateau}}+v_{\text{surplus}}|_1}{v_{\text{initial}}(x)+v_{\text{plateau}}(x)} - 2 \qquad (3.7) \\
&\geq \sum_x \frac{v_{\text{initial}}(x)+v_{\text{plateau}}(x)}{|v_{\text{initial}}+v_{\text{plateau}}+v_{\text{surplus}}|_1} \times \\
&\qquad \log \frac{|v_{\text{initial}}+v_{\text{plateau}}|_1}{v_{\text{initial}}(x)+v_{\text{plateau}}(x)} - 2 \\
&= \frac{|v_{\text{initial}}+v_{\text{plateau}}|_1}{|v_{\text{initial}}+v_{\text{plateau}}+v_{\text{surplus}}|_1} H\left(\frac{v_{\text{initial}}+v_{\text{plateau}}}{|v_{\text{initial}}+v_{\text{plateau}}|_1}\right) - 2 \\
&= \frac{|v_{\text{initial}}+v_{\text{plateau}}|_1}{|v_{\text{initial}}+v_{\text{plateau}}|_1 + z_{y'}} H\left(\frac{v_{\text{initial}}+v_{\text{plateau}}}{|v_{\text{initial}}+v_{\text{plateau}}|_1}\right) - 2 \\
&\geq \frac{1}{1+2c_{\text{lb}}} \cdot H\left(\frac{v_{\text{initial}}+v_{\text{plateau}}}{|v_{\text{initial}}+v_{\text{plateau}}|_1}\right) - 2 \qquad (3.8) \\
&= \Omega(\log(\log(n))) \qquad (3.9)
\end{aligned}$$

To obtain Step 3.10, we note that all summands are manipulated from the form $\sum_x p_x \log(\frac{1}{p_x})$ to $\sum_x p'_x \log(\frac{1}{p'_x})$ where $p'_x \leq p_x$ for all x. As the derivative of $p\log(\frac{1}{p})$ is non-negative for $0 \leq p \leq \frac{1}{e}$, the value of at most two summands can decrease, and they can each decrease by at most one. To obtain Step 3.11, we use Claim 20. To obtain Step 3.12, we use Claim 22. □

Thus, we have shown $H(X|Y = y') = \Omega(\log(\log(n)))$. □

Corollary 8. *Under our assumptions,* $H(X|Y = y') = \Omega(\log(\log(n)))$ *and thus* $H(\tilde{E}) = \Omega(\log(\log(n)))$*.*

Proof of Theorem 7

Proof Outline.

For much of this proof, we follow intuitions and use terminology from the proof of Theorem 6. Consider a pair of variables X and Y such that X is a source and there is a path from X to Y. We aim to show that $\text{MEC}(Y|X) < \text{MEC}(X|Y)$. It is simple for us to show that $\text{MEC}(Y|X) = o(\log(\log(n)))$. To show $\text{MEC}(X|Y) = \Omega(\log(\log(n)))$, we will use an approach similar to Theorem 6 in that we will show existence of a state y' of Y such that $H(X|Y = y')$ is large. For showing there is a large $H(X|Y = y')$, we will show that there is a y' where its surplus is small and it receives many plateau balls. While we can factor Y as a function of X and small-entropy E (i.e., $Y = f(X, E)$), this is *not* a uniformly random function so we cannot simply apply the result of Theorem 6. In fact, a key difficulty is that this graph setting with more than two variables results in correlations between mappings. For example, in a graph such as the line graph (Figure 3-1a) with each node being a uniformly random deterministic function of its parents, one can show that conditioning on $f_Y(f_{X_2}(X = x)) = y$ almost doubles the probability that $f_Y(f_{X_2}(X = x')) = y$. Our new proof method must be able to withstand the dependencies that are introduced by this setting.

To provide some intuition, we give a very high-level overview for how to show existence of a large $H(X_{\text{src}}|Y = y')$ for two particular graphs, and we then expand to generalize these intuitions.

First, we consider the line graph. For simplicity, suppose that all nodes are deterministic functions of their parents (i.e., all $H(E_i) = 0$). Using the method from Theorem 6, we can see that there exists a large $H(X_{\text{src}}|X_2 = x_2')$. This is because we can show there is a bin of X_2 that has small surplus and receives $\Omega(\frac{\log(n)}{\log(\log(n))})$ balls. However, this analysis is actually loose in a sense. For a c where $0 < c < 1$, we can actually show there are n^c such bins that have small surplus and receive $\Omega(\frac{\log(n)}{\log(\log(n))})$ plateau balls. Now, when we look at how X_2 is mapped to Y, each of the bins of X_2 will "stick together." More formally, each bin of X_2 will have all of its mass mapped together to a uniformly random state of Y. This is because it is a deterministic function, but our proof will utilize a similar idea for when the function is not deterministic but the entropy is still small. It is then our hope that a good fraction of the bins with our desired properties (small surplus and many plateau balls) at X_2, will be mapped to a state of Y that does not have much surplus. In this sense, we have "heavy bins" and a non-negligible proportion of them are "surviving" from one node to the next because they aren't mapped to a bin with too much surplus. Through careful analysis, we are able to show that at least one such bin survives to the node of Y, and thus $H(X|Y = y')$ is large. This proof method would hold if we extend this line graph to any constant length.

Second, we consider the diamond graph (Figure 3-1b). Again, we assume all functions are deterministic for simplicity. Recall that for the line graph, our proof method was to show that there were many heavy bins at X_2, and then some heavy bins kept "sticking together" and "surviving" until we reached Y. This was because if two states of X were mapped to the same state of X_2, then they would "stick together" and would always be mapped to the same state for later nodes (e.g. if $f_{X_2}(x) = f_{X_2}(x')$ then $f_{X_3}(f_{X_2}(x)) = f_{X_3}(f_{X_2}(x'))$). However, this is very far from what is happening in diamond graph. In diamond graph, observe that $Y = f_Y(X_2, X_3)$. By definition of our graph, X_2 and X_3 are independent deterministic functions of X. Two states x and x' of X will be mapped to Y independently unless both $f_{X_2}(x) = f_{X_2}(x')$ and $f_{X_3}(x) = f_{X_3}(x')$. As these are independent, the probability of this happening is $\frac{1}{n^2}$. Thus, the expected number of pairs that are not mapped to Y independently of each other is $\binom{n}{2} \times \frac{1}{n^2} < \frac{1}{2}$. Accordingly, essentially all states of X will be mapped to a state of Y i.i.d. uniformly randomly. This enables us to more directly use the result and techniques of Theorem 6 and treat X and Y as a bivariate problem.

While we are able to show how both of these graphs will result in a large $H(X_{\text{src}}|Y = y')$, we do so very differently. For the line graph we show that there are bins with the properties we desire (small surplus and many plateau balls), that they will "stick together" as we move down through the graph, and at least one will "survive" to Y and thus $H(X_{\text{src}}|Y = y')$. For the diamond graph we show that when we get to Y, almost everything will be mapped independently randomly again, and that we can more directly use our bivariate techniques. There is a strong sense in which these two proof methods are opposites of each other (utilizing probability mass staying together throughout the graph as opposed to being independent at the end), yet we would like one unified approach for handling general graphs. To accomplish this, we introduce the *Random Function Graph Decomposition* to combine intuitions of these two settings into a characterization for all graphs.

Definition 10 (Random Function Graph Decomposition)**.** *For the Random Function Graph Decomposition we specify a source X and a node Y such that there is a path from X to Y. We ignore all nodes not along a path from X to Y. We define the remaining nodes as the set V_{decomp}. Then, we consider the nodes of V_{decomp} an arbitrary valid topological ordering and color each node as follows:*

- *If X is a parent of the node, or if the node has multiple parents and they are not all the same color, we create a new color for this node.*
- *Otherwise, all of the node's parent(s) have the same color, and this node will inherit said color.*

At a high-level, when a new color is created for a node, then everything is being mapped to the node almost-independently (similar to the intuition of the diamond graph). When a node inherits its color, there is a sense in which things "stick together" (similar to the intuition of the line graph). Let color-root(Y) be the earliest node in any topological ordering that has the same color as Y in the Random Function Graph Decomposition (it can be shown that color-root(Y) is unique). We aim to use the Random Function Graph Decomposition to show that everything will be mapped to color-root(Y) mostly independently. This will result in there being some bins with our desired properties (small surplus, many plateau balls) at color-root(Y). Then, we will show that at least one of these bins survives throughout all bins with the same color from color-root(Y) to Y, implying existence of a large $H(X_{\text{src}}|Y = y')$.

In particular, to show that balls are mapped to color-root(Y) mostly independently, we introduce the notion of *related* mass. More concretely, we define related$_1(x)$ mass as the amount of mass of balls that are ever mapped to the same state as x among any variable. We define related$_2(x)$ mass as the amount of mass of balls that are mapped to the same state as x for variables of at least two distinct colors in the Random Function Graph Decomposition. Inductively, we will show there are $\Omega(n)$ plateau balls such that related$_1(x) = O(\frac{1}{n})$ and related$_2(x) = O(\frac{1}{n^2})$. Moreover, we show that the quantity related$_2(x)$ upper-bounds mass that can have some dependence with x in how it is mapped to color-root(Y). With this upper-bound on dependence, we are able to use techniques of Theorem 6 to show there are many bins of color-root(Y) with many plateau balls and not much surplus. Finally, we show that, within the color of color-root(Y) and Y, at least one of these bins "survives" to Y and accordingly MEC($X_{\text{src}}|Y$) is large. *Complete Proof.* We must show that for a source X_{src} and a node Y such that there is a path from X_{src} to Y, MEC($X_{\text{src}}|Y$) > MEC($Y|X_{\text{src}}$).

Upper bounding MEC($Y|X_{\text{src}}$). It is simple to show that MEC($Y|X_{\text{src}}$) is small:

Claim 9. MEC($Y|X_{\text{src}}$) $\leq o(|V|\log(\log(n)))$

Proof. Y can be written as a function of X_{src} and the set of all E_i excluding $E_{X_{\text{src}}}$. As X_{src} is independent of these E_i, and their total entropy is $\sum_i H(E_i) = o(|V|\log(\log(n)))$, the claim holds since $|V| = \mathcal{O}(1)$. □

Bounding MEC($X_{\text{src}}|Y$) via $H(X_{\text{src}}|Y = y)$. Our method for lower-bounding MEC($X_{\text{src}}|Y$) is substantially more involved. As in Theorem 6, we will lower-bound it by MEC($X_{\text{src}}|Y$) $\geq \max_y H(X_{\text{src}}|Y = y)$ (see Theorem 6 for proof). Our proof aims to show there is a conditional entropy such that $\max_y H(X_{\text{src}}|Y = y) = \Omega(\log(\log(n)))$.

Showing existence of a near-uniform plateau. A key step in our approach, as in the proof of Theorem 6, is that we will find a subset of the support of X whose probabilities are multiplicative close to one another. In particular, we will find a subset of X_{src} where their probabilities are within a factor of $\log^{c_{\text{close}}}(n)$ of each other, where $0 < c_{\text{close}} < 1$. For our analysis, we require a value of c_{close} that is $\Omega(1)$ yet below some threshold. While there are multiple values of c_{close} that satisfy this condition, we will use $c_{\text{close}} = 1/4$. This set of states of X_{src} that are multiplicatively close to one another will be called the *plateau* of X_{src}. We use Lemma 10 proven in Theorem 6 to show how the $(\Omega(n), \Omega(\frac{1}{n\log(n)}))$-support assumptions implies a plateau of states of X:

Lemma 10 (Plateau existence)**.** *Suppose X has $(c_{support}n, \frac{1}{c_{lb}n\log(n)})$-support for constants $0 < c_{support} \leq 1$ and $c_{lb} \geq 1$. Additionally, assume n is sufficiently large such that $\frac{\log(2c_{lb}/c_{support})}{\log(\log(n))} \leq 1$. Then, there exists a subset $S \subseteq [n]$ of the support of X, such that the following three statements hold:*

1. $\frac{\max_{i\in S} P(X=i)}{\min_{i\in S} P(X=i)} \leq \log^{c_{close}}(n)$

2. $\min_{i\in S} P(X = i) \geq \frac{1}{c_{lb}n\log(n)}$

3. $|S| \geq \frac{c_{close}c_{support}n}{6}$, *for any* $0 < c_{close} < 1$.

Characterization as a balls-and-bins game. Our proof method of Theorem 6 characterizes a balls-and-bins game where states of $X \times E$ are balls and states of Y are bins. As we realized an entry $f(x, e)$ as a uniformly random state of Y, we characterized this as a ball (a state of $X \times E$) being assigned to a uniformly random bin (a state of Y). In the graph setting of this theorem, such a characterization is more complicated. Any node X_i is a uniformly random function of $\text{Pa}(X_i)$ and E_i. We define E^* to be the Cartesian product of all E_i other than E_X. Using this, we characterize balls as being states of $X \times E^*$. Note how any random variable in our SCM is a deterministic function of $X \times E^*$. In particular, it is the composition of (potentially many) f_i terms. For simplicity of notation, we let $f^*_T(x \times e^*)$ denote the value of a set of variables T for a particular state of $x \times e^*$. In the characterization of our balls-and-bins game, all balls with the same configuration of $\mathrm{Pa}(X_i)$ and E_i are mapped uniformly randomly together to a state of X_i. In other words, configurations are realized i.i.d. uniformly randomly. Using our notation, this means two balls (x_a, e^*_a) and (x_b, e^*_b) are mapped independently to variable X_i if any only if $f^*_{\mathrm{Pa}(X_i)\cup E_i}(x_a, e^*_a) \neq f^*_{\mathrm{Pa}(X_i)\cup E_i}(x_b, e^*_b)$.

Lower-bounding the most probable state of E_i and E^*. We focus first on *plateau balls*, which are balls corresponding to states of S (the set of plateau states of X) and the highest probability state of E^*. In particular, they are balls of the form $(X \in S, E^* = e^*_1)$ where e^*_1 is the most probable state of E^*. To show that these plateau balls have enough probability mass to be helpful, we first use Lemma 11 proven in Theorem 6 that implies all $\max_e H(E_i = e) \geq \frac{1}{\log^{c_{\text{close}}}(n)}$:

Lemma 11. *If $H(E) \leq c_{close} \log(\log(n))$ then $P(E = e_1) \geq \frac{1}{\log^{c_{close}}(n)}$*

This implies a lower-bound on the probability $P(E^* = e^*_1)$:

Lemma 18. *If all $\max_e P(E_i = e) \geq \frac{1}{\log^{c_{close}}(n)}$, then $P(E^* = e^*_1) \geq \frac{1}{\log^{c_{close}|V|}(n)}$.*

Proof. As E^* is the Cartesian product of $|V|-1$ variables E_i, it holds that $\max P(E^* = e^*) \geq (\min_i \max_e P(E_i = e))^{|V|-1} \geq (\frac{1}{\log^{c_{\text{close}}}(n)})^{|V|-1} \geq \frac{1}{\log^{c_{\text{close}}|V|}(n)}$. □

Introducing surplus. In Theorem 6, we prove how there exists a bin that receives a large amount of mass that helps the bin have large conditional entropy (such helpful mass includes the plateau balls), and not much mass that hurts the conditional entropy making it small. To formalize this hurtful mass, we introduced the *surplus* quantity described in Definition 6. This surplus is a way of quantifying the probability mass received by a state of Y that is hurtful towards making the conditional entropy large. The proof of Theorem 6 achieves a lower-bound for $\max_y H(X|Y = y)$ by proving existence of a state y' of Y where y' receives many plateau balls and the surplus is small. Likewise, we will also prove existence of such a state of Y with many plateau balls and small surplus, in the graph setting. We formalize the notion of surplus as follows:

Definition 11 (Surplus, $\mathcal{T} = \frac{120}{n\log(n)}$)**.** *We define the surplus of a state i of Y as $z_i = \sum_{j \notin S} \max(0, P(X = j, Y = i) - \frac{120}{n\log(n)})$.*

Introducing the Random Function Graph Decomposition. In Section 3.9.1, we introduced intuitions from considering the diamond graph in Figure 3-1b and the line graph in Figure 3-1a. In the proof outline for a diamond graph, we utilized the intuition that almost all balls were independently assigned to Y. This enables us to use techniques from Theorem 6, as almost all balls were independently assigned to a uniformly random state of Y, closely mirroring the setting of Theorem 6. In the proof outline for line graph, we used techniques of Theorem 6 to show that there would be many bins that received many plateau balls and small surplus. Then, we showed that at least one of these bins would mostly "survive" and remain in-tact to Y. While our intuitions for both of these graphs enabled us to show existence of a large $H(X_{\text{src}}|Y=y)$, but they did so with near-opposite methods. Our intuition for the diamond graph exploits independence (everything is assigned almost independently to Y), while our intuition for the line graph exploits dependence (some bins with our desired properties "survive" from the second node onwards). We introduce the *Random Function Graph Decomposition* to combine intuitions of these two graphs into a characterization for all graphs:

Definition 10 (Random Function Graph Decomposition)**.** *For the Random Function Graph Decomposition we specify a source X and a node Y such that there is a path from X to Y. We ignore all nodes not along a path from X to Y. We define the remaining nodes as the set V_{decomp}. Then, we consider the nodes of V_{decomp} an arbitrary valid topological ordering and color each node as follows:*

- *If X is a parent of the node, or if the node has multiple parents and they are not all the same color, we create a new color for this node.*
- *Otherwise, all of the node's parent(s) have the same color, and this node will inherit said color.*

At a high-level, when a new color is created for a node, then we will see that plateau balls are being mapped to the node almost-independently (similar to the intuition of the diamond graph). When a node inherits its color, there is a sense in which things “stick together” (similar to the intuition of the line graph). For some node $X_i \in V_{\text{decomp}}$, we define color$(X_i)$ to be the node's color in the Random Function Graph Decomposition. Under a fixed topological ordering, let color-root(Y) be the earliest node that has the same color as Y in the Random Function Graph Decomposition (it can be shown that color-root(Y) is unique). We aim to use the Random Function Graph Decomposition to show that everything will be mapped to color-root(Y) mostly independently. This will result in there being some bins with our desired properties (small surplus, many plateau balls) at color-root(Y). Then, we will show that at least one of these bins survives throughout all bins with the same color from color-root(Y) to Y, implying existence of a large $H(X_{\text{src}}|Y = y')$.

Introducing related mass. To show how plateau balls are mapped to color-root(Y) mostly independently, we introduce the concept of *related* mass. Related mass introduces a measure of how much mass has come into contact with a particular plateau ball:

Definition 12 (Related mass)**.** *We define related mass of two types as follows.*

- *For a plateau state x of X, we define related$_1(x)$ mass as the amount of mass of balls from non-plateau states of X that are ever mapped to the same state as the plateau ball of x among any variable in the Random Function Graph Decomposition. In other words, x', e^* contributes to related$_1(x)$ if it satisfies the following for some X_i: x' together with some realization e^* contributes to the same bin of X_i that x is mapped to together with e^*_1. More formally, we define $\mathcal{B}_1(x)$ as the set of balls whose mass counts towards related$_1(x)$, where $\mathcal{B}_1(x) = \{x' \in X \backslash S, e^* \in E^* | \exists X_i \in V_{decomp} s.t. f^*_{X_i}(x', e^*) = f^*_{X_i}(x, e^*_1)\}$. Accordingly, related$_1(x) = \sum_{x',e^* \in \mathcal{B}_1(x)} P(X = x') \cdot P(E^* = e^*)$.*
- *For a plateau state x of X, we define related$_2(x)$ mass as the amount of mass of balls from non-plateau states of X that are ever mapped to the same state as the plateau ball of x among variables of at least two distinct colors in the Random Function Graph Decomposition. In other words, x', e^* contributes to related$_2(x)$ if it satisfies the following for some X_i, X_j with distinct colors:*

x' together with some realization e^* contributes to the same bin of X_i that x is mapped to together with e^*_1; same holds for X_j. More formally, we define $\mathcal{B}_2(x)$ as the set of balls whose mass counts towards related$_2(x)$, where $\mathcal{B}_2(x) = \{x' \in X \backslash S, e^* \in E^* | \exists X_i, X_j \in V_{decomp} s.t. f^*_{X_i}(x', e^*) = f^*_{X_i}(x, e^*_1), f^*_{X_j}(x', e^*) = f^*_{X_j}(x, e^*_1), color(X_i) \neq color(X_j)\}$. Accordingly, related$_2(x) = \sum_{x', e^* \in \mathcal{B}_2(x)} P(X = x') \cdot P(E^* = e^*)$.

Now, we will consider an arbitrary topological ordering of V_{decomp}. In this ordering, we define order(X_i) for $X_i \in V_{\text{decomp}}$ as the index of X_i in the topological ordering. We introduce a modification of related$_1(x)$ where related$_1^{\text{order}(X_i)}(x)$ only considers nodes of V_{decomp} that are strictly earlier in the topological ordering than X_i. We define related$_2^{\text{order}(X_i)}(x)$ analogously. It is our goal to show that there are many plateau states $x \in S$ such that related$_2^{\text{order(color-root}(Y))}(x)$ is small. This will enable us to show how there are many plateau balls that are mapped to color-root(Y) independently of almost all other mass.

Upper-bounding related mass. To show independence in how some plateau balls are mapped to color-root(Y), we bound related$_2^{\text{order(color-root}(Y))}(x)$ for some plateau states $x \in S$.

To show this, we will process nodes in the topological ordering. After processing the first i nodes, we will argue that there is a large set S^i_{indep} with upper-bounds on all related$_1^i(x)$ and related$_2^i(x)$.

Lemma 19. *With high probability, after processing the first i nodes in the topological ordering, there exists a set S^i_{indep} such that $|S^i_{indep}| = \frac{|S|}{6^i}$, all $x \in S^i_{indep}$ satisfy related$_1^i(x) \leq \frac{6i}{n}$ and related$_2^i(x) \leq \frac{18 \times i \times (i-1)}{n^2}$, and all $x, x' \in S^i_{indep}$ satisfy $f^*_{X_j}(x, e^*_1) \neq f^*_{X_j}(x', e^*_1)$ for all $1 \leq j \leq i$.*

Proof. We begin with the following claims.

Claim 10. Lemma 19 holds for $i = 1$.

Proof. To find a subset of S to be S^i_{indep}, we will choose any arbitrary subset of size $\frac{S}{6}$. By definition of related$_1$ and related$_2$, all plateau balls are different states of X_{src}

so $\text{related}_1^1(x) = \text{related}_2^1(x) = 0$ for every $x \in S$, and $f^*_{X_{\text{src}}}(x, e_1^*) \neq f^*_{X_{\text{src}}}(x', e_1^*)$ for all $x, x' \in S$. □

Claim 11. Lemma 19 holds for i if it holds for all $j < i$.

Proof. First, we realize f_{X_i} for all cells other than those corresponding to configurations of $\text{Pa}(X_i) \cup E_{X_i}$ that contain an element of S^{i-1}_{indep}. Now, we consider the process of realizing the entries of f_{X_i} corresponding to elements of S^{i-1}_{indep} in an arbitrary order. We define a random variable for every element of S^{i-1}_{indep}. For the j-th element, we define $\mathcal{S}_j$ as follows:

- If the element is mapped to a bin that another element of S^{i-1}_{indep} has been mapped to, then $\mathcal{S}_j = -1$.
- Otherwise, if the element $x \in S^{i-1}_{\text{indep}}$ is mapped to a bin that contains total mass at least $\frac{6}{n}$, or total mass from $\mathcal{B}_1^{i-1}(x)$ of at least $\frac{6\text{related}_1^{i-1}(x)}{n}$, then $\mathcal{S}_j = 0$.
- Else, then $\mathcal{S}_j = 1$.

The intuition behind $\mathcal{S}_j$ is that we will count an element of $x \in S^i_{\text{indep}}$ as being eligible for S^i_{indep} if it lands in a bin with no other value of S^{i-1}_{indep}, and if it lands in a bin that will not increase $\text{related}_1^i(x)$ or $\text{related}_2^i(x)$ by too much.

Claim 12. Consider the set comprised of each element $x \in S^{i-1}_{\text{indep}}$ that satisfies the following. Suppose x is assigned to a bin such that before x is mapped to the bin, the bin has total mass at most $\frac{6}{n}$ and total mass intersecting from $\mathcal{B}^{i-1}(x)$ of at most $\frac{6\text{related}_1^{i-1}(x)}{n}$. Moreover, suppose x is the only element of S^{i-1}_{indep} that is ever assigned to this bin. Then, this set of all such x would meet the desired properties required of S^i_{indep}.

Proof. The increase of the quantity $\text{related}_1^i(x)$ is bounded by the amount of other mass in the bin that x is assigned to. Accordingly, $\text{related}_1^i(x) \leq \text{related}_1^{i-1}(x) + \frac{6}{n} \leq \frac{6\times(i-1)}{n} + \frac{6}{n} = \frac{6i}{n}$. The increase of the quantity $\text{related}_2^i(x)$ is bounded by the amount of mass from $\mathcal{B}_1^{i-1}(x)$ in the bin x is assigned to. Accordingly, $\text{related}_2^i(x) \leq \text{related}_2^{i-1}(x) + \frac{6\text{related}_1^{i-1}(x)}{n} \leq \frac{18\times(i-1)\times(i-2)}{n^2} + \frac{36(i-1)}{n^2} = \frac{18\times i\times(i-1)}{n^2}$ □

Moreover, we claim that $\sum \mathcal{S}_i$ serves as a lower bound for the set of elements eligible for S^i_{indep} referenced in Claim 12.

Claim 13. The number of elements of S^{i-1}_{indep} that are eligible for S^i_{indep} by satisfying Claim 12 is at least $\sum_j \mathcal{S}_j$.

Proof. For each bin, consider the sum of $\mathcal{S}_j$ for variables corresponding to elements of S^{i-1}_{indep} that were assigned to the bin (if any). If the sum is nonpositive, then we trivially claim the set of elements meeting the criteria in this bin is at least the sum, as there will be at least 0 such elements. Otherwise, the sum must be 1, This implies there is exactly one element of S^{i-1}_{indep} assigned to the bin, and that it met the criteria when it was assigned, because its corresponding $\mathcal{S}_j = 1$. Moreover, as no other elements could have been assigned to the bin later, it still meets the criteria. Combining both cases, we see that the sum of $\mathcal{S}_j$ for each bin is a lower-bound for the number of elements satisfying the criteria in said bin, and thus globally the sum of all $\mathcal{S}_j$ is a lower-bound for how many elements meet the criteria in total. □

We aim to now use the sum of $\mathcal{S}_j$ as a lower-bound for the size of the set of elements meeting the criteria. To do so, we will first lower-bound $E[\mathcal{S}_j]$.

Claim 14. Regardless of the realization of any previous randomness, $E[\mathcal{S}_j] \geq \frac{1}{3}$.

Proof. $\mathcal{S}_j$ is equal to -1 only if it is assigned to a bin with another element of S^{i-1}_{indep}. The number of such bins is upper-bounded by $|S^{i-1}_{\text{indep}}| \leq |S^1_{\text{indep}}| \leq \frac{n}{6}$. Otherwise, $\mathcal{S}_j$ is equal to 0 only if the bin had mass at least $\frac{n}{6}$ or it has mass from the corresponding $\mathcal{B}^{i-1}_1(x)$ of at least $\frac{6\text{related}^{i-1}_1(x)}{n}$. There can only be at most $\frac{n}{6}$ bins satisfying the former, and at most $\frac{n}{6}$ bins satisfying the latter. Accordingly, there are at least $n - 3 \times \frac{n}{6} = \frac{n}{2}$ where if the corresponding element is assigned to it, then $\mathcal{S}_j = 1$. Hence $E[\mathcal{S}_j] \geq \frac{1}{2} - \frac{1}{6} = \frac{1}{3}$. □

As we need a set S^i_{indep} with cardinality $|S^i_{\text{indep}}| = \frac{|S^{i-1}_{\text{indep}}|}{6}$, we show the following:

Claim 15. $\sum_j \mathcal{S}_j \geq \frac{|S^{i-1}_{\text{indep}}|}{6}$ with high probability.

Proof. We will modify the variables to make a martingale and then utilize Azuma's inequality. We define $\mathcal{S}'_j = \mathcal{S}'_{j-1} + \mathcal{S}_j - E[(\mathcal{S}_j | \mathcal{S}_1, \ldots, \mathcal{S}_{j-1})]$. Accordingly, the sequence of $\mathcal{S}'$ is a martingale of length $|S^{i-1}_{\text{indep}}|$ where $|\mathcal{S}'_{j-1} - \mathcal{S}'_j| \leq 1$. Thus, we can use Azuma's inequality to show $P(|\mathcal{S}'_{|S^{i-1}_{\text{indep}}|} - \mathcal{S}'_1| \geq \frac{|S^{i-1}_{\text{indep}}|}{6}) \leq 2e^{\frac{-|S^{i-1}_{\text{indep}}|}{72}} = 2e^{\frac{-|S|}{726^{i-1}}} = 2e^{-\Omega(n)}$. By definition, $\sum_j \mathcal{S}_j = \sum_j \mathcal{S}'_j + \sum_j E[S_j]$. By our result with Azuma's inequality, we then claim that with high probability it holds that $\sum_j \mathcal{S}_j \geq -\frac{|S^{i-1}_{\text{indep}}|}{6} + \frac{|S^{i-1}_{\text{indep}}|}{3} = \frac{|S^{i-1}_{\text{indep}}|}{6}$. □

Combining Claim 12 and Claim 15, we have now shown that there exists a valid set S^{i}_{indep} of size $|S^{i}_{\text{indep}}| = \frac{|S^{i-1}_{\text{indep}}|}{6}$, completing the proof of Claim 11. □

By induction Lemma 19 holds.

□

Corollary 9. *There exists a subset of plateau states $S_{indep} \subseteq S$ such that $|S_{indep}| \geq \frac{|S|}{6^{|V|}} = \Omega(n)$ and every $x \in S_{indep}$ satisfies $\text{related}_2^{order(color\text{-}root(Y))}(x) \leq \frac{18 \times |V| \times (|V|-1)}{n^2}$. Moreover, for all pairs $x, x' \in S_{indep}$ it holds that they never share a state, meaning $f^*_{X_i}(x) \neq f^*_{X_i}(x')$ for all $X_i \in V_{decomp}$.*

Proof. One such set is simply $|S^{|V|}_{\text{indep}}|$ as shown in Lemma 19. □

Characterizing balls. Recall that each variable assigns balls with the same configuration of its parents and exogenous variable together. We aim to show a similar result at color-root(Y). To do so, we will characterize the balls within configurations into types:

Definition 7 (Ball characterizations)**.** *We characterize three types of balls:*

1. *Dense balls. Consider a set L of states of X, where a state of X is in L if $P(X = x) \geq \frac{1}{\log^3(n)}$. Dense balls are all balls of the form $(x \in L, e \in E)$. We call these dense balls, because the low-entropy of E will prevent the collective mass of these balls from "expanding" well.*

2. *Large balls. For all balls of the form $(x \in X \backslash (S \cup L), e \in E)$ where the ball has mass $\geq \frac{\mathcal{T}}{2}$.*

3. *Small balls. For all balls of the form $(x \in X \backslash (S \cup L), e \in E)$ where the ball has mass $< \frac{\mathcal{T}}{2}$.*

We use $\mathcal{T} = \frac{120|V|}{n\log(n)}$.

Now, we will show that for every variable X_i there are many bins without too much surplus, such that the plateau configurations have many bins that they may be assigned to that will help us obtain a bin with small surplus and many plateau configurations.

Definition 13 (Configuration and ball characterizations)**.** *We characterize three types of configurations/balls:*

1. *Large* configurations. *For all configurations of the form $(Pa(X_i) \cup E_i)$ where the configuration has balls of total mass $\geq \frac{\mathcal{T}}{2}$.*

2. *Dense* ball. *Consider a set L of states of X_{src}, where a state of X_{src} is in L if $P(X_{src} = x) \geq \frac{1}{\log^3(n)}$. Dense balls are all balls of the form $(x \in L, e \in E^*)$. We call these dense balls, because the low-entropy of E will prevent the collective mass of these balls from being distributed well throughout.*

3. *Small* ball. *For all balls of the form $(x \in X_{src} \backslash (S \cup L), e \in E^*)$ where the ball has mass $< \frac{\mathcal{T}}{2}$.*

Bounding dense ball surplus. Recall the following used in Theorem 6 to bound contributions from dense balls:

Lemma 12 (Limited expansion)**.** *Suppose Y can be written as a function $f(X, E)$ and $X \perp\!\!\!\perp E$. Consider any subset R of the support of X. For any subset T of the support of Y that satisfies $\forall t \in T : P(X \in R, Y = t) > \delta$, the cardinality of T is upper bounded as $|T| \leq \frac{H(E)+\log(|R|)+2}{\delta \log(\frac{1}{\delta})}$.*

We use the following corollary:

Corollary 5. *There exist no subset $|T| = n/4$ such that $\forall t \in T : P(X \in L, Y = t) \geq \frac{1}{n\log(\log(n))\log^{2c_{close}}(n)}$*

These imply the following for our graph setting. While this may seems strictly weaker than Corollary 5, we will utilize that the event of a bin having too much mass from dense balls is now independent from how large configurations are mapped.

Corollary 10. *Let $\mathcal{C}_{large}$ denote the set of large configurations of $Pa(X_i) \cup E_i$ as defined in Definition 13. Let C be a random variable denoting the configuration of the corresponding ball of $Pa(X_i) \cup E^*$. We claim that dense balls in configurations other than $\mathcal{C}_{large}$ are not distributed well throughout X_i. In particular, there exists no subset $|T| = n/4$ such that $\forall t \in T : P(X \in L, Y = t, C \notin \mathcal{C}_{large}) \geq \frac{1}{n \log(\log(n)) \log^{2c_{close}}(n)}$.*

Proof. Note that the results of Lemma 12 and Corollary 5 still hold in this setting as any $X_i \in V_{\text{decomp}}$ can be written as a function of X_{src} and $\cup_j E_j$. This corollary trivially follows from Corollary 5, as it is strictly weaker in that we add a restriction that $C \notin \mathcal{C}_{\text{large}}$. Any set T that contradicts Corollary 10 would immediately contradict Corollary 5. □

Bounding large configuration surplus. To bound contribution to surplus by large configurations, we bound the number of bins that receive any mass from large configurations. Recall Lemma 13 from the proof of Theorem 6:

Lemma 13 (Avoided big)**.** *Given a balls-and-bins game with $c \cdot n \ln(n)$ balls mapped uniformly randomly to n bins, at least $\frac{n^{1-c}}{2}$ bins will receive no balls with high probability if c is a constant such that $0 < c \leq \frac{1}{3}$.*

Accordingly, we can use the following:

Corollary 11. *As there are at most $\frac{1}{\mathcal{T}/2} \leq \frac{n \log(n)}{60|V|} \leq \frac{1}{40|V|} \cdot n \ln(n)$ large balls, with high probability there are at least $\frac{n^{1-\frac{1}{40|V|}}}{2}$ bins that receive no large balls.*

Bounding small ball surplus. Here we will bound the surplus from small balls. Note that, while this proof is not short, it is using the same ideas as the corresponding section in the proof of Theorem 6. However, there are some subtle differences that necessitate a separate proof for the graph setting. We use identical text from the proof of Theorem 6 when applicable.

For the small balls, we will also show that they cannot contribute too much surplus to too many states of any X_i. We will notably use that all small balls correspond to a state of X_{src} where $P(X_{\text{src}} = x) \leq \frac{1}{\log^3(n)}$. We will utilize this to show that most small balls are assigned to a state of X_i that has not yet received $> \frac{\mathcal{T}}{2}$ mass from its corresponding state of X_{src}, and accordingly would not increase the surplus. To accomplish this, we define a surplus quantity that only takes into account small balls:

Definition 14 (Small ball surplus). *We define the small ball surplus of a state y of Y as*

$$z_j^{small} = \sum_{x_{src} \notin (S \cup L)} \max \Bigg(0, -\mathcal{T} +$$

$$\sum_{\substack{e^*: \\ (X = x_{src}, E^* = e. C \notin \mathcal{C}_{large})}} P(X_{src} = x, E^* = e, X_i = j) \Bigg) .$$

With this notion of surplus constrained to small balls, we show the following:

Lemma 20 (Small ball limited surplus). *With high probability, there are at most $\frac{n}{4}$ values of i, i.e., number of bins, where $z_i^{small} \geq \frac{1}{n \log(\log(n)) \log^{2c_{close}}(n)}$.*

Proof. We will consider configurations $C \notin \mathcal{C}_{\text{large}}$ in an arbitrary order, and within each configuration consider balls in an arbitrary order. Let $x_i(c)$ be the corresponding state of X_i for the c-th configuration, let $x_{\text{src}_c}(t)$ be the corresponding state of X_{src} for the t-th ball in the c-th configuration. $e_{\text{src}_c}(t)$ be the corresponding state of E^* for the t-th ball in the c-th configuration, and $w_{c,\text{ball}}(t)$ be the t-th ball's probability mass in the c-th configuration (i.e., $P(X_{\text{src}} = x_{\text{src}_c}(t), E = e_{\text{src}_c}(t))$). Moroever, we define $w_{\text{config}}(c)$ as the weight of all such balls with configuration c. Recall that for all small balls it must hold that $x_{\text{src}_c}(t) \notin L$ and thus $P(X_{\text{src}} = x_{\text{src}_c}(t)) < \frac{1}{\log^3(n)}$. We define the total small ball surplus as $Z^{\text{small}} = \sum_{j \in X_i} z_j^{\text{small}}$. Now, we will consider all non-large configurations in an arbitrary order and realize their corresponding entry of f to map them to a state of X_i. Initially, we have not realized the entry of f for any balls and thus all $z_j^{\text{small}} = 0$ and $Z^{\text{small}} = 0$. As we map configurations to states of X_i, we define $\Delta(c)$ as the increase of Z^{small} after mapping the c-th configuration to a state of X_i. By definition, $\sum_c \Delta(c)$ is equal to Z^{small} after all values of f have been completely realized.

Our primary intuition is that we will show for many small balls it holds that they have zero contribution towards their configuration's quantity $\Delta(c)$. As f is realized for each configuration, let $w_{X_i}(x_i, x_{\text{src}})$ denote the total mass of balls assigned to state x_i of X_i so far from state x_{src} of X_{src}, i.e., $w_{X_i}(x'_i, x'_{\text{src}}) := \sum\limits_{c'<c, t: x_i(c)=x'_i} w_{c,\text{ball}}(t')$. Note that this quantity is shared among all configurations.

Claim 16. Regardless of the realizations of all $\Delta(c')$ for $c' < c$, it holds that $\Delta(c)$ is a random variable with values in range $[0, w_{\text{config}}(c)]$ and $E[\Delta(c)] \leq \frac{w_{\text{config}}(c)}{\log^2(n)}$.

Proof. Let us define $\Delta_t(c)$ as the contribution of the t-th ball to $\Delta(c)$. By definition, $\sum_t \Delta_t(c) = \Delta(c)$. We aim to show $E[\Delta_t(c)] \leq \frac{w_{c,\text{ball}}(t)}{\log^2(n)}$. This would immediately imply the desired bound on $E[\Delta(t)]$ by linearity of expectation.

The only conditions under which $\Delta_t(c)$ takes a non-negative value (which is upper-bounded by $w_{c,\text{ball}}(t)$), is when $w_{X_i}(x_i(c), x_{\text{src}_c}(t)) > \frac{\mathcal{T}}{2}$ before the entry of f for the c-th configuration is realized (other. Recall that $P(x_{\text{src}_c}(t)) \leq \frac{1}{\log^3(n)}$. Accordingly, the number of states x'_i of X_i where $w^c_{X_i}(x'_i, x_{\text{src}_c}(t)) > \frac{\mathcal{T}}{2}$ is upper-bounded by $\frac{P(X_{\text{src}}=x_{\text{src}}))}{\mathcal{T}/2} \leq \frac{1/\log^3(n)}{60|V|/(n\log(n))} = \frac{n\log(n)}{60|V|\log^3(n)} \leq \frac{n}{\log^2(n)}$. This is due to the fact that balls partition the total mass of $P(X_{\text{src}} = x_{\text{src}}(t))$ since we have $P(X_{\text{src}} = x_{\text{src}}(t)) = \sum_e P(X = x_{\text{src}}(t), E = e)$. This implies that the probability that the t-th ball of configuration c will be mapped to a state x'_i of X_i such that $w^c_{X_i}(x'_i, x_{\text{src}_c}(t))$ already exceeds the threshold of $\mathcal{T}/2$ (in other words where we will have $\Delta_t(c) > 0$) is upper-bounded by $\frac{n/\log^2(n)}{n} = \frac{1}{\log^2(n)}$ due to the fact that the function f is realized uniformly randomly. Accordingly, $E[\Delta_t(c)] \leq \frac{w_{c,\text{ball}}(t)}{\log^2(n)}$ and thus $E[\Delta(c)] \leq \frac{w_{\text{config}}(c)}{\log^2(n)}$. □

This enables us to upper-bound the sum of $\Delta(t)$:

Claim 17. $\sum_c \Delta(c) \leq \frac{1}{4\log(n)}$ with high probability.

Proof. We will transform $\Delta(c)$ into a martingale. In particular, we define $\Delta'(c) = \Delta'(c-1) + \Delta(c) - E[\Delta(c)|\Delta(1), \ldots, \Delta(c-1)]$. We define $\Delta'(0) = 0$, and note that $\Delta'(c)$ is a martingale. By Azuma's inequality, we show $|\sum_c \Delta'(c)| \leq \frac{1}{8\log(n)}$ with high probability:

$$P[|\Delta(c)| > \varepsilon] < 2e^{-\frac{\varepsilon^2}{2\sum c_i^2}}$$

$$\leq 2e^{-\frac{\left(\frac{1}{8\log(n)}\right)^2}{2(\max_i c_i)\cdot\sum c_i}}$$

$$
\begin{aligned}
&\leq 2e^{-\frac{\left(\frac{1}{8\log(n)}\right)^2}{2\times\mathcal{T}/2\cdot 1}} \\
&= 2e^{\frac{-n\log(n)}{120\times 8\times|V|\times\log(n)}}
\end{aligned}
$$

Accordingly, by definition of $\Delta'(c)$ this implies $|(\sum_c \Delta(c)) - \sum_c E[\Delta(c)|\Delta(1), \ldots, \Delta(c-1)]| \leq \frac{1}{8\log(n)}$. By Claim 16 we know all $E[\Delta(c)|\Delta(1), \ldots, \Delta(c-1)] \leq \frac{w_{\text{config}(c)}}{\log^2(n)}$ and accordingly, $\sum_c E[\Delta(c)|\Delta(1), \ldots, \Delta(c-1)] \leq \frac{1}{\log^2(n)}$. Together, these imply $\sum_c \Delta(c) \leq \frac{1}{8\log(n)} + \frac{1}{\log^2(n)}$ with high probability, and for sufficiently large n it holds that $\frac{1}{\log^2(n)} \leq \frac{1}{8\log(n)}$. Thus, our high-probability on $|\Delta'(c)|$ implies that $\sum_t \Delta(t) \leq \frac{1}{4\log(n)}$ with high probability. □

Finally, we conclude that our upper-bound on $\sum_t \Delta(t)$ implies an upper-bound on the number of states of Y with non-negligible small ball support:

Claim 18. If $\sum_t \Delta(t) \leq \frac{1}{4\log(n)}$, then there are at most $\frac{n}{4}$ bins where $z_i^{\text{small}} \geq \frac{1}{n\log(\log(n))\log^{2c_{\text{close}}}(n)}$.

Proof. By definition, $Z^{\text{small}} = \sum_t \Delta(t) \leq \frac{1}{4\log(n)}$. Given this upper-bound for total small ball surplus, we can immediately upper-bound the number of states of X_i with small ball surplus greater than $\frac{1}{n\log(\log(n))\log^{2c_{\text{close}}}(n)}$ by the quantity $\frac{1/(4\log(n))}{1/(n\cdot\log(\log(n))\cdot\log^{2c_{\text{close}}}(n))} \leq \frac{n\cdot\log(\log(n))\cdot\log^{1/2}(n)}{4\log(n)} \leq \frac{n}{4}$. We obtain this by using $c_{\text{close}} = \frac{1}{4}$ and for sufficiently large n such that $\log(\log(n)) \leq \log^{1/2}(n)$.

□

This concludes the proof of the lemma. □

Concluding many bins with small surplus. Now, we combine all these intuitions to show there are many bins that have a small amount of surplus. We have shown that, with high probability, the are at most $n/4$ bins with non-negligible mass from dense balls by Corollary 5, and at most $n/4$ bins with non-negligible surplus from small balls from non-large configurations Lemma 20. Combining these sets, there are at most $n/2$ bins with non-negligible mass from dense balls or surplus from small balls. By Corollary 11, with high probability at least $\frac{n^{1-\frac{1}{40|V|}}}{2}$ bins will receive no large configurations mapped to it. Our goal is to show the intersection of the sets is large, so there are many bins that have small surplus. We use Lemma 15 proven in Theorem 6:

Lemma 15. *Let there be two sets $A, B \subseteq [n]$, where $|A| \geq \frac{n}{2}$ and A and B are both independently uniformly random subsets of size $|A|$ and $|B|$, respectively. It holds that $P(|A \cap B| \geq \frac{|B|}{4}) \geq 1 - 2e^{\frac{-|B|}{8}}$.*

Corollary 12. *With high probability, there are at least $\frac{n^{1-\frac{1}{40|V|}}}{8}$ bins with surplus $z_y \leq \frac{2}{n\log(\log(n))\log^{2c_{close}}(n)}$.*

Proof. We have defined three types of balls, and have proven results that show how there are many bins with negligible bad contribution for each type of ball. Now, we combine these with Lemma 15 to show there are many bins where there is not much bad contribution in total. By Corollary 5 there are at most $n/4$ bins with more than $\frac{1}{n\log(\log(n))\log^{2c_{\text{close}}(n)}(n)}$ mass from dense balls. By Lemma 20, there are at most $n/4$ bins with small ball surplus more than $\frac{1}{n\log(\log(n))\log^{2c_{\text{close}}(n)}(n)}$. Let A be the set of bins with at most $\frac{1}{n\log(\log(n))\log^{2c_{\text{close}}(n)}(n)}$ mass from dense balls and at most $\frac{1}{n\log(\log(n))\log^{2c_{\text{close}}(n)}(n)}$ small ball surplus. By combining Corollary 5 and Lemma 20 we know $|A| \geq \frac{n}{2}$ with high probability. Let B be the set of bins that receive no large configurations. By Corollary 11, it holds that $|B| \geq \frac{n^{1-\frac{1}{40|V|}}}{2}$ with high probability. A and B are independent, as A is undetermined by the mapping of large configurations. By Lemma 15, it holds that $|A \cap B| \geq \frac{n^{1-\frac{1}{40|V|}}}{8}$ with failure probability at most $2e^{\frac{-n^{1-\frac{1}{40|V|}}}{16}}$. Moreover, all such bins will have total surplus at most $\frac{2}{n\log(\log(n))\log^{2c_{\text{close}}}(n)}$, because they receive no large configurations and total surplus is then upper-bounded by the sum of small ball surplus and total mass from dense balls. □

Existence of many desirable bins at color-root(Y).

We have shown that at each node X_i there are many bins without much surplus. If we restrict this calculation of surplus to not include mass from plateau configurations in S_{indep} for color-root(Y), then the set of bins that do not have much surplus is independent of the assignment of such configurations with plateau balls not having much related mass. Consider the set $S_{\text{indep}}^{\text{order(color-root}(Y))}$. By Corollary 9, we know

$|S_{\text{indep}}^{\text{order(color-root}(Y))}| \geq \frac{|S|}{6|V|}$, no two corresponding plateau balls ever share a state before color-root(Y), and all $\text{related}_2^{\text{order(color-root}(Y))}(x) \leq \frac{18\times|V|\times(|V|-1)}{n^2}$. Recall that color-root$(Y)$ by its definition must create a new color, and thus either have X_{src} as a parent, or have at least two distinct colors in its parent set. Given these properties, we know that each plateau ball in this set has at most $\text{related}_2^{\text{order(color-root}(Y))}(x) \leq \frac{18\times|V|\times(|V|-1)}{n^2}$ mass in its configuration for color-root(Y), and all elements in the set will be in different configurations. We aim to show that there are many bins with small surplus that receive many of these configurations corresponding to the plateau balls in the set. To do so, we use the following results shown in Theorem 6. First, we use negative association:

Claim 4. Indicator variables for if a bin receives some threshold of balls in a i.i.d. uniformly random balls-and-bins game are NA.

Second, we lower bound the probability of a bin researching a certain threshold:

Claim 5. Suppose cn balls $(c \leq 1)$ are thrown i.i.d. uniformly randomly into n bins. The probability that a particular bin receives at least $k = \frac{d\log(n)}{\log(\log(n))}$ balls is at least $\frac{1}{en^d}$ given that $\frac{d}{c} \leq \log(\log(n))$.

Corollary 13. *With high probability, there are* $\frac{n^{1-\frac{1}{20|V|}}}{16e}$ *bins with* $z_j \leq \frac{2}{n\log(\log(n))\log^{2c_{close}}(n)}$ *and at least* $\frac{\log(n)}{40|V|\log(\log(n))}$ *configurations mapped from states of* $S_{\textit{indep}}^{\textit{order(color-root}(Y))}$.

Proof. Consider the indicator variable $\mathcal{B}_i$ if a bin met the threshold of plateau configurations. We show that with high probability $\sum_i \mathcal{B}_i$ is large enough, considering just the bins with small surplus. By Corollary 12 we know there are at least $\frac{n^{1-\frac{1}{40|V|}}}{8}$ such bins with high probability. Now let us focus on just the configurations corresponding to $S_{\text{indep}}^{\text{order(color-root}(Y))}$ that we know has cardinality $\Omega(n)$. By Claim 5, the probability of a bin receiving at least $\frac{\log(n)}{40|V|\log(\log(n))}$ such configurations is at least $\frac{1}{en^{\frac{1}{40|V|}}}$. Accordingly, it holds that $\sum_i E[\mathcal{B}_i] \geq \frac{n^{1-\frac{1}{20|V|}}}{8e}$.

By Hoeffding's inequality, we show $|\sum_i \mathcal{B}_\rangle - \sum_i E[\mathcal{B}_\rangle]| \leq \frac{n^{1-\frac{1}{20|V|}}}{16e}$ with high probability:

$$\begin{aligned} P[|S_n - E_n| > t] &< 2e^{-\frac{2t^2}{\sum c_i^2}} \\ &\leq 2e^{-\frac{2n^{2-\frac{1}{10|V|}}}{16^2 e^2 \frac{n^{1-\frac{1}{40|V|}}}{8}}} \\ &\leq 2e^{-\frac{n^{1-\frac{3}{40|V|}}}{16e^2}}. \end{aligned}$$

Therefore, $\sum_i \mathcal{B}_i \geq \frac{n^{1-\frac{1}{20|V|}}}{16e}$ with high probability. □

Survival of desirable bins to Y. Now we aim to show that of the bins that received many plateau configurations and had small surplus, that enough will "survive" and keep these properties as we process nodes within the same color as Y, and that eventually at least one such bin will survive to Y with high probability.

Lemma 21. *After processing i nodes of the same color as Y, with high probability there are at least $\frac{n^{1-\frac{i}{20|V|}}}{16e}$ sets of plateau balls, such that each set has cardinality at least $\frac{\log(n)}{100\log(\log(n))}$, have been assigned together to a bin with surplus $z_j \leq \frac{2}{n\log(\log(n))\log^{2c_{close}}(n)}$, and no two sets were ever mapped to the same state within this color.*

Proof. Trivially, this holds for $i = 1$ from Corollary 13.

First, we note that all sets of plateau balls will again be mapped together. This is because they are in the same configuration, as for any node X_i, as it inherits its color, it must be true that they all have the same values for $\text{Pa}(X_i)$ and because they are plateau balls they must have $E^* = e_1^*$.

Now, we will make a random variable $\mathcal{S}_i$ for whether the i-th configuration survived together. Roughly, we desire $\mathcal{S}_i$ to be 1 if it is assigned to a bin with small surplus with none of the other bins that has survived to this stage, we desire $\mathcal{S}_i$ to be 0 if it is assigned to a bin with non-small surplus, and $\mathcal{S}_i$ to be -1 if it lands in a small surplus bin with another bin that had survived (the intuition is that said bin would likely have a positive $\mathcal{S}_j$ and now we must cancel them out). Now, we slightly modify $\mathcal{S}_i$ so all $\mathcal{S}_i$ are independent. By Corollary 12 we know there will be at least $\frac{n^{1-\frac{1}{40|V|}}}{8}$ small surplus bins with high probability. Let us create a subset of bad bins B_{bad} for which $\mathcal{S}_i$ will take value -1. Before realizing the assignment for the i-th configuration, add all small surplus bins that have already received a bin that survived to this round. Arbitrarily fill the remainder of B_{bad} so that $|B_{\text{bad}}| = \frac{n^{1-\frac{i}{20|V|}}}{16e}$. Let us define $|B_{\text{good}}| = \frac{n^{1-\frac{1}{40|V|}}}{8} - \frac{n^{1-\frac{i}{20|V|}}}{16e}$. So, if the configuration is assigned to B_{bad} then $\mathcal{S}_i = -1$, if assigned to B_{good} then $\mathcal{S}_i = 1$, and otherwise $\mathcal{S}_i = 0$. By Hoeffding's inequality, it holds that $\sum_i \mathcal{S}_\rangle \geq \frac{n^{1-\frac{i+1}{20|V|}}}{16e}$ with high probability and thus the lemma holds. □

Concluding large conditional entropy from desirable bin. By Lemma 21, it is clear that with high probability there is at least $\frac{n^{19/20}}{16e} > \sqrt{n}$ bin y' of Y satisfying the desired properties. Now, we seek to prove that this implies $H(X_{\text{src}}|Y = y')$. Consider a looser definition of surplus:

Definition 15 (Relaxed Surplus, $\mathcal{T}^{\text{relax}} = \frac{120|V|}{n\log(n)}$)**.** *We define the surplus of a state i of Y as $z_i^{relax} = \sum_{j \notin S} \max(0, P(X = j, Y = i) - \frac{120|V|}{n\log(n)})$.*

Claim 19. There exists a bin with at least $\frac{\log(n)}{100\log(\log(n))}$ plateau balls and relaxed surplus at most $\frac{2|V|}{n\log(\log(n))\log^{2c_{\text{close}}}(n)}$

Proof. There are two contributors towards relaxed surplus. First, when configurations were assigned to color-root(Y), each plateau ball brought related$_2(x)$ mass with it that could contribute to the surplus. By Lemma 19, each of the plateau balls we considered satisfied related$_2(x) \leq \frac{18|V|(|V|-1)}{n^2}$. Accordingly, there is at most $n \times \frac{18|V|(|V|-1)}{n^2} \leq \frac{18|V|^2}{n}$ such mass in total. Among the at least $\sqrt{n}$ bins that survived to Y, let us choose the one with the least initial mass from related$_2$. Accordingly, it must have at most $\frac{18|V|^2}{n^{1.5}}$ such mass.

Now, consider the at most $|V - 1|$ times that mass may have been acquired by landing in a bin with at most $\frac{2}{n\log(\log(n))\log^{2c_{\text{close}}}(n)}$ surplus. Combining all these masses and calculating the worst-case relaxed mass results in an upper-bound of $\frac{2}{n\log(\log(n))\log^{2c_{\text{close}}}(n)} \times (|V| - 1) + \frac{18|V|^2}{n^{1.5}} \leq \frac{2|V|}{n\log(\log(n))\log^{2c_{\text{close}}}(n)}$. This is because the definition of relaxed surplus gives enough threshold to fit within it all the mass that was within the regular surplus threshold for each of the groups we are aggregating. □

Now, we show that this implies $H(X|Y = y')$ is large, with almost exactly the same proof as Lemma 17:

Lemma 22 (High-entropy conditional)**.** *Given a bin y' that has $z_{y'}^{relax} \leq \frac{2|V|}{n \cdot \log(\log(n)) \cdot \log^{2c_{close}}(n)}$, and receives $\frac{\log(n)}{100 \log(\log(n))}$ plateau balls, then $H(X_{src}|Y = y') = \Omega(\log(\log(n)))$.*

Proof. To show $H(X_{\text{src}}|Y = y')$ is large, we first define the vector v such that $v(x) = P(X_{\text{src}} = x, Y = y')$. Similarly, we define $\overline{v}(x) = \frac{v}{P(Y=y')}$, meaning $\overline{v}(x) = P(X_{\text{src}} = x|Y = y')$ and $|\overline{v}|_1 = 1$. Our underlying goal is to show $H(\overline{v})$ is large. To accomplish this, we will split the probability mass of v into three different vectors $v_{\text{initial}}, v_{\text{plateau}}, v_{\text{surplus}}$ such that $v = v_{\text{initial}} + v_{\text{plateau}} + v_{\text{surplus}}$. The entries of v_{plateau} will correspond to mass from plateau states of X, v_{initial} will correspond to the first $\mathcal{T}^{\text{relaxed}}$ mass from non-plateau states of X_{src}, and v_{surplus} will correspond to mass that contributes to the surplus $z_{y'}$. We more formally define the three vectors as follows:

- v_{plateau}. The vector of probability mass from plateau states of X_{src}. $v_{\text{plateau}}(x)$ is 0 if $x \notin S$ and $v_{\text{plateau}}(x) = P(X_{\text{src}} = x, Y = y')$ if $x \in S$.

- v_{initial}. For non-plateau states of X_{src}, their first $\mathcal{T}$ probability mass belongs to v_{initial}. $v_{\text{initial}}(x) = \min(P(X = x, Y = y'), \mathcal{T}^{\text{relaxed}})$ if $x \notin S$ and $v_{\text{initial}}(x) = 0$ otherwise.

- v_{surplus}. For non-plateau states of X_{src}, their probability mass beyond the first $\mathcal{T}^{\text{relaxed}}$ mass belongs to v_{surplus}. This corresponds to the surplus quantity. $v_{\text{surplus}}(x) = \max(0, P(X_{\text{src}} = x, Y = y') - \mathcal{T}^{\text{relaxed}})$ if $x \notin S$ and $v_{\text{surplus}}(x) = 0$ otherwise. By this definition, $z_{y'} = |v_{\text{surplus}}|_1$.

To show $H(X_{\text{src}}|Y = y') = H(\overline{v})$ is large, we divide our approach into two steps:

1. Show there is substantial helpful mass: $|v_{\text{initial}} + v_{\text{plateau}}|_1 = \Omega\left(\frac{1}{n \cdot \log(\log(n)) \cdot \log^{2c_{\text{close}}}(n)}\right)$
2. Show the distribution of helpful mass has high entropy: $H\left(\frac{v_{\text{initial}} + v_{\text{plateau}}}{|v_{\text{initial}} + v_{\text{plateau}}|_1}\right) = \Omega(\log(\log(n)))$
3. Show that, even after adding the hurtful mass, the conditional entropy is large: $H(X_{\text{src}}|Y = y') = H(\overline{v}) \geq H\left(\frac{v_{\text{initial}} + v_{\text{plateau}}}{|v_{\text{initial}} + v_{\text{plateau}}|_1}\right) - O(1) = \Omega(\log(\log(n)))$

In the first step, we are showing that the distribution when focusing on just the helpful mass of $v_{\text{initial}}, v_{\text{plateau}}$ has high a substantial amount of probability mass. In the second step, we prove how this distribution of helpful mass has high entropy. In the third step, we show that the hurtful mass of v_{surplus} does not decrease entropy more than a constant.

First, we show that there is a substantial amount of helpful mass:

Claim 20. $|v_{\text{initial}} + v_{\text{plateau}}|_1 = \frac{1}{100c_{\text{lb}}n \cdot \log(\log(n)) \cdot \log^{2c_{\text{close}}}(n)}$

Proof. Recall that the bin y' received $\frac{\log(n)}{100\log(\log(n))}$ plateau balls. As defined in Lemma 10, the set S of plateau states is defined such that $\frac{\max_{x \in S} P(X=x)}{\min_{x \in S} P(X=x)} \leq \log^{c_{\text{close}}}(n)$ and $\min_{x \in S} P(X = x) \geq \frac{1}{c_{\text{lb}} n \log(n)}$. Also recall that by Lemma 11 the most probably state of E has large probability. In particular, $P(E = e_1) \geq \frac{1}{\log^{c_{\text{close}}}(n)}$. Let the subset $S' \subseteq S$ be the subset of plateau states of X such that their plateau ball is mapped to y'. In particular, for every $x \in S'$ it holds that $f(x, e_1) = y'$. Accordingly, $P(X_{\text{src}} = x, Y = y') \geq P(X_{\text{src}} = x) \cdot P(E = e_1)$ for $x \in S'$. Thus, the total weight from plateau states of X_{src} is at least $|S'| \cdot \min_{x \in S'} P(X_{\text{src}} = x) \cdot P(E = e_1) \geq |S'| \cdot \frac{\max_{x \in S'} P(X_{\text{src}}=x)}{\log^{c_{\text{close}}}(n)} \cdot P(E = e_1) \geq \frac{1}{100c_{\text{lb}} n \log(\log(n)) \log^{2c_{\text{close}}}(n)}$. □

Next, we show the distribution of helpful mass has high entropy:

Claim 21. $H\left(\frac{v_{\text{initial}}+v_{\text{plateau}}}{|v_{\text{initial}}+v_{\text{plateau}}|_1}\right) \geq \frac{\log(\log(n))}{4}$

Proof. Let us define $\overline{v}_{\text{helpful}} = \frac{v_{\text{initial}}+v_{\text{plateau}}}{|v_{\text{initial}}+v_{\text{plateau}}|_1}$ to be the vector of helpful mass, and we will show $H(\overline{v}_{\text{helpful}})$ is large by upper-bounding $\max_x \overline{v}_{\text{helpful}}(x)$.

For non-plateau states of X_{src}, it follows from Claim 20 that $\max_{x \notin S} \overline{v}_{\text{helpful}}(x) \leq \frac{\mathcal{T}^{\text{relaxed}}}{|v_{\text{initial}}+v_{\text{plateau}}|_1} \leq \frac{\mathcal{T}^{\text{relaxed}}}{\frac{1}{100c_{\text{lb}}n \cdot \log(\log(n)) \cdot \log^{2c_{\text{close}}}(n)}} = \frac{100|V|c_{\text{lb}} \log(\log(n)) \cdot \log^{2c_{\text{close}}}(n)}{\log(n)}$. For plateau states of X, in Claim 20 we also developed the lower-bound of $|v_{\text{initial}} + v_{\text{plateau}}|_1 \geq |S'| \cdot \frac{\max_{x \in S'} P(X=x)}{\log^{c_{\text{close}}}(n)} \cdot P(E = e_1) \geq \frac{\log(n) \cdot \max_{x \in S'} P(X=x)}{2\log^{2c_{\text{close}}}(n) \log(\log(n))}$. Accordingly, we can upper-bound $\max_{x \in S'} \overline{v}_{\text{helpful}}(x) \leq \frac{\max_{x \in S'} P(X=x)}{|v_{\text{initial}}+v_{\text{plateau}}|_1} \leq \frac{100 \log(\log(n)) \log^{2c_{\text{close}}}(n)}{\log(n)}$.

Accordingly, we can lower-bound the entropy of $H(\overline{v}_{\text{helpful}}) = \sum_x \overline{v}_{\text{helpful}}(x) \cdot \log(\frac{1}{\overline{v}_{\text{helpful}}(x)}) \geq \sum_x \overline{v}_{\text{helpful}}(x) \cdot \log(\frac{1}{\max_{x'} \overline{v}_{\text{helpful}}(x')}) = \log(\frac{1}{\max_{x'} \overline{v}_{\text{helpful}}(x')}) \geq \log(\frac{1200c_{\text{lb}} \log(n)}{\log^{2c_{\text{close}}}(n) \log(\log(n))}) = (1 - 2c_{\text{close}}) \log(\log(n)) - \log(\log(\log(n))) - \log(1200c_{\text{lb}}) =$

$\frac{\log(\log(n))}{2} - \log(\log(\log(n))) - \log(1200c_{\text{lb}}) \geq \frac{\log(\log(n))}{4}$ for sufficiently large n where $\frac{\log(\log(n))}{2} \geq \log(\log(\log(n))) + \log(1200c_{\text{lb}})$. □

Finally, we show the hurtful mass does not decrease entropy much, and thus our conditional distribution has high entropy:

Claim 22. $H(X|Y = y') = H(\overline{v}) \geq \Omega(1) \cdot H\left(\frac{v_{\text{initial}}+v_{\text{plateau}}}{|\overline{v}_{\text{initial}}+\overline{v}_{\text{plateau}}|_1}\right) - O(1) = \Omega(\log(\log(n)))$

Proof. We lower-bound $H(\overline{v})$ with the main intuitions that $H\left(\frac{v_{\text{initial}}+v_{\text{plateau}}}{|v_{\text{initial}}+v_{\text{plateau}}|_1}\right) = \Omega(\log(\log(n)))$ and $\frac{|v_{\text{initial}}+v_{\text{plateau}}|_1}{|v_{\text{initial}}+v_{\text{plateau}}+v_{\text{surplus}}|_1} = \Omega(1)$. We more precisely obtain this lower-bound for $H(\overline{v})$ as follows:

$$\begin{aligned}
H(\overline{v}) &= H\left(\frac{v_{\text{initial}} + v_{\text{plateau}} + v_{\text{surplus}}}{|v_{\text{initial}} + v_{\text{plateau}} + v_{\text{surplus}}|_1}\right) \\
&= \sum_x \frac{v_{\text{initial}}(x) + v_{\text{plateau}}(x) + v_{\text{surplus}}(x)}{|v_{\text{initial}} + v_{\text{plateau}} + v_{\text{surplus}}|_1} \times \\
&\qquad \log \frac{|v_{\text{initial}} + v_{\text{plateau}} + v_{\text{surplus}}|_1}{v_{\text{initial}}(x) + v_{\text{plateau}}(x) + v_{\text{surplus}}(x)} \\
&\geq \sum_x \frac{v_{\text{initial}}(x) + v_{\text{plateau}}(x)}{|v_{\text{initial}} + v_{\text{plateau}} + v_{\text{surplus}}|_1} \times \\
&\qquad \log \frac{|v_{\text{initial}} + v_{\text{plateau}} + v_{\text{surplus}}|_1}{v_{\text{initial}}(x) + v_{\text{plateau}}(x)} - 2 \qquad (3.10) \\
&\geq \sum_x \frac{v_{\text{initial}}(x) + v_{\text{plateau}}(x)}{|v_{\text{initial}} + v_{\text{plateau}} + v_{\text{surplus}}|_1} \times \\
&\qquad \log \frac{|v_{\text{initial}} + v_{\text{plateau}}|_1}{v_{\text{initial}}(x) + v_{\text{plateau}}(x)} - 2 \\
&= \frac{|v_{\text{initial}} + v_{\text{plateau}}|_1}{|v_{\text{initial}} + v_{\text{plateau}} + v_{\text{surplus}}|_1} H\left(\frac{v_{\text{initial}} + v_{\text{plateau}}}{|v_{\text{initial}} + v_{\text{plateau}}|_1}\right) - 2 \\
&= \frac{|v_{\text{initial}} + v_{\text{plateau}}|_1}{|v_{\text{initial}} + v_{\text{plateau}}|_1 + z_{y'}} H\left(\frac{v_{\text{initial}} + v_{\text{plateau}}}{|v_{\text{initial}} + v_{\text{plateau}}|_1}\right) - 2 \\
&\geq \frac{1}{1 + 50c_{\text{lb}}|V|} \cdot H\left(\frac{v_{\text{initial}} + v_{\text{plateau}}}{|v_{\text{initial}} + v_{\text{plateau}}|_1}\right) - 2 \qquad (3.11) \\
&= \Omega(\log(\log(n))) \qquad (3.12)
\end{aligned}$$

To obtain Step 3.10, we note that all summands are manipulated from the form $\sum_x p_x \log(\frac{1}{p_x})$ to $\sum_x p'_x \log(\frac{1}{p'_x})$ where $p'_x \leq p_x$ for all x. As the derivative of $p\log(\frac{1}{p})$ is non-negative for $0 \leq p \leq \frac{1}{e}$, the value of at most two summands can decrease, and they can each decrease by at most one. To obtain Step 3.11, we use Claim 20. To obtain Step 3.12, we use Claim 22. □

Thus, we have shown $H(X|Y = y') = \Omega(\log(\log(n)))$. □

Corollary 14. *Under our assumptions, $H(X_{src}|Y = y') = \Omega(\log(\log(n)))$ and thus $H(\tilde{E}) = \Omega(\log(\log(n)))$.*

Counterexample for General Identifiability with Unconfounded-Pairwise Oracles

We formalize the oracle first discussed in Section 3.6:

Definition 16 (Unconfounded-pairwise oracle)**.** *An unconfounded-pairwise oracle is an oracle that returns the correct orientation of an edge if the edge exists in the true graph and if there is no confounding for the edge.*

Consider a causal graph G_1 with four nodes and the edge set $\{X_1 \to X_2, X_1 \to X_3, X_2 \to X_4, X_2 \to X_3, X_3 \to X_4\}$. Likewise, consider G_2 with four nodes and the edge set $\{X_2 \to X_1, X_3 \to X_1, X_4 \to X_2, X_2 \to X_3, X_4 \to X_3\}$. Note that these graphs are in the same Markov equivalence class. Finally, consider a causal graph G_3 with four nodes and the edge set $\{X_1 \to X_2, X_1 \to X_3, X_4 \to X_2, X_2 \to X_3, X_4 \to X_3\}$. If we orient all edges in the skeleton for G_1 and G_2 without conditioning using an unconfounded-pairwise oracle, G_3 is a consistent output of the oracle for both G_1 and G_2. As a result, the peeling approach would see the same set of edge orientations for G_1 and G_2 and not be able to identify the true source.

Proof of Theorem 3

We need to show that in each iteration of the **while** loop in *line 4*, the algorithm correctly identifies all non-source nodes and only the true non-sources. In the first iteration of the loop, there is no node to condition on and therefore tests are independence (not conditional independence) tests. Consider a non-source node N in the initial graph. Then there must be a directed path from some source node X_{src}

to N. Due to faithfulness assumption, two nodes with a directed path cannot be unconditionally independent. Then either Oracle$(X_{\text{src}}, N|\emptyset)$ will return $X_{\text{src}} \to N$ in *line 12* or Oracle$(N, X_{\text{src}}|\emptyset)$ will return $X_{\text{src}} \to N$. In either case, N is added to the non-source list. Now suppose S is a source node. It is never identified as a non-source since it is either conditionally independent with the node it is compared with, found in *line 9* , or it is conditionally dependent with a non-source (for which it has a path to) and the oracle orients correctly in *line 11* or *line 13.* Since all non-sources are correctly identified and only the true non-sources are identified as non-sources, all sources are correctly identified as well in *line 15.* Since sources are incomparable in the partial order, the order in which they are added to the topological order does not impact the validity of topological order in *line 18.*

Suppose the **while** loop identified all sources correctly for all iterations j, $\forall j < i$. Let $\mathcal{S}_i$ be the sources in $\mathcal{R}$, i.e., the sources that are to be discovered in iteration i. Let $Pa_{\mathcal{S}_i}$ be the set of parents of $\mathcal{S}_i$ in the initial graph. Then $Pa_{\mathcal{S}_i} \subseteq \mathcal{C}$, i.e., the set of conditioned nodes include all the parents of the current source nodes. Therefore, $\mathcal{C}$ blocks all backdoor paths from $\mathcal{S}_i$, effectively disconnecting the previous found sources from the graph: This is because $G \backslash \mathcal{C}$ is a valid Bayesian network for the conditional distribution $p(.|\mathcal{C} = c)$. Therefore, in $G \backslash \mathcal{C}$, $\mathcal{S}_i$ are source nodes and the source pathwise oracle correctly identifies all non-source nodes, similar to the base case.

This implies after the **while** loop terminates, we have a valid topological order for the nodes in the graph. Finally, the for loop in *line 19* converts the obtained total order into a partial order since either it removes an edge, or adds an edge that is consistent with the topological order. Therefore we only need to show that non-edges are correct. Suppose X_i, X_j are non-adjacent in the true graph. Then conditioned on all ancestors of X_i, X_j they are independent due to d-separation in the graph. Therefore all non-edges are identified at some iteration of the **for loop**. Furthermore, under the faithfulness assumption, no edge can be mistakenly identified as a non-edge: No two adjacent nodes at any stage of the algorithm are independent conditioned on the ancestors of the cause variable[3].

[3] The faithfulness assumption, which was overlooked by us at the submission deadline, will be

3.9.2 Additional Experiments and Experimental Details

Further Synthetic Experiments and Comparisons with ANM

Figure 3-4 compares the performance under the additive noise model assumption, i.e., the data is sampled from $X = f(\mathrm{Pa}_X) + N_X$ for all variables. The noise term is chosen as a uniform, zero-mean cyclic noise in the supports $\{-1, 0, 1\}, \{-2, -1, 0, 1, 2\}$, $\{-3, -2, -1, 0, 1, 2, 3\}$ which corresponds to different entropies $H(N_X)$. This entropy is shown on the $x-$axis in Figure 3-4. The corresponding plots where the $x-$axis represents the number of samples are given in the Appendix. While this is the generative model that discrete ANM is designed for, its performance is either approximately matched or exceeded by entropic enumeration.

For further experiments with different number of nodes, please see Figures 3-4, 3-5, 3-7, 3-10.

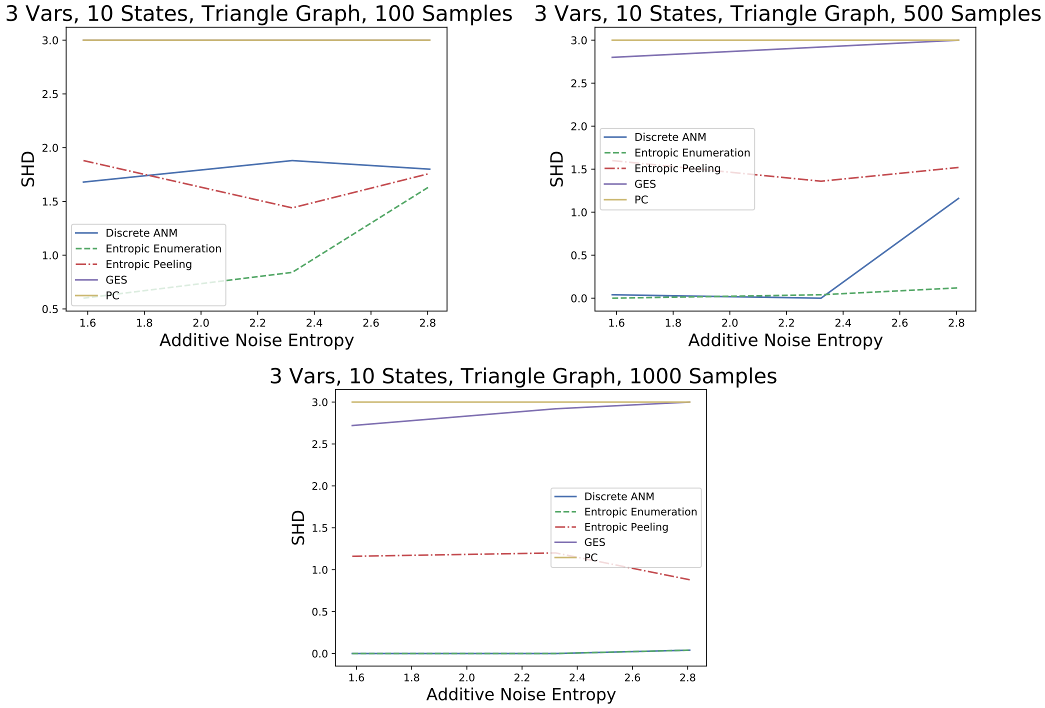


Figure 3-4: Performance of methods in the ANM setting in the triangle graph $X \to Y \to Z, X \to Z$: 25 datasets are sampled for each configuration from the ANM model $X = f(\mathrm{Pa}_X) + N$. The $x-$axis shows entropy of the additive noise.

added to the main paper in camera-ready.

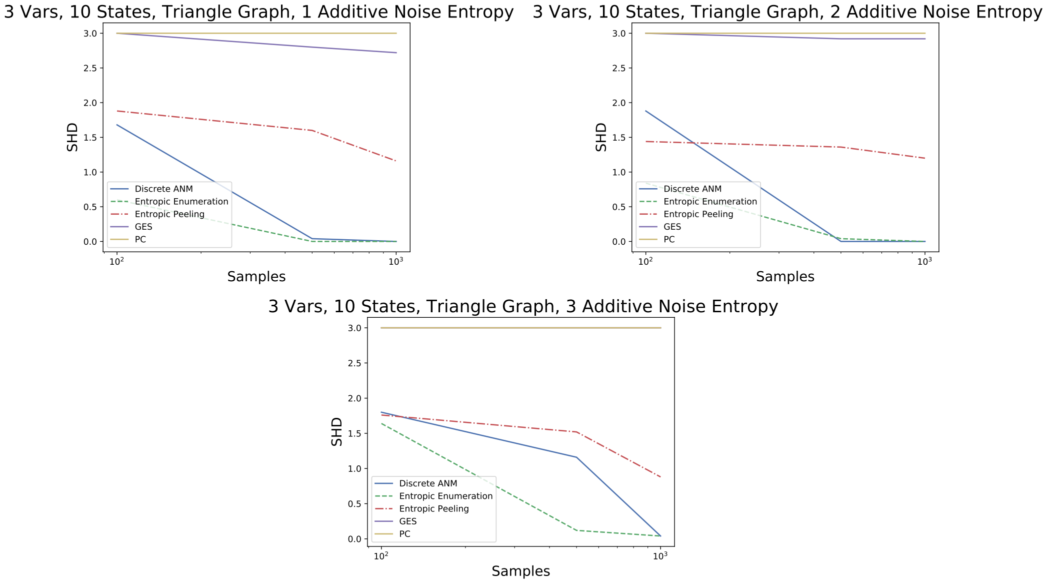


Figure 3-5: Performance of methods in the ANM setting in the triangle graph $X \to Y \to Z, X \to Z$: 25 datasets are sampled for each configuration from the ANM model $X = f(\mathrm{Pa}_X) + N$. The $x-$axis shows the number of samples in each dataset. Entropic enumeration outperforms ANM algorithm in the low-noise low-sample regime.

Entropy Measure for Peeling Algorithm

In this section, we compare different versions of peeling algorithm, one that uses only the exogenous entropy and one that uses the total entropy in pairwise comparisons. We randomly sample exogenous distributions according to symmetric Dirichlet distribution, which is characterized by a single parameter α. By varying α and with rejection sampling, we are able to generate distributions for the exogenous nodes E such that $H(E) \leq \theta$ for some θ. For each distribution, we then compare the structural Hamming distance of the output of our peeling algorithm with the true graph. The structural Hamming distance (SHD) is the number of edge modifications (insertions, deletions, flips) required to change one graph to another. Results are given in Figure 3-12. Let $H(E)$ and $H(\tilde{E})$ be the minimum exogenous entropy needed to generate Y from X and the minimum exogenous entropy needed to generate X from Y, respectively. At each step of the algorithm for every pair X, Y, the red curve compares $H(X)$ with $H(Y)$, the blue curve compares $H(E)$ with $H(\tilde{E})$, and the green curve compares $H(X) + H(E)$ with $H(Y) + H(\tilde{E})$ and orients the edge based on the minimum. As

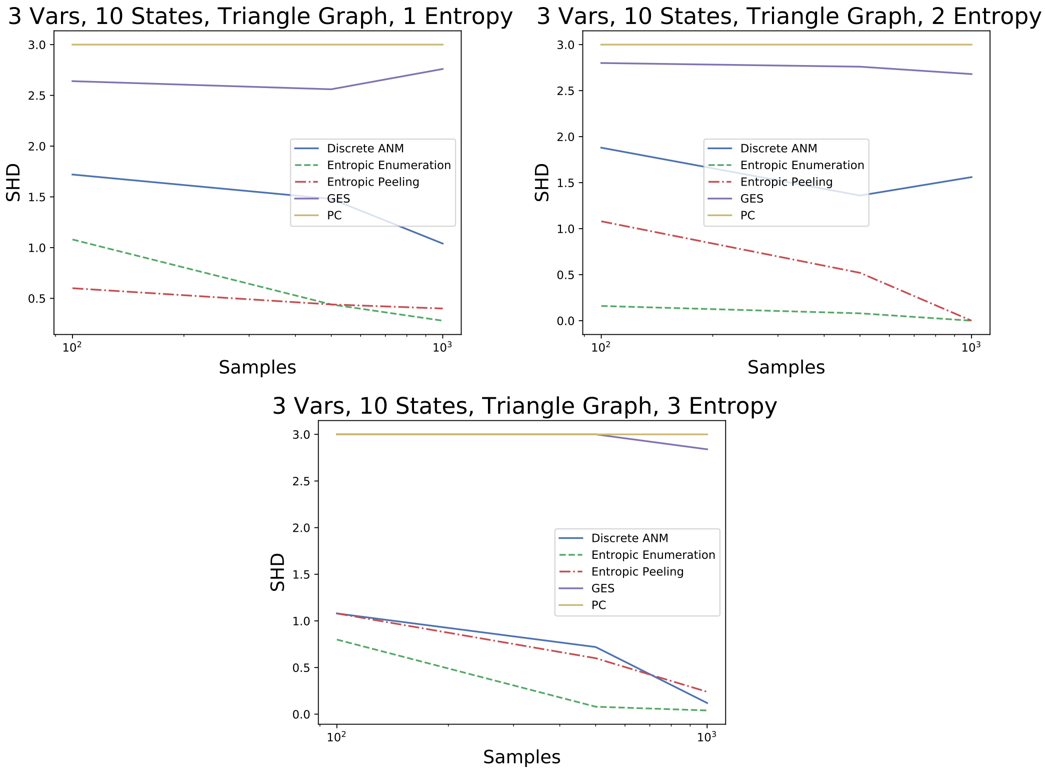


Figure 3-6: Performance of methods in the unconstrained setting in the triangle graph $X \to Y \to Z, X \to Z$: 25 datasets are sampled for each configuration from the unconstrained model $X = f(\text{Pa}_X, E_X)$. The $x-$axis shows the number of samples in each dataset. Entropic enumeration and peeling algorithms consistently outperform the ANM algorithm in almost all regimes.

expected, comparing exogenous entropies perform better than comparing observed variables' entropies in the low-entropy regime. Interestingly, we observe that comparing total entropies consistently performs much better than either.

Entropy Percentile of True Graph

In this section, We test the hypothesis that *when true exogenous entropies are small, the true causal graph is the DAG that minimizes the total entropy*. To test this, we find the minimum entropy needed to generate the joint distribution for every directed acyclic graph that is consistent with the true graph skeleton. We then look at the percentile of the entropy of the true graph. For example, if there are 5 DAGs with less entropy than the true graph out of 100 distinct DAGs, then the percentile is

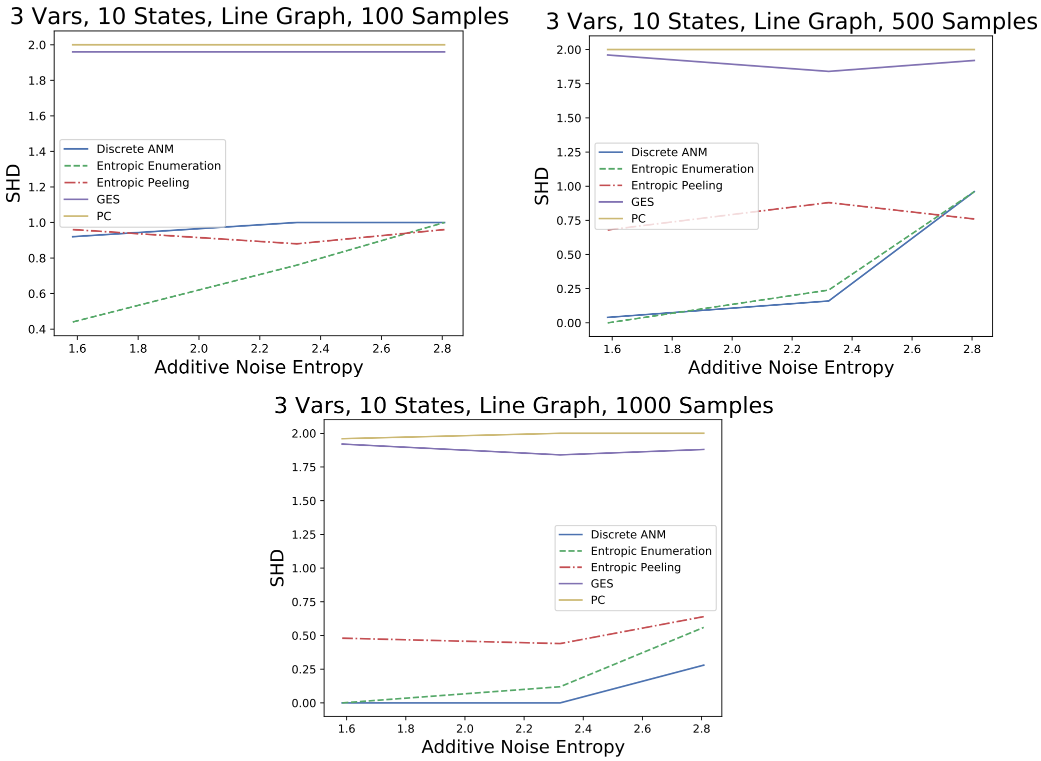
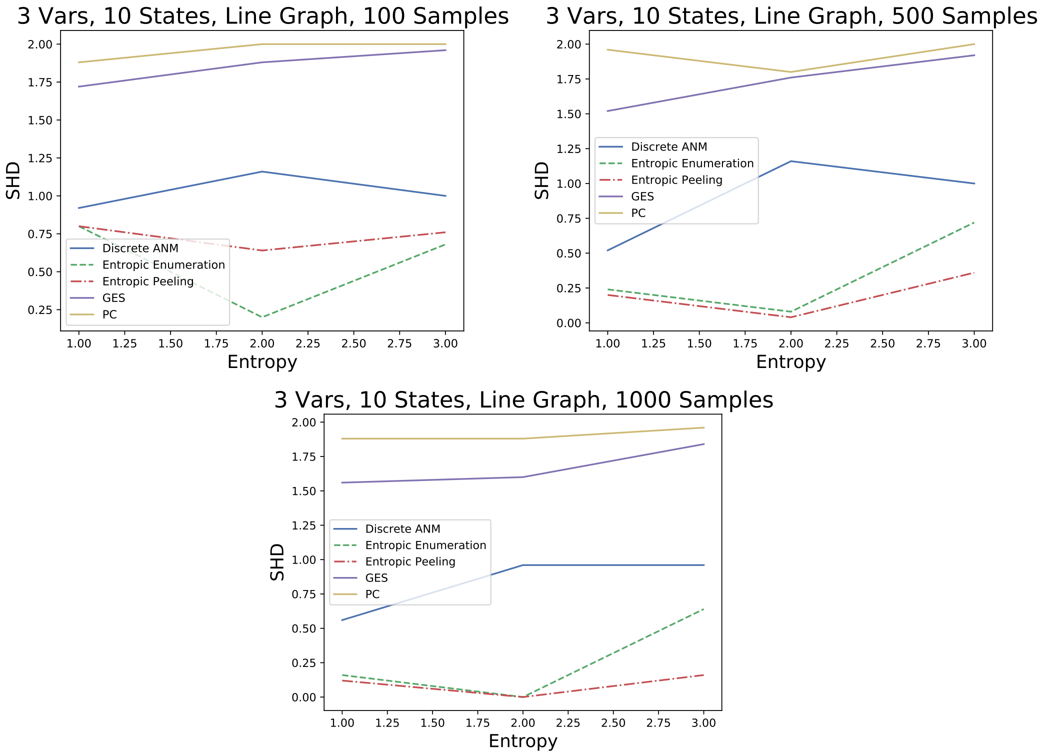


Figure 3-7: Performance of methods in the ANM setting on the line graph $X \rightarrow Y \rightarrow Z$: 25 datasets are sampled for each configuration from the ANM model $X = f(\mathrm{Pa}_X) + N$. The $x-$axis shows entropy of the additive noise.

$1 - 5/100 = 0.95$.

Synthetic Data

Figure 3-13 shows the results for various graphs. It can be seen that, especially for dense graphs, the true graph is the unique minimizer of total entropy needed to generate the joint distribution for a very wide range of entropy values. This presents exhaustive search as a practical algorithm for graphs with a small number of nodes (or generally, those for which the MEC is small). Our experiments, contrary to our theory, show that even when the number of nodes is the same as the number of states, entropic causality can be used for learning the graph.



3 Vars, 10 States, Line Graph, 100 Samples

3 Vars, 10 States, Line Graph, 500 Samples

3 Vars, 10 States, Line Graph, 1000 Samples

Figure 3-8: Performance of methods in the unconstrained setting on the line graph $X \rightarrow Y \rightarrow Z$: 25 datasets are sampled for each configuration from the unconstrained model $X = f(\mathrm{Pa}_X, E_X)$. The $x-$axis shows entropy of the exogenous noise. Entropic enumeration and peeling algorithms consistently outperform the ANM algorithm in all regimes.

Semi-synthetic Data

We use the *bnlearn* repository[4] which contains a selection of Bayesian network models curated from real data [57]. Using these models, we can generate any number of samples and test the accuracy of our algorithms. The datasets however are typically binary which makes them less suitable for Algorithm 2. We therefore limit our use of this data to test our hypothesis that the true graph has the smallest entropy among all graphs consistent with the skeleton. The results are given in Figure 3-14. As can be seen, for most of the small-sized graphs that we tested, the true graph has one of the smallest entropies. Indeed for *Cancer* dataset, the true graph is the unique minimizer when the number of samples is large enough. Only in *Sachs* data does the

[4] `https://www.bnlearn.com/bnrepository/`

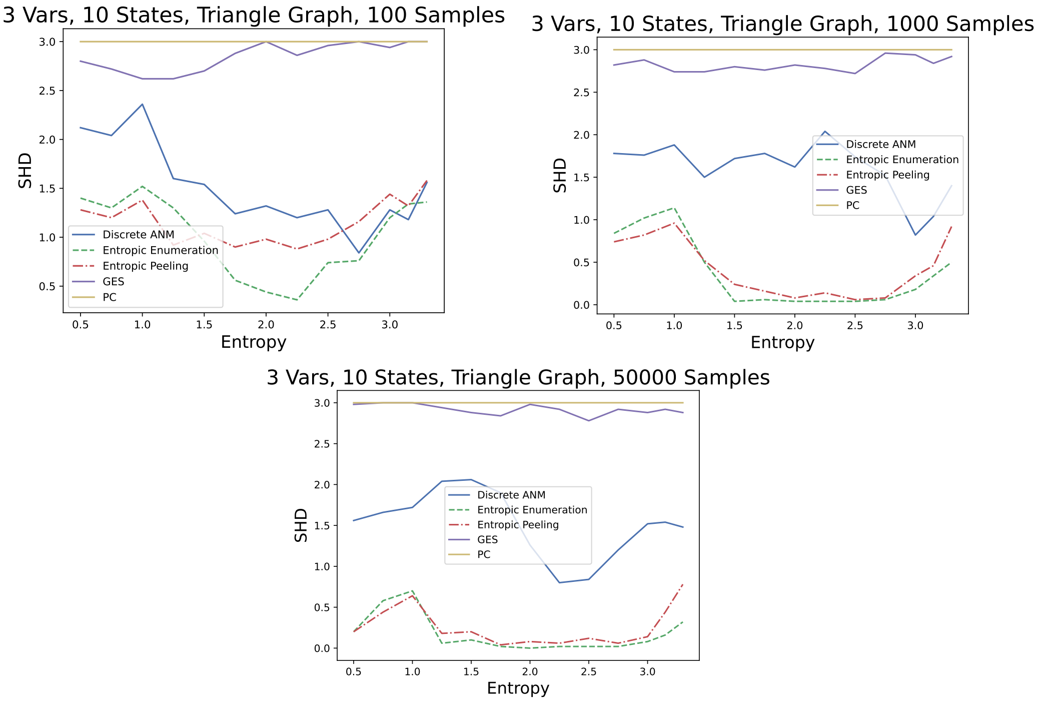


Figure 3-9: Performance of methods in the unconstrained setting in the triangle graph $X \to Y \to Z, X \to Z$: 50 datasets are sampled for each configuration from the unconstrained model $X = f(\mathrm{Pa}_X, E_X)$. The $x-$axis shows entropy of the exogenous noise. Entropic enumeration and peeling algorithms consistently outperform the ANM algorithm in almost all regimes. Note how unlike Figure 3-2, we do not fix the source to have high-entropy or treat the source differently than the other nodes.

true causal graph require one of the largest entropies. This shows that our low-entropy assumption is viable for some real datasets.

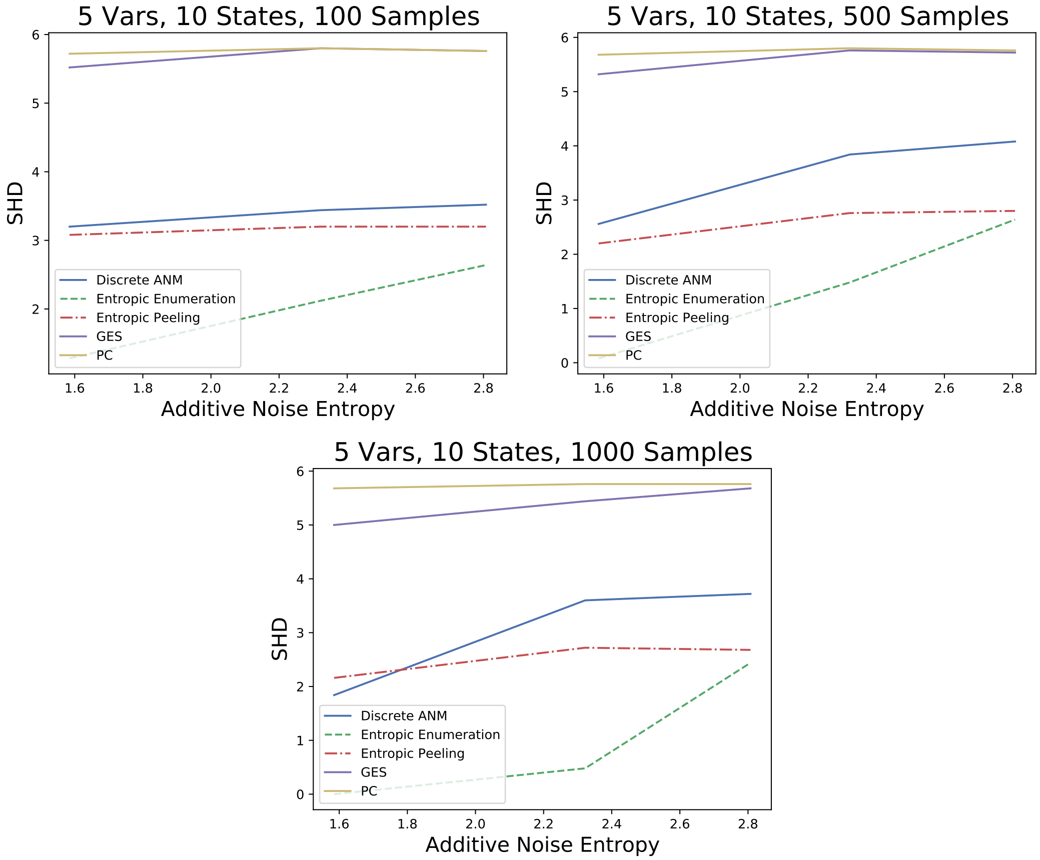


Figure 3-10: Performance of methods in the ANM setting on random 5−node graphs: 25 datasets are sampled for each configuration from the ANM $X = f(\mathrm{Pa}_X) + N$. The x−axis shows entropy of the exogenous noise. Entropic enumeration outperforms consistently in all regimes.

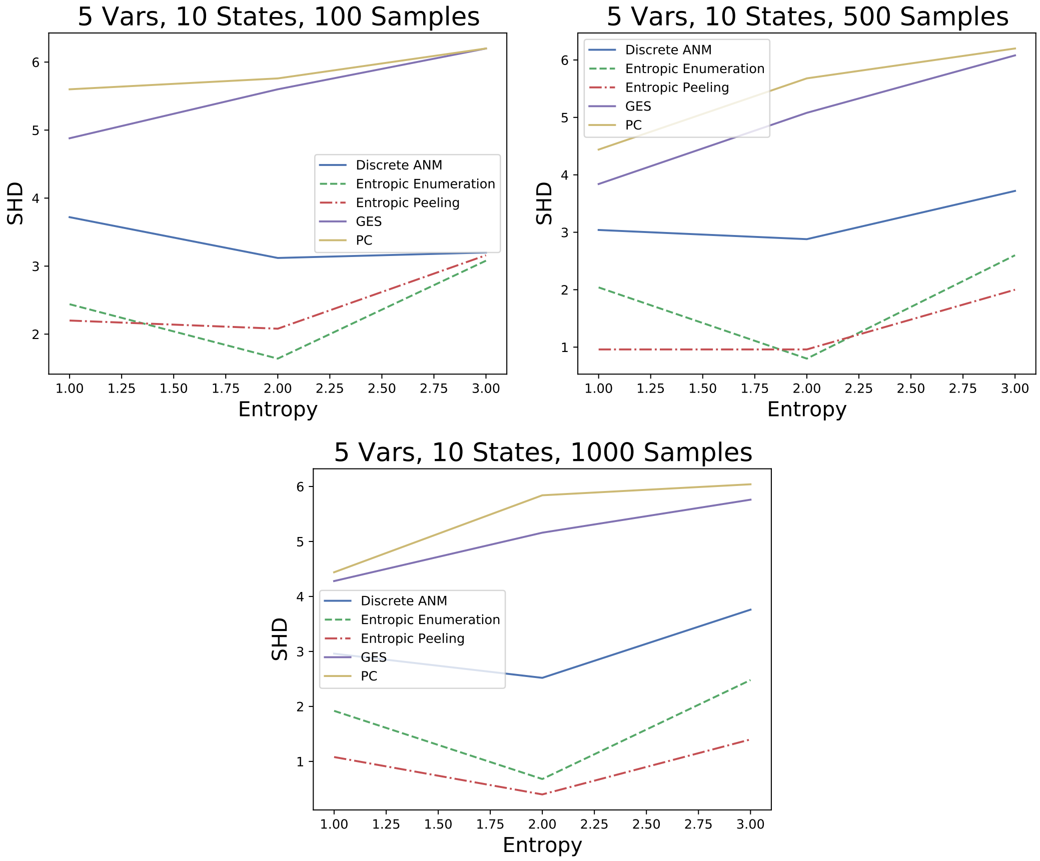
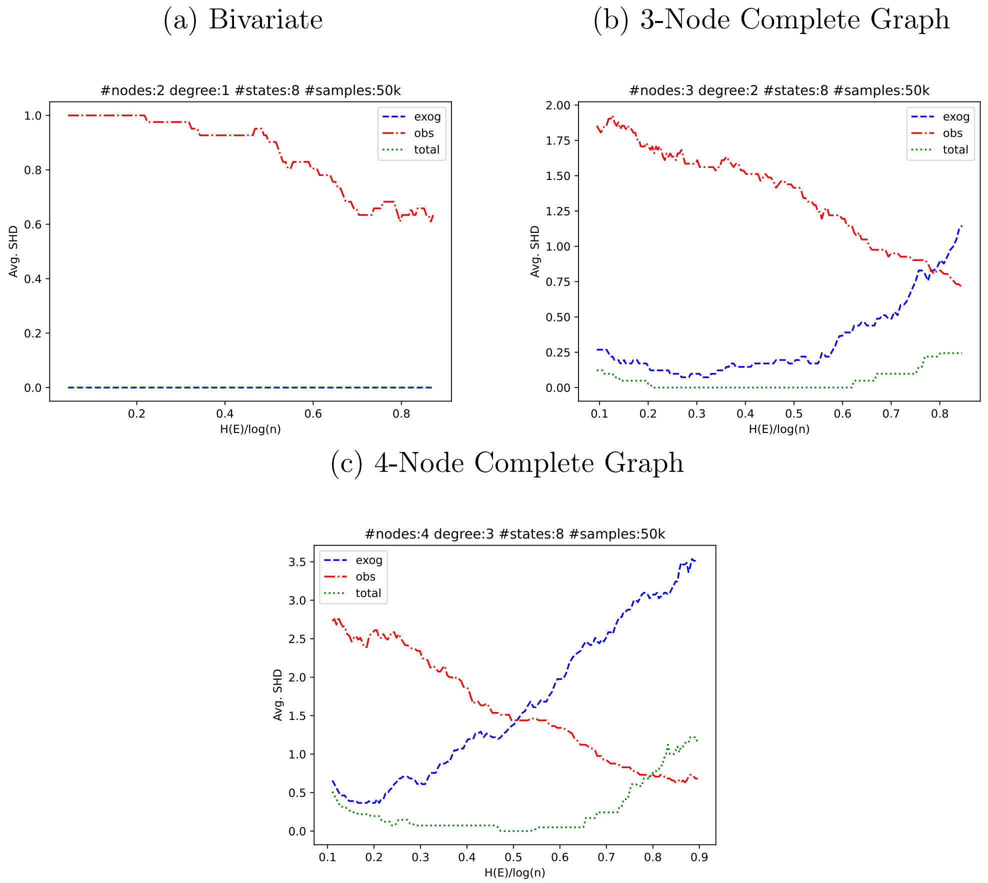


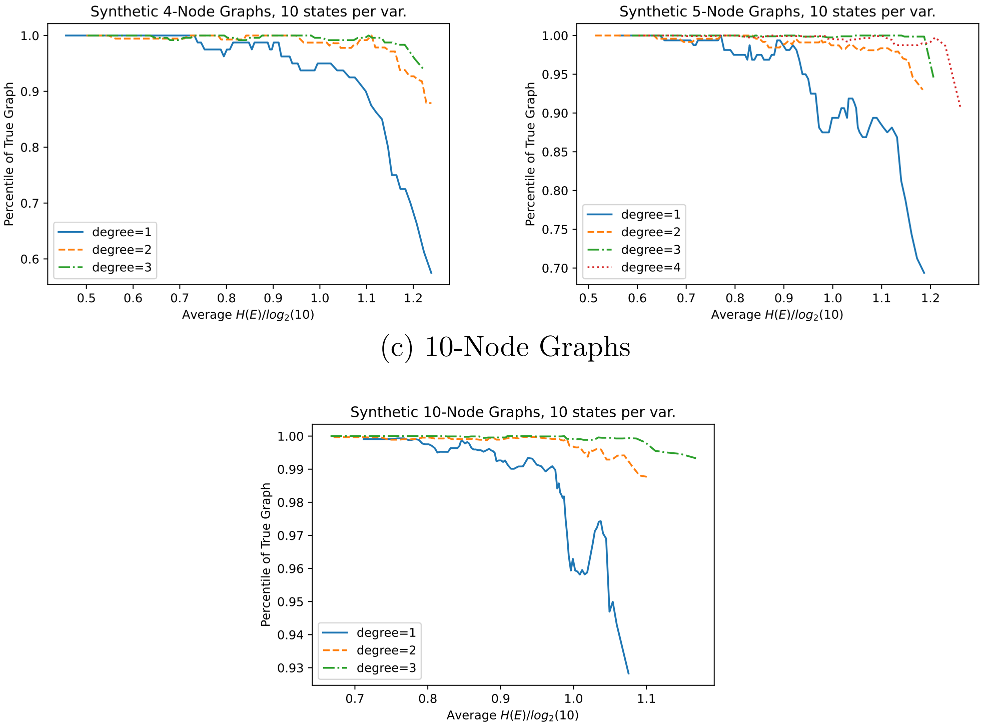
Figure 3-11: Performance of methods in the unconstrained setting on random 5−node graphs: 25 datasets are sampled for each configuration from a random graph from the unconstrained model $X = f(PA_X, N)$. The $x-$axis shows entropy of the additive noise. Entropic enumeration and peeling algorithms consistently outperform the ANM algorithm in all regimes.


(a) Bivariate

(b) 3-Node Complete Graph

(c) 4-Node Complete Graph

Figure 3-12: Average structural Hamming distance (SHD) of peeling algorithm on synthetic data for comparing *i)* exogenous entropies (blue, dashed), *ii)* entropies of observed variables (red, dotted-dashed) and *iii)* total entropies (green, dotted) in line 11 as the Oracle for Algorithm 2. Randomly orienting all edges would result in an average SHD equal to half the number of edges (0.5, 1.5, and 3.0 for (a), (b), and (c), respectively).



(a) 4-Node Graphs

(b) 5-Node Graphs

(c) 10-Node Graphs

Figure 3-13: Percentile of the true graph's entropy compared to minimum entropy required to fit every other incorrect possible causal graph that is consistent with the skeleton (synthetic data).

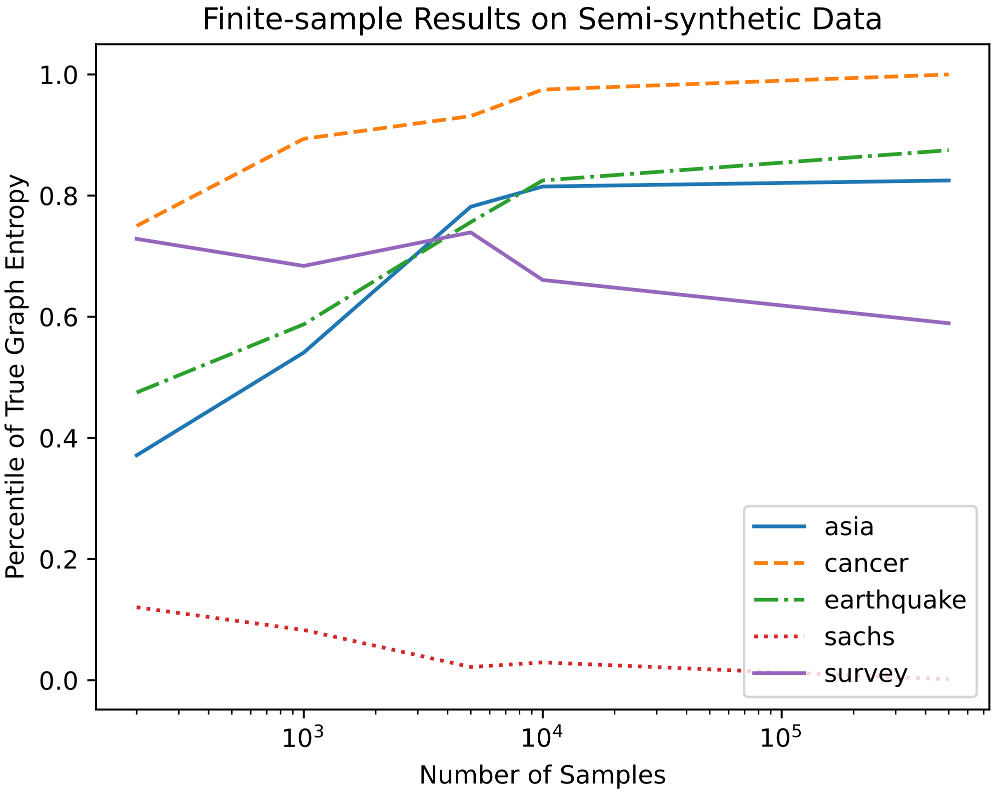


Figure 3-14: Percentile of true graphs entropy compared to minimum entropy required to fit wrong causal graphs in semi-synthetic data from *Bayesian Network Repository* [57].

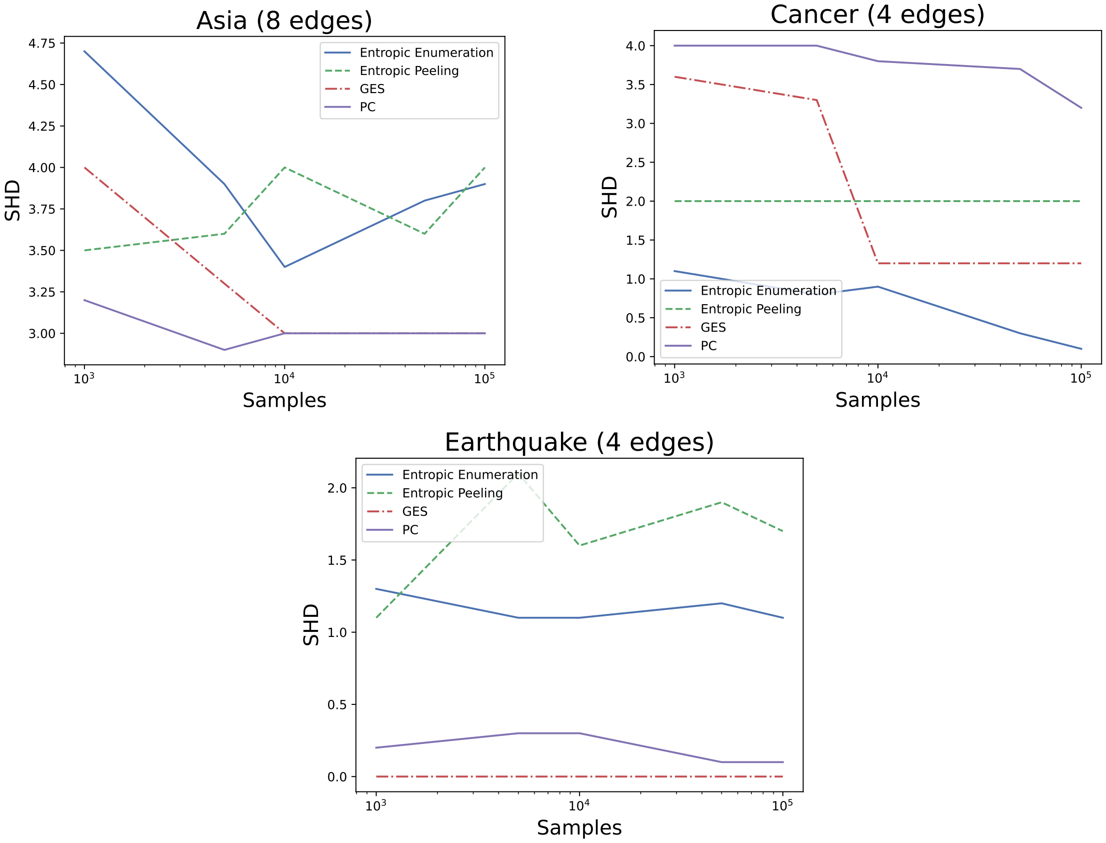


Figure 3-15: Performance of methods on more networks from the *bnlearn* repository with varying samples: 10 datasets are sampled for each configuration from the *bnlearn* network.

Chapter 4

A Tighter Approximation Guarantee for Greedy Minimum Entropy Coupling

4.1 Overview

We examine the minimum entropy coupling problem, where one must find the minimum entropy variable that has a given set of distributions $S = \{p_1, \ldots, p_m\}$ as its marginals. Although this problem is NP-Hard, previous works have proposed algorithms with varying approximation guarantees. In this paper, we show that the greedy coupling algorithm of [Kocaoglu et al., AAAI'17] is always within $\log_2(e)$ (≈ 1.44) bits of the minimum entropy coupling. In doing so, we show that the entropy of the greedy coupling is upper-bounded by $H(\bigwedge S) + \log_2(e)$. This improves the previously best known approximation guarantee of 2 bits within the optimal [Li, IEEE Trans. Inf. Theory '21]. Moreover, we show our analysis is tight by proving there is no algorithm whose entropy is upper-bounded by $H(\bigwedge S) + c$ for any constant $c < \log_2(e)$. Additionally, we examine a special class of instances where the greedy coupling algorithm is exactly optimal.

4.2 Introduction

An instance of the minimum entropy coupling problem is represented by a set S of m distributions, each with n states (i.e., $S = \{p_1, \ldots, p_m\}$). The objective is to find a variable of minimum entropy that "couples" S, meaning its marginals are equal to S. Equivalently, this can be described as finding a minimum entropy joint distribution over variables $p_1, \ldots, p_m$.

This has a variety of applications, including areas such as causal inference [12, 26, 29, 32] and dimension reduction [6, 66]. In the context of random number generation as discussed in [35], the minimum entropy coupling is equivalent to determining the minimum entropy variable such that one sample from this variable enables us to generate one sample from any distribution of S.

While the problem is NP-Hard [33], previous works have designed algorithms with varying approximation guarantees. [10] showed a 1-additive algorithm for $m = 2$ and $\lceil \log(m) \rceil$-additive for general m. [29] introduced the greedy coupling algorithm, [32] showed this is a local optima and [55] showed this is a 1-additive algorithm for $m = 2$. Most recently, [35] introduced a new $(2 - 2^{2-m})$-additive algorithm.

Our Contributions: Our work provides novel perspectives and analytical tools to demonstrate a tighter approximation guarantee for the greedy coupling algorithm. In Section 4.4, we show a closed-form characterization that lower-bounds each state of the greedy coupling. In Section 4.5, we study a class of instances where the greedy coupling is exactly optimal and the lower-bound characterization given in Section 4.4 is tight. Finally, in Section 4.6 we show the greedy coupling is always within $\log_2(e)$ bits of the optimal coupling by proving it is upper-bounded by $H(\bigwedge S) + \log_2(e)$. This improves the best-known approximation guarantee for the minimum entropy coupling problem, and we accomplish this by developing techniques involving a stronger notion of majorization and splitting distributions in an infinitely-fine manner. We show how this analysis is tight and that no algorithm can be upper-bounded by $H(\bigwedge S) + c$ for any constant $c < \log_2(e)$. This resolves that the largest possible gap between $H(\bigwedge S)$ and $H(\mathrm{OPT}_S)$ is $\log_2(e)$.

Table 4.1: Best-Known Additive Approximation Guarantee

	Algorithm (prior/now)		
	Greedy (prior)	***Best (prior)***	***Greedy/Best (now)***
$m = 2$	1 [55]	1 [10]	1 [10, 55]
$m > 2$	$\lceil \log(m) \rceil$[a] [10, 55]	$2 - 2^{2-m}$ [35]	$\mathbf{\log_2(e) \approx 1.44}$

[a] Not explicitly shown before to our knowledge, but can combine [10, 55].

4.3 Background

Notation: The base of log is always 2. H denotes Shannon entropy. The states of any distribution p are sorted such that $p(1) \geq \cdots \geq p(|p|)$. $[n]$ denotes $\{1, \ldots, n\}$. OPT_S denotes the minimum entropy coupling of a set of distributions S.

Greedy Minimum Entropy Coupling: We show approximation guarantees for the greedy coupling algorithm of [29] (formally described in Algorithm 3). At a high-level, the algorithm builds a coupling by repeatedly creating a state of the coupling output that corresponds to the currently largest state of each distribution $p_i \in S$, with weight corresponding to the smallest of these m maximal states. Intuitively, this greedily adds the largest possible state to the coupling at each step. We use $\mathcal{G}_S$ to denote the sequence of states produced by the algorithm. The algorithm runs in $O(m^2 n \log(n))$ time.

Algorithm 3 Greedy Coupling (pseudocode from [32])

```
Input: Marginal distributions of $m$ variables each with $n$ states $\{\mathbf{p_1}, \mathbf{p_2}, ..., \mathbf{p_m}\}$.
Initialize the tensor $\mathbf{P}(i_1, i_2, \ldots, i_n) = 0, \forall i_j \in [n], \forall j \in [n]$.
Initialize $r = 1$.
while $r > 0$ do
    $(\{\mathbf{p_i}\}_{i \in [m]}, r) = \mathbf{UpdateRoutine}(\{\mathbf{p_i}\}_{i \in [m]}, r)$
end while
return $\mathbf{P}$
UpdateRoutine($\{\mathbf{p_1}, \mathbf{p_2}, ..., \mathbf{p_m}\}, r$)
Find $i_j := \arg\max_k \{\mathbf{p_j}(k)\}, \forall j \in [m]$.
Find $u = \min\{\mathbf{p_k}(i_k)\}_{k \in [n]}$.
Assign $\mathbf{P}(i_1, i_2, \ldots, i_n) = u$.
Update $\mathbf{p_k}(i_k) \leftarrow \mathbf{p_k}(i_k) - u, \forall k \in [m]$.
Update $r = \sum_{k \in [n]} \mathbf{p_1}(k)$
return $\{\mathbf{p_1}, \mathbf{p_2}, ..., \mathbf{p_m}\}, r$
```

Majorization: We use ideas from majorization theory [40]. A distribution p is majorized by another distribution q (i.e., $p \preceq q$) if $\sum_{j=1}^{i} p(j) \leq \sum_{j=1}^{i} q(j)$ $\forall i \in [|p|]$. It is known that if $p \preceq q$ then $H(q) \leq H(p)$ [40]. $\bigwedge S$ denotes the greatest lower-bound in regards to majorization such that $\bigwedge S \preceq p$ $\forall p \in S$. Meaning, for any r where $r \preceq p$ $\forall p \in S$, it must hold that $r \preceq \bigwedge S$. For ease of notation, we also use $\mathcal{M}_S$ to refer to $\bigwedge S$. It is known that $\mathcal{M}_S(i) = \min_{p \in S} \sum_{j=1}^{i} p(j) - \sum_{j=1}^{i-1} \mathcal{M}_S(i)$ [11] and that $H(\bigwedge S) \leq H(\mathrm{OPT}_S)$ [10].

4.4 Characterization of Greedy Coupling

To help analyze the performance of the greedy coupling algorithm, we show this closed-form characterization that lower-bounds each element of its output:

Theorem 9. $\mathcal{G}_S(i) \geq \max_j \frac{\sum_{k=1}^{j} \mathcal{M}_S(k) - \sum_{k=1}^{i-1} \mathcal{G}_S(k)}{j}$

Proof. We denote p_ℓ before the t-th step of $\mathcal{G}_S$ as p_ℓ^t. We observe that $\mathcal{G}_S(i)$ is determined by Line 10 of Algorithm 3 to be $\min_\ell \max_k p_\ell^i(k)$. We will lower-bound this quantity:

Claim 23. $\max_{1 \leq k \leq n} p_\ell^i(k) \geq \frac{\sum_{k=1}^{j} \mathcal{M}_S(k) - \sum_{k=1}^{i-1} \mathcal{G}_S(k)}{j}$ $\forall j, \ell$

Proof.

$$\max_{1 \leq k \leq n} p_\ell^i(k) \tag{4.1}$$

$$\geq \max_{1 \leq k \leq j} p_\ell^i(k) \tag{4.2}$$

$$\geq \frac{\sum_{k=1}^{j} p_\ell^i(k)}{j} \tag{4.3}$$

$$= \frac{\sum_{k=1}^{j} p_\ell^1(k) - \sum_{k=1}^{j} (p_\ell^1(k) - p_\ell^i(k))}{j} \tag{4.4}$$

$$\geq \frac{\sum_{k=1}^{j} \mathcal{M}_S(k) - \sum_{k=1}^{i-1} \mathcal{G}_S(k)}{j} \tag{4.5}$$

□

By the definition of $\mathcal{G}_S$ and Claim 23, our theorem holds. □

4.5 Minimum Entropy Coupling of Majorizing Sets

Many related works show guarantees for the minimum entropy coupling problem by showing a relation to the lower-bound of $H(\bigwedge S)$. It is natural to wonder, if we only fix $\bigwedge S$, what is the most challenging that S can be? We introduce a special-case of the minimum entropy coupling problem, where for a fixed value of $\bigwedge S$ we consider the set S to include all distributions that are consistent with $\bigwedge S$ (i.e., all distributions that majorize $\bigwedge S$). More formally, in this variant $S = \textsc{Majorizing-Set}(p) = \{p' | p \preceq p'\}$ for some p. This corresponds to coupling the set of all distributions that majorize a given distribution. We show that in this setting, the greedy coupling produces the optimal solution:

Theorem 10. *When* $S = \textsc{Majorizing-Set}(p)$ *for some* p*, then* $H(\mathcal{G}_S) = H(\mathrm{OPT}_S)$*.*

Proof. First, we clarify:

Claim 24. $\mathcal{M}_S = p$

Proof. For sake of notation, suppose $p(0) = \mathcal{M}_S(0) = 0$. We will inductively show $\mathcal{M}_S(i) = p(i)$ for all $i \in [n]$. First:

$$\begin{aligned}
&\mathcal{M}_S(i) && (4.6)\\
&= \left(\min_{p' \in \textsc{Majorizing-Set}(p)} \sum_{j=1}^{i} p'(j)\right) - \left(\sum_{j=1}^{i-1} \mathcal{M}_S(j)\right) && (4.7)\\
&\geq \left(\min_{p' \in \textsc{Majorizing-Set}(p)} \sum_{j=1}^{i} p(j)\right) - \left(\sum_{j=1}^{i-1} p(j)\right) && (4.8)\\
&= p(i) && (4.9)
\end{aligned}$$

(4.8) follows as all $p' \in \textsc{Majorizing-Set}(p)$ majorize p. Next:

$$\begin{aligned}
&\mathcal{M}_S(i) && (4.10)\\
&= \left(\min_{p' \in \textsc{Majorizing-Set}(p)} \sum_{j=1}^{i} p'(j)\right) - \left(\sum_{j=1}^{i-1} \mathcal{M}_S(j)\right) && (4.11)
\end{aligned}$$

$$\leq \left(\sum_{j=1}^{i} p(j)\right) - \left(\sum_{j=1}^{i-1} p(j)\right) = p(i) \tag{4.12}$$

(4.12) follows as $p \in$ Majorizing-Set(p). □

We now define a distribution $\mathcal{G}'_S$ that mirrors Theorem 9:

Definition 17. $\mathcal{G}'_S(i) = \max_j \frac{\sum_{k=1}^{j} \mathcal{M}_S(k) - \sum_{k=1}^{i-1} \mathcal{G}'_S(k)}{j}$

Clearly $\mathcal{G}'_S$ is a valid distribution as $\mathcal{G}'_S(i) \leq 1 - \sum_{k=1}^{i-1} \mathcal{G}'_S(k)$ and each $\mathcal{G}'_S(i) \geq \frac{1-\sum_{k=1}^{i-1} \mathcal{G}'_S(k)}{n}$. We show that any coupling for S must be majorized by $\mathcal{G}'_S$:

Lemma 23. *If a distribution $\mathcal{C}_S$ couples S, then $\mathcal{C}_S \preceq \mathcal{G}'_S$.*

Proof. For sake of contradiction, suppose $\mathcal{C}_S \npreceq \mathcal{G}'_S$. Then, there must exist an i' where $\sum_{k=1}^{i'} \mathcal{C}_S(k) > \sum_{k=1}^{i'} \mathcal{G}'_S(k)$. Let i' be the earliest such value. Additionally, let $j' = \arg\max_j \frac{\sum_{k=1}^{j} \mathcal{M}_S(k) - \sum_{k=1}^{i'-1} \mathcal{G}'_S(k)}{j}$. We use these to define a distribution $\tilde{p} \in S$ such that $\mathcal{C}_S$ cannot couple $\tilde{p}$:

Definition 18. *$\tilde{p}(k)$ is $\sum_{\ell=1}^{i'} \mathcal{G}'_S(\ell)$ for $k = 1$, is $\mathcal{G}'_S(i')$ for $1 < k \leq j'$, and is $\mathcal{M}_S(k)$ for $k > j'$*

Claim 25. $\tilde{p}$ is a valid probability distribution.

Proof. All states are non-negative. Also, they sum to 1:

$$\sum_{\ell=1}^{n} \tilde{p}(\ell) \tag{4.13}$$

$$= \tilde{p}(1) + \sum_{\ell=2}^{j'} \tilde{p}(\ell) + \sum_{\ell=j'+1}^{n} \tilde{p}(\ell) \tag{4.14}$$

$$= \left(\sum_{\ell=1}^{i'} \mathcal{G}'_S(\ell)\right) + ((j'-1) \times \mathcal{G}'_S(i')) + \left(\sum_{\ell=j'+1}^{n} \mathcal{M}_S(\ell)\right) \tag{4.15}$$

$$= \sum_{\ell=1}^{i'-1} \mathcal{G}'_S(\ell) + \sum_{\ell=1}^{j'} \mathcal{M}_S(\ell) - \sum_{\ell=1}^{i'-1} \mathcal{G}'_S(\ell) + \sum_{\ell=j'+1}^{n} \mathcal{M}_S(\ell) \tag{4.16}$$

$$= \sum_{\ell=1}^{n} \mathcal{M}_S(\ell) = 1 \tag{4.17}$$

(4.16) is obtained by definition of $\mathcal{G}'_S(i')$ and j'. □

Claim 26. $p \preceq \tilde{p}$

Proof. We will show that p is majorized by $\tilde{p}$. To begin:

Subclaim 1. For $k \geq j'$, it holds that $\sum_{\ell=1}^{k} \tilde{p}(\ell) \geq \sum_{\ell=1}^{k} p(\ell)$

Proof.

$$\sum_{\ell=1}^{k} \tilde{p}(\ell) \tag{4.18}$$

$$= \sum_{\ell=1}^{j'} \tilde{p}(\ell) + \sum_{\ell=j'+1}^{k} \tilde{p}(\ell) \tag{4.19}$$

$$= \left(\sum_{\ell=1}^{i'-1} \mathcal{G}'_S(\ell) + j' \times \frac{\sum_{\ell=1}^{j'} \mathcal{M}_S(\ell) - \sum_{\ell=1}^{i'-1} \mathcal{G}'_S(\ell)}{j'} \right)$$

$$+ \sum_{\ell=j'+1}^{k} \mathcal{M}_S(\ell) = \sum_{\ell=1}^{k} \mathcal{M}_S(\ell) \tag{4.20}$$

$$= \sum_{\ell=1}^{k} p(\ell) \tag{4.21}$$

(4.21) is obtained by Claim 24. □

Still, we must show this holds for $k < j'$. We start with:

Subclaim 2. If $j' > 1$, it holds that $\mathcal{G}'_S(i') \leq \mathcal{M}_S(j')$.

Proof. For sake of contradiction, suppose $\mathcal{G}'_S(i') > \mathcal{M}_S(j')$:

$$\mathcal{G}'_S(i') \tag{4.22}$$

$$= \frac{\sum_{\ell=1}^{j'} \mathcal{M}_S(\ell) - \sum_{\ell=1}^{i'-1} \mathcal{G}'_S(\ell)}{j'} \tag{4.23}$$

$$= \frac{j'-1}{j'} \times \frac{\sum_{\ell=1}^{j'-1} \mathcal{M}_S(\ell) - \sum_{\ell=1}^{i'-1} \mathcal{G}'_S(\ell)}{j'-1} + \frac{1}{j'} \times \mathcal{M}_S(j') \tag{4.24}$$

$$\leq \frac{j'-1}{j'} \times \frac{\sum_{\ell=1}^{j'} \mathcal{M}_S(\ell) - \sum_{\ell=1}^{i'-1} \mathcal{G}'_S(\ell)}{j'} + \frac{1}{j'} \times \mathcal{M}_S(j') \tag{4.25}$$

$$= \frac{j'-1}{j'} \times \mathcal{G}'_S(i') + \frac{1}{j'} \times \mathcal{M}_S(j') \tag{4.26}$$

$$< \frac{j'-1}{j'} \times \mathcal{G}'_S(i') + \frac{1}{j'} \times \mathcal{G}'_S(i') = \mathcal{G}'_S(i') \tag{4.27}$$

This is a contradiction. (4.25) follows by definition of j' and (4.27) by supposing $\mathcal{G}'_S(i') > \mathcal{M}_S(j')$. □

Using this, we take the next step:

Subclaim 3. If $1 \leq k < j'$, then $\sum_{\ell=1}^{k} \tilde{p}(\ell) - \sum_{\ell=1}^{k} p(\ell) \geq \sum_{\ell=1}^{k+1} \tilde{p}(\ell) - \sum_{\ell=1}^{k+1} p(\ell)$

Proof.

$$\sum_{\ell=1}^{k} \tilde{p}(\ell) - \sum_{\ell=1}^{k} p(\ell) \tag{4.28}$$

$$= \sum_{\ell=1}^{k+1} \tilde{p}(\ell) - \sum_{\ell=1}^{k+1} p(\ell) + (p(k+1) - \tilde{p}(k+1)) \tag{4.29}$$

$$= \sum_{\ell=1}^{k+1} \tilde{p}(\ell) - \sum_{\ell=1}^{k+1} p(\ell) + (\mathcal{M}_S(k+1) - \mathcal{G}'_S(i')) \tag{4.30}$$

$$\geq \sum_{\ell=1}^{k+1} \tilde{p}(\ell) - \sum_{\ell=1}^{k+1} p(\ell) + (\mathcal{M}_S(j') - \mathcal{G}'_S(i')) \tag{4.31}$$

$$\geq \sum_{\ell=1}^{k+1} \tilde{p}(\ell) - \sum_{\ell=1}^{k+1} p(\ell) \tag{4.32}$$

(4.32) is obtained by Subclaim 2. □

We now show majorization for smaller indices:

Subclaim 4. If $1 \leq k < j'$, then $\sum_{\ell=1}^{k} \tilde{p}(\ell) \geq \sum_{\ell=1}^{k} p(\ell)$

Proof. We can equivalently write this subclaim as how it must hold that for $1 \leq k < j'$, it holds that $\sum_{\ell=1}^{k} \tilde{p}(\ell) - \sum_{\ell=1}^{k} p(\ell) \geq 0$. By Subclaim 1, this holds for $k = j'$. By Subclaim 3, the left-hand side is non-decreasing as we decrease k from j' to 1. Thus, our subclaim is shown inductively. □

It follows from Subclaims 1 and 4 that $p \preceq \tilde{p}$. □

As we now know $\tilde{p} \in S$, we show that $\mathcal{C}_S$ cannot couple $\tilde{p}$:

Claim 27. $\mathcal{C}_S$ cannot couple $\tilde{p}$

Proof. We have designed $\tilde{p}$ such that all states other than $\tilde{p}(1)$ will be too small for any of $\mathcal{C}_S(1), \dots, \mathcal{C}_S(i')$ to be assigned to them in a valid coupling. Additionally, we have set $\tilde{p}(1)$ to be small enough such that not all of $\mathcal{C}_S(1), \dots, \mathcal{C}_S(i')$ can all be assigned to $\tilde{p}(1)$ simultaneously. We prove as follows:

Subclaim 5. $\mathcal{C}_S(1) \geq \cdots \geq \mathcal{C}_S(i') > \mathcal{G}'_S(i')$

Proof. This holds if $\mathcal{C}_S(i') > \mathcal{G}'_S(i')$:

$$\begin{aligned}
&\mathcal{C}_S(i') && (4.33)\\
&= \sum_{k=1}^{i'} \mathcal{C}_S(k) - \sum_{k=1}^{i'-1} \mathcal{C}_S(k) && (4.34)\\
&\geq \sum_{k=1}^{i'} \mathcal{C}_S(k) - \sum_{k=1}^{i'-1} \mathcal{G}'_S(k) && (4.35)\\
&> \sum_{k=1}^{i'} \mathcal{G}'_S(k) - \sum_{k=1}^{i'-1} \mathcal{G}'_S(k) && (4.36)\\
&= \mathcal{G}'_S(i') && (4.37)
\end{aligned}$$

(4.36) is obtained by definition of i'. □

Subclaim 6. For any coupling of $\tilde{p}$ with $\mathcal{C}_S$, all of $\mathcal{C}_S(1), \dots, \mathcal{C}_S(i')$ must be assigned to $\tilde{p}(1)$.

Proof. By definition, $\tilde{p}(2), \dots, \tilde{p}(n) \geq \mathcal{G}'_S(i')$. By Subclaim 5, we then know $\mathcal{C}_S(1) \geq \cdots \geq \mathcal{C}_S(i') > \tilde{p}(2), \dots, \tilde{p}(n)$. As such, all of $\mathcal{C}_S(1), \dots, \mathcal{C}_S(i')$ could only be assigned to $\tilde{p}(1)$. □

Further, not all of $\mathcal{C}_S(1), \dots, \mathcal{C}_S(i')$ can be assigned to $\tilde{p}(1)$:

Subclaim 7. $\tilde{p}(1) < \sum_{k=1}^{i'} \mathcal{C}_S(k)$

Proof. $\tilde{p}(1) = \sum_{k=1}^{i'} \mathcal{G}'_S(k) < \sum_{k=1}^{i'} \mathcal{C}_S(k)$. □

By Subclaim 6 all of $\mathcal{C}_S(1), \dots, \mathcal{C}_S(i')$ can only be assigned to $\tilde{p}(1)$, yet by Subclaim 7 they cannot all be assigned to $\tilde{p}(1)$ simultaneously. Accordingly, $\mathcal{C}_S$ cannot couple $\tilde{p}$. □

Thus, by contradiction, $\mathcal{C}_S \preceq \mathcal{G}'_S$ for any valid $\mathcal{C}_S$. □

By Lemma 23, we conclude $H(\mathrm{OPT}_S) \geq H(\mathcal{G}'_S)$. Now, we show how in this setting $\mathcal{G}_S$ is exactly $\mathcal{G}'_S$:

Lemma 24. *For all i, it holds that $\mathcal{G}_S(i) = \mathcal{G}'_S(i)$.*

Proof. We show this inductively. Using Theorem 9 we know $\mathcal{G}_S(i) \geq \max_j \frac{\sum_{k=1}^{j} \mathcal{M}_S(k) - \sum_{k=1}^{i-1} \mathcal{G}_S(k)}{j} = \max_j \frac{\sum_{k=1}^{j} \mathcal{M}_S(k) - \sum_{k=1}^{i-1} \mathcal{G}'_S(k)}{j} = \mathcal{G}'_S(k)$. Using Lemma 23 we know $\mathcal{G}_S(i) = \sum_{k=1}^{i} \mathcal{G}_S(k) - \sum_{k=1}^{i-1} \mathcal{G}_S(k) \leq \sum_{k=1}^{i} \mathcal{G}'_S(k) - \sum_{k=1}^{i-1} \mathcal{G}_S(k) = \sum_{k=1}^{i} \mathcal{G}'_S(k)$ $- \sum_{k=1}^{i-1} \mathcal{G}'_S(k) = \mathcal{G}'_S(i)$. □

Thus, $H(\mathcal{G}_S) = H(\mathrm{OPT}_S)$, meaning $\mathcal{G}_S$ is optimal. □

We emphasize that in Lemma 24 we have shown how in this setting, the characterization of Theorem 9 is actually exact.

4.6 Greedy Coupling is a $\log_2(e) \approx 1.44$ Additive Approximation

We now show our primary result:

Theorem 11. $H(\mathcal{G}_S) \leq H\left(\bigwedge S\right) + \log_2(e)$

Proof. We will split $\bigwedge S$ in a particular way, and show that $\mathcal{G}_S$ majorizes this modified distribution. Moreover, we will show that it majorizes said distribution in a very strong manner. This will enable a good approximation guarantee for $\mathcal{G}_S$. To split $\bigwedge S$, we introduce the geometric distribution with parameter γ as $\textsc{Geom}_\gamma(x) = \gamma \times (1-\gamma)^{x-1}$. We split $\bigwedge S$ as follows:

Definition 19. $\mathcal{M}_S^\gamma = (\bigwedge S) \times \textsc{Geom}_\gamma$

We will show that $\mathcal{G}_S$ not only majorizes $\mathcal{M}_S^\gamma$ for particular γ, but also satisfies the following stronger notion:

Definition 20. *A distribution p is α-strongly majorized by a distribution q (i.e., $p \preceq_\alpha q$) if for all $i \in [|p|]$ there exists a j such that $\sum_{k=1}^{i} p(k) \leq \sum_{k=1}^{j} q(k)$ and $\alpha \times p(i) \leq q(j)$.*

In other words, p is α-strongly majorized by q if for every prefix of $p(1), \ldots, p(i)$ there is a prefix of q that has at least the same sum, and only contains values at least a factor of α greater than $p(i)$. We show that as we decrease γ to split $\bigwedge S$ more finely, it is increasingly strongly majorized by $\mathcal{G}_S$:

Lemma 25. *For any integer $z \geq 2$, $\mathcal{M}_S^{1/z} \preceq_{z-1} \mathcal{G}_S$*

Proof. We will prove this by contradiction. Suppose that $\mathcal{M}_S^{1/z} \not\preceq_{z-1} \mathcal{G}_S$. This means there exists an i, j such that $\sum_{k=1}^{j} \mathcal{G}_S(k) < \sum_{k=1}^{i} \mathcal{M}_S^{1/z}(k)$ and $\mathcal{G}_S(j+1) < (z-1) \times \mathcal{M}_S^{1/z}(i)$. We show that this cannot occur:

Claim 28. For integer $z \geq 2$ and any i', j', if $\sum_{k=1}^{j'} \mathcal{G}_S(k) < \sum_{k=1}^{i'} \mathcal{M}_S^{1/z}(k)$, then $\mathcal{G}_S(j'+1) \geq (z-1) \times \mathcal{M}_S^{1/z}(i')$.

Proof. Every element of $\mathcal{M}_S^{1/z}$ corresponds to the product of an element of $\bigwedge S$ and an element of $\textsc{Geom}_{1/z}$. We define:

Definition 21. *$\textsc{Index}_{\bigwedge S}(k)$ is the corresponding index of $\bigwedge S$ for $\mathcal{M}_S^{1/z}(k)$. Likewise, $\textsc{Index}_{\textsc{Geom}_{1/z}}(k)$ is the corresponding index of $\textsc{Geom}_{1/z}$ for $\mathcal{M}_S^{1/z}(k)$.*

We define a set $\mathcal{T}^{i'}(k)$ for each index k of $\bigwedge S$, denoting the set of indices of $\textsc{Geom}_{1/z}$ in $\mathcal{M}_S^{1/z}(1), \ldots, \mathcal{M}_S^{1/z}(i')$ corresponding to the k-th element of $\bigwedge S$:

Definition 22. $\mathcal{T}^{i'}(k) = \{\ell | \exists i \leq i' : \textsc{Index}_{\bigwedge S}(i) = k, \textsc{Index}_{\textsc{Geom}_{1/z}}(i) = \ell\}$

Also, we define the set $\mathcal{N}$ as the set of non-empty $\mathcal{T}^{i'}$:

Definition 23. $\mathcal{N} = \{k \in [n] || \mathcal{T}^{i'}(k)| > 0\}$

Finally, we show our claim by:

$$\mathcal{G}_S(j'+1) \tag{4.38}$$

$$\geq \max_k \frac{\sum_{\ell=1}^{k} \mathcal{M}_S(\ell) - \sum_{\ell=1}^{j'} \mathcal{G}_S(\ell)}{k} \tag{4.39}$$

$$\geq \frac{\sum_{\ell=1}^{|\mathcal{N}|}\mathcal{M}_S(\ell)}{|\mathcal{N}|} - \frac{\sum_{\ell=1}^{j'}\mathcal{G}_S(\ell)}{|\mathcal{N}|} \tag{4.40}$$

$$> \frac{\sum_{\ell=1}^{|\mathcal{N}|}\mathcal{M}_S(\ell)}{|\mathcal{N}|} - \frac{\sum_{\ell=1}^{i'}\mathcal{M}_S^{1/z}(\ell)}{|\mathcal{N}|} \tag{4.41}$$

$$\geq \frac{1}{|\mathcal{N}|}\left(\sum_{\ell\in\mathcal{N}}\mathcal{M}_S(\ell) - \sum_{\ell=1}^{i'}\mathcal{M}_S^{1/z}(\ell)\right) \tag{4.42}$$

$$= \frac{1}{|\mathcal{N}|}\sum_{\ell\in\mathcal{N}}\left(\mathcal{M}_S(\ell) - \mathcal{M}_S(\ell)\times\sum_{k\in\mathcal{T}^{i'}(\ell)}\textsc{Geom}_{1/z}(k)\right) \tag{4.43}$$

$$= \frac{1}{|\mathcal{N}|}\sum_{\ell\in\mathcal{N}}\left(\sum_{k=\max(\mathcal{T}^{i'}(\ell))+1}^{\infty}\mathcal{M}_S(\ell)\times\textsc{Geom}_{1/z}(k)\right) \tag{4.44}$$

$$= \frac{1}{|\mathcal{N}|}\sum_{\ell\in\mathcal{N}}\frac{(1-1/z)\times\mathcal{M}_S(\ell)\times\textsc{Geom}_{1/z}(\max(\mathcal{T}^{i'}(\ell)))}{1-(1-1/z)} \tag{4.45}$$

$$\geq \frac{1}{|\mathcal{N}|}\times\sum_{\ell\in\mathcal{N}}\frac{(1-1/z)\times\mathcal{M}_S^{1/z}(i')}{1-(1-1/z)} \tag{4.46}$$

$$= (z-1)\times\mathcal{M}_S^{1/z}(i') \tag{4.47}$$

(4.39) follows from Claim 9. (4.41) follows from the conditions of Claim 28. (4.43) follows by definition of $\mathcal{T}^{i'}$. (4.46) follows from $\mathcal{M}_S(\ell)\times\textsc{Geom}_{1/z}(\max(\mathcal{T}^{i'}(\ell)))\geq\mathcal{M}_S^{1/z}(i')$ because by definition of $\mathcal{T}^{i'}$ there is an element in the prefix of $\mathcal{M}_S^{1/z}(1),\ldots,\mathcal{M}_S^{1/z}(i')$ that corresponds to the ℓ-th element of $\mathcal{M}_S$ and the $\max(\mathcal{T}^{i'}(\ell))$-th element of $\textsc{Geom}_{1/z}$. □

Thus, this contradiction shows that $\mathcal{M}_S^{1/z}\preceq_{z-1}\mathcal{G}_S$. □

We could use Lemma 25 to immediately conclude (by setting $z=2$) that $\mathcal{M}_S^{1/2}\preceq\mathcal{G}_S$ and thus $H(\mathcal{G}_S)\leq H(\bigwedge S)+2$, giving a 2-additive approximation. However, we can do better.

Lemma 26. *If $p\preceq_\alpha q$, then $H(q)\leq H(p)-\log(\alpha)$*

Proof. For any distribution D, we define $\beta_D(x)$ as the set of all indices of D corresponding to the minimum length prefix required to sum to at least x. More formally:

Definition 24. $\beta_D(x) = \{i \in [|D|] | \sum_{j=1}^{i-1} D(j) < x\}$

With this, we show:

$$H(q) \tag{4.48}$$

$$= \sum_{i=1}^{|q|} q(i) \log\left(\frac{1}{q(i)}\right) \tag{4.49}$$

$$= \sum_{i=1}^{|p|} \sum_{j \in (\beta_q(\sum_{k=1}^{i} p(k)) \setminus \beta_q(\sum_{k=1}^{i-1} p(k)))} q(j) \log\left(\frac{1}{q(j)}\right) \tag{4.50}$$

$$\leq \sum_{i=1}^{|p|} \sum_{j \in (\beta_q(\sum_{k=1}^{i} p(k)) \setminus \beta_q(\sum_{k=1}^{i-1} p(k)))} q(j) \log\left(\frac{1}{\alpha \times p(i)}\right) \tag{4.51}$$

$$= \sum_{i=1}^{|p|} \log\left(\frac{1}{\alpha \times p(i)}\right) \times \sum_{j \in (\beta_q(\sum_{k=1}^{i} p(k)) \setminus \beta_q(\sum_{k=1}^{i-1} p(k)))} q(j) \tag{4.52}$$

$$\leq \sum_{i=1}^{|p|} \log\left(\frac{1}{\alpha \times p(i)}\right) \times p(i) \tag{4.53}$$

$$= H(p) - \log(\alpha) \tag{4.54}$$

(4.53) is obtained by noticing how the sequence of the values of the inner summation must majorize p by definition of β_q. As the inner summation's coefficient is non-decreasing, the equation is maximized when sequence of the values of the inner summation is exactly p. □

Corollary 15. *For $z \geq 2$, it holds that $H(\mathcal{G}_S) \leq H(\bigwedge S) + H(\textsc{Geom}_{1/z}) - \log(z-1)$*

Proof. This follows from Lemma 25 and Lemma 26. □

We show this upper-bound approaches $\log_2(e)$ as $z \to \infty$:

Claim 29. $\lim_{z\to\infty} H(\textsc{Geom}_{1/z}) - \log(z-1) = \log_2(e)$

Proof.

$$\lim_{z\to\infty} H(\textsc{Geom}_{1/z}) - \log(z-1) \tag{4.55}$$

$$= \lim_{z\to\infty} \sum_{i=0}^{\infty} \frac{(1-1/z)^i}{z} \times \log\left(\frac{z}{(1-1/z)^i}\right) - \log(z-1) \tag{4.56}$$

$$= \lim_{z\to\infty} \sum_{i=0}^{\infty} \frac{(1-1/z)^i}{z} \times i \times \log\left(\frac{1}{1-1/z}\right) + \log\left(\frac{z}{z-1}\right) \tag{4.57}$$

$$= \lim_{z\to\infty} (z-1) \times \log\left(\frac{1}{1-1/z}\right) + \log\left(\frac{z}{z-1}\right) \tag{4.58}$$

$$= \log_2(e) \tag{4.59}$$

□

Finally, we show that $H(\mathcal{G}_S) \leq H(\bigwedge S) + \log_2(e)$ by contradiction. Suppose there exists an S where $H(\mathcal{G}_S) = H(\bigwedge S) + \log_2(e) + \varepsilon$ for some $\varepsilon > 0$. By combining Corollary 15 and Claim 29 we can immediately conclude there is a sufficiently large z where we can bound $H(\mathcal{G}_S) < H(\bigwedge S) + \log_2(e) + \varepsilon$. This is a contradiction, so it must hold for all S that $H(\mathcal{G}_S) \leq H(\bigwedge S) + \log_2(e)$. □

Moreover, this gap between $H(\mathcal{G}_S)$ and $H(\bigwedge S)$ is tight:

Theorem 12. *There exists no algorithm $\mathcal{A}$ where it holds for all S that $H(\mathcal{A}_S) \leq H(\bigwedge S) + c$ for any $c < \log_2(e)$.*

Proof. Consider the instance $S =$ Majorizing-Set$(\mathcal{U}_n)$ where $\mathcal{U}_n$ is the uniform distribution over n states.

Claim 30. If $S = \mathcal{U}_n$, $\mathcal{G}_S(i) = (1-1/n)^{i-1} \times 1/n \ \forall i \geq 1$.

Proof. By Lemma 24, we know $\mathcal{G}_S(i) = \max_j \frac{\sum_{k=1}^{j} \mathcal{M}_S(k) - \sum_{k=1}^{i-1} \mathcal{G}_S(k)}{j} = \max_{1\leq j\leq n} \frac{j/n - \sum_{k=1}^{i-1} \mathcal{G}_S(k)}{j} = 1/n - \frac{\sum_{k=1}^{i-1} \mathcal{G}_S(k)}{n}$. For $i = 1$, $\mathcal{G}_S(1) = 1/n - \frac{0}{n} = (1 - 1/n)^0 \times 1/n$. For $i > 1$ we can inductively show, $\mathcal{G}_S(i) = 1/n - \frac{\sum_{k=1}^{i-1} \mathcal{G}_S(k)}{n} = 1/n - \frac{n((1/n)-(1-1/n)^{i-1}/n)}{n} = 1/n - \frac{1-(1-1/n)^{i-1}}{n} = (1-1/n)^{i-1} \times 1/n$.

□

Claim 31. If $S = \mathcal{U}_n$, $\lim_{n\to\infty} H(\mathcal{G}_S) = H(\bigwedge S) + \log_2(e)$

Proof. Using Claim 30 we determine that $H(\mathcal{G}_S) = \sum_{i=1}^{\infty} \mathcal{G}_S(i) \times \log(\frac{1}{\mathcal{G}_S(i)}) = \sum_{i=1}^{\infty}(1 - 1/n)^{i-1} \times 1/n \times \log(\frac{1}{1/n \times (1-1/n)^{i-1}}) = \log(n) + \sum_{i=1}^{\infty}(1-1/n)^i \times 1/n \times i \times \log(\frac{1}{1-1/n}) =$

$\log(n) + (n-1) \times \log(\frac{n}{n-1}) = H(\bigwedge S) + (n-1) \times \log(\frac{n}{n-1})$. Finally, $\lim_{n\to\infty} H(\mathcal{G}_S) = H(\bigwedge S) + \lim_{n\to\infty}(n-1) \times \log(\frac{n}{n-1}) = H(\bigwedge S) + \log_2(e)$. □

By Theorem 10, we know $H(\mathcal{G}_S) = H(\mathrm{OPT}_S)$. Accordingly, for any $c < \log_2(e)$ there exists an n where if $S = \textsc{Majorizing-Set}(\mathcal{U}_n)$ then $H(\mathrm{OPT}_S) > H(\bigwedge S) + c$. □

Bibliography

[1] David Alvarez-Melis and Tommi Jaakkola. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 412–421, 2017.

[2] Susan Athey. Machine learning and causal inference for policy evaluation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 5–6, 2015.

[3] Elias Bareinboim and Judea Pearl. Transportability of causal effects: Completeness results. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, 2012.

[4] Kailash Budhathoki and Jilles Vreeken. Mdl for causal inference on discrete data. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 751–756. IEEE, 2017.

[5] Ruichu Cai, Jie Qiao, Kun Zhang, Zhenjie Zhang, and Zhifeng Hao. Causal discovery from discrete data using hidden compact representation. In *Advances in Neural Information Processing Systems*, pages 2671–2679, 2018.

[6] Ferdinando Cicalese, Luisa Gargano, and Ugo Vaccaro. Approximating probability distributions with short vectors, via information theoretic distance measures. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 1138–1142. IEEE, 2016.

[7] Ferdinando Cicalese, Luisa Gargano, and Ugo Vaccaro. How to find a joint probability distribution of minimum entropy (almost), given the marginals. *arXiv preprint 1701.05243*, 2017.

[8] Ferdinando Cicalese, Luisa Gargano, and Ugo Vaccaro. How to find a joint probability distribution of minimum entropy (almost) given the marginals. In *IEEE International Symposium on Information Theory (ISIT)*, pages 2173–2177, 2017.

[9] Ferdinando Cicalese, Luisa Gargano, and Ugo Vaccaro. H(x) vs. h(f(x)). In *IEEE International Symposium on Information Theory (ISIT)*, pages 51–55. IEEE, 2017.

[10] Ferdinando Cicalese, Luisa Gargano, and Ugo Vaccaro. Minimum-entropy couplings and their applications. *IEEE Transactions on Information Theory*, 65(6):3436–3451, 2019.

[11] Ferdinando Cicalese and Ugo Vaccaro. Supermodularity and subadditivity properties of the entropy on the majorization lattice. *IEEE Transactions on Information Theory*, 48(4):933–938, 2002.

[12] Spencer Compton, Murat Kocaoglu, Kristjan Greenewald, and Dmitriy Katz. Entropic causal inference: Identifiability and finite sample results. In *Advances in Neural Information Processing Systems*, volume 33, pages 14772–14782, 2020.

[13] João Cunha. Lecture notes of 15-859m: Randomized algorithms; lec 8: Balls and bins/two choices, February 2011.

[14] P Daniusis, D Janzing, J Mooij, J Zscheischler, B Steudel, K Zhang, and B Schölkopf. Inferring deterministic causal relations. In *26th Conference on Uncertainty in Artificial Intelligence (UAI 2010)*, pages 143–150. AUAI Press, 2010.

[15] Jalal Etesami and Negar Kiyavash. Directed information graphs: A generalization of linear dynamical graphs. In *2014 American Control Conference*, pages 2563–2568. IEEE, 2014.

[16] Jalal Etesami and Negar Kiyavash. Discovering influence structure. In *IEEE International Symposium on Information Theory (ISIT)*, 2016.

[17] AmirEmad Ghassami and Negar Kiyavash. Interaction information for causal inference: The case of directed triangle. In *IEEE International Symposium on Information Theory (ISIT)*, pages 1326–1330, 2017.

[18] Siu-Wai Ho and Raymond W Yeung. The interplay between entropy and variational distance. *IEEE Transactions on Information Theory*, 56(12):5906–5929, 2010.

[19] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems*, 21:689–696, 2008.

[20] Patrik O Hoyer, Dominik Janzing, Joris Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, 2008.

[21] Aapo Hyvärinen and Stephen M Smith. Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *Journal of Machine Learning Research*, 14(Jan):111–152, 2013.

[22] Dominik Janzing. The cause-effect problem: Motivation, ideas, and popular misconceptions. In *Cause Effect Pairs in Machine Learning*, pages 3–26. Springer, 2019.

[23] Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

[24] Dominik Janzing and Bernhard Scholkopf. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

[25] Dominik Janzing, Bastian Steudel, Naji Shajarisales, and Bernhard Schölkopf. Justifying information-geometric causal inference. In *Measures of Complexity*, pages 253–265. Springer, 2015.

[26] Mohammad Ali Javidian, Vaneet Aggarwal, Fanglin Bao, and Zubin Jacob. Quantum entropic causal inference. *arXiv preprint arXiv:2102.11764*, 2021.

[27] Kumar Joag-Dev and Frank Proschan. Negative association of random variables with applications. *The Annals of Statistics*, pages 286–295, 1983.

[28] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

[29] Murat Kocaoglu, Alexandros Dimakis, Sriram Vishwanath, and Babak Hassibi. Entropic causal inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[30] Murat Kocaoglu, Alexandros G. Dimakis, Sriram Vishwanath, and Babak Hassibi. Entropic causal inference. In *AAAI*, 2017.

[31] Murat Kocaoglu, Alexandros G. Dimakis, Sriram Vishwanath, and Babak Hassibi. Entropic causality and greedy minimum entropy coupling. In *IEEE International Symposium on Information Theory (ISIT)*, 2017.

[32] Murat Kocaoglu, Alexandros G Dimakis, Sriram Vishwanath, and Babak Hassibi. Entropic causality and greedy minimum entropy coupling. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1465–1469. IEEE, 2017.

[33] Mladen Kovačević, Ivan Stanojević, and Vojin Šenk. On the entropy of couplings. *Information and Computation*, 242:369–382, 2015.

[34] Samory Kpotufe, Eleni Sgouritsa, Dominik Janzing, and Bernhard Schölkopf. Consistency of causal inference under the additive noise model. In *International Conference on Machine Learning*, pages 478–486, 2014.

[35] Cheuk Ting Li. Efficient approximate minimum entropy coupling of multiple probability distributions. *IEEE Transactions on Information Theory*, 2021.

[36] Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, 5:3065–3105, 2014.

[37] Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15(1):3065–3105, 2014.

[38] Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems*, pages 10869–10879, 2018.

[39] G Marsaglia. Uniform distributions over a simplex. Technical report, Boeing Scientific Research Labs, Seattle WA, 1961.

[40] Albert W Marshall, Ingram Olkin, and Barry C Arnold. *Inequalities: theory of majorization and its applications*, volume 143. Springer, 1979.

[41] Alexander Marx and Jilles Vreeken. Formally justifying mdl-based inference of cause and effect. *arXiv preprint arXiv:2105.01902*, 2021.

[42] Joris Mooij, Dominik Janzing, Jonas Peters, and Bernhard Schölkopf. Regression by dependence minimization and its application to causal inference in additive noise models. In *Proceedings of the 26th International Conference on Machine Learning*, pages 745–752, 2009.

[43] Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204, 2016.

[44] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(1):18–33, 2020.

[45] Stephen L Morgan and Christopher Winship. *Counterfactuals and causal inference.* Cambridge University Press, 2015.

[46] Amichai Painsky, Saharon Rosset, and Meir Feder. Innovation representation of stochastic processes with application to causal inference. *IEEE Transactions on Information Theory*, 66(2):1136–1154, 2019.

[47] Judea Pearl. *Causality.* Cambridge university press, 2009.

[48] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.

[49] Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595, 2014.

[50] Jonas Peters and Peter Bühlman. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101:219–228, 2014.

[51] Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

[52] Jonas Peters, Dominik Janzing, and Bernhard Scholkopf. Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2436–2450, 2011.

[53] Iosif Pinelis. Coordinates of dirichlet distribution negatively associated? MathOverflow. (version: 2018-04-10).

[54] Martin Raab and Angelika Steger. "Balls into bins"—a simple and tight analysis. In *International Workshop on Randomization and Approximation Techniques in Computer Science*, pages 159–170. Springer, 1998.

[55] Massimiliano Rossi. Greedy additive approximation algorithms for minimum-entropy coupling problem. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 1127–1131. IEEE, 2019.

[56] Federica Russo. *Causality and causal modelling in the social sciences*. Springer, 2010.

[57] Marco Scutari. Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010.

[58] S Shimizu, P. O Hoyer, A Hyvarinen, and A. J Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003––2030, 2006.

[59] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

[60] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, Prediction, and Search*. MIT Press, 2000.

[61] Chandler Squires, Dennis Shen, Anish Agarwal, Devavrat Shah, and Caroline Uhler. Causal imputation via synthetic interventions. *arXiv preprint arXiv:2011.03127*, 2020.

[62] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127, 2019.

[63] Mervyn Susser. Glossary: causality in public health science. *Journal of Epidemiology & Community Health*, 55(6):376–378, 2001.

[64] Gregory Valiant and Paul Valiant. Estimating the unseen: Improved estimators for entropy and other properties. *Journal of the ACM (JACM)*, 64(6):1–41, 2017.

[65] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[66] Mathukumalli Vidyasagar. A metric between probability distributions on finite sets of different cardinalities and applications to order reduction. *IEEE Transactions on Automatic Control*, 57(10):2464–2477, 2012.

[67] David Wajc. Negative association: definition, properties, and applications. *Manuscript, available from https://goo. gl/j2ekqM*, 2017.

[68] Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *AAAI*, 2018.

[69] Kun Zhang, Mingming Gong, Petar Stojanov, Biwei Huang, QINGSONG LIU, and Clark Glymour. Domain adaptation as a problem of inference on graphical models. *Advances in Neural Information Processing Systems*, 33, 2020.