Department of Mathematical Sciences
UNIVERSITY OF COPENHAGEN

Phillip Bredahl Mogensen

Master Thesis in Mathematics-Economics

Greedy Learning of Causal Structures in Additive Noise Models

Advisor

Jonas Martin Peters

March 4th, 2019

Abstract

Drawing causal inference from data is an important, but also very difficult task due to the vastness of the space of directed acyclic graphs. Taking an unpublished article by Jonas Peters and Martin Wainwright as a starting point, we present the Greedy entropy-search, a score-based, Greedy search algorithm for causal discovery in Additive Noise Models. Greedily searching the space of graphs reduces the complexity of causal discovery from super-exponential to polynomial. We prove that the Greedy entropy-search recovers the true causal structure of an Additive Noise Model in the population case and illustrate its finite-sample performance by simulation.

Contributions and acknowledgements

The basis of this thesis is rooted in an unpublished manuscript by my thesis adviser Jonas Peters, Associate Professor in statistics at the Department of Mathematical Sciences at University of Copenhagen, and by Martin Wainwright, Chancellor's professor at the University of California at Berkeley. I thank Jonas for letting me be a part of what has been an educative and exciting project for the past six months, and for his helpful guidance along the way.

CONTENTS

INTRODUCTION

The notion of causality is one that is deeply ingrained in both the scientific fields and the human mind. Scientific studies and investigations are performed on a daily basis in an attempt to understand – or even alter – the world that surrounds us. Exactly how causal inference can be drawn has long been a point of debate, and the field of causality itself has been an object of mystery. Only recently has causality transformed into a well-defined, mathematical discipline. As a field in mathematics, causality builds on the theory of graphical models, in which probabilistic systems are represented as graphs and causations as arrows. Causal discovery is the statistical discipline of learning these graphical structures only from data. Being a recently birthed field, causal discovery is a field that is rapidly gaining interest and expanding every day. The act of learning causal structures from data is an ambitious, but immensely important task with wide-reaching applications. As we learn the causal foundations of a system, we gain insight into how the system can be altered. For instance, if we learn that a specific gene causes a disease, we can attempt to create an intervention which deletes the gene and potentially cure the disease. If we do not know the cause of the disease, the task of finding a cure becomes considerably more difficult.

Still, methods of learning causal structures from data are not fully developed and currently plagued by immense computational difficulties, among other things. In fact, if we are given just 22 variables and we are instructed to draw every possible graph that can represent their causal structure, we are going to draw $1.07 \cdot 10^{87}$ different graphs – which is roughly a million times more than the estimated number of atoms in the observable universe. Clearly, we cannot simply list every possible causal structure and see which one fits the best. Instead, different methods need to be employed.

In this thesis, we consider a Greedy algorithm for causal discovery, proposed by Jonas Peters and Martin Wainwright in an unpublished manuscript. The main task of this thesis is prove that this algorithm is capable of recovering the causal structure from a distribution.

CHAPTER 1

AN INTRODUCTION TO CAUSAL INFERENCE

The purpose of this chapter is to give the reader an introduction to the framework graphical models and causality on which we build the subsequent chapters. What we present in this chapter is by no means an exhaustive account of the field of causality. In Section 1.1, we give a primer in graph theory. This is meant to familiarize the reader with notation and basic concepts of graph theory and is to serve as an internal thesaurus on graph theory. It is dense in information and the reader may benefit from only skimming it section to familiarize themselves with the notation and return whenever necessary. In Section 1.2, we introduce the Markov properties, which establish the link between probabilistic independence and the concept of separation in graphs. In Section 1.3 we introduce a class of models, Structural Causal Models, which generate probability distributions in a manner such that these can always be associated with a graph. We define in Section 1.4 the notion of a causal effect in a probabilistic setting and discuss when and how knowledge of a a graph associated with a model can provide insight into causal relationships. In Section 1.5 we narrow our attention to a special case of structural causal models called Additive Noise Models, which are considerable easier to work with than structural causal models.

1.1 A PRIMER IN GRAPH THEORY

We give here a brief overview of graph theory. The theory of graphs is vast and the below is but a small excerpt. Most of the notation follows that of Lauritzen [2017].

A **graph**, denoted $\mathcal{G} = (V, E)$, consists of an ordered pair of nodes, V, and edges, E. An edge is a pair of nodes. If all pairs in E are unordered, we say that $\mathcal{G}$ is an undirected graph, and if all pairs in E are ordered, we say that $\mathcal{G}$ is a directed graph. If E contains both ordered and unordered pairs, we say that $\mathcal{G}$ is mixed. Given a graph, we write $V(\mathcal{G})$ and $E(\mathcal{G})$ to indicate the nodes and edges of $\mathcal{G}$ respectively. If (α, β) is a directed edge, we use the notation $(\alpha \to \beta)$. If it is an undirected edge, we use the notation $(\alpha - \beta)$. Whenever there is an edge between α and β, we say that α and β are **directly connected**. The **skeleton** of $\mathcal{G}$, denoted $\mathrm{ske}(\mathcal{G})$, is the graph obtained by undirecting all directed edges. In the below, we use $\mathcal{G}$ to indicate a graph and α and β to indicate generic nodes in $\mathcal{G}$. A **subgraph** of $\mathcal{G}$, is a graph $\tilde{\mathcal{G}}$ such that every node and edge present in $\tilde{\mathcal{G}}$ is also present in $\mathcal{G}$. We write $\tilde{\mathcal{G}} \subseteq \mathcal{G}$ to indicate that $\tilde{\mathcal{G}}$ is a subgraph. If there is an edge present in $\mathcal{G}$ that is not present in $\tilde{\mathcal{G}}$, we call $\tilde{\mathcal{G}}$ a **true subgraph** of $\mathcal{G}$. The **parents** of α are all the nodes that have a directed edge which goes into α. We write this as the set

$$\mathbf{PA}_{\mathcal{G}}(\alpha) := \left\{\beta \in V : \ (\beta \to \alpha) \in E(\mathcal{G})\right\}.$$

Similarly, the **children** of α are defined as the set

$$\mathbf{CH}_{\mathcal{G}}(\alpha) := \left\{\beta \in V : \ (\alpha \to \beta) \in E(\mathcal{G})\right\}.$$

A **walk** of length n between two nodes, α_0 and α_n, is a sequence of $n+1$ nodes such that every node is directly connected to the next – no matter the direction of the connecting edge. We will in most cases use the symbol ϵ to denote a walk. Formally, a walk $\epsilon := (\alpha_0, \dots, \alpha_n)$ satisfies

$$\forall i \in \{1, \dots, n\} : \qquad (\alpha_{i-1} - \alpha_i) \in \mathrm{ske}(\mathcal{G}).$$

If ϵ has no repeating nodes, we say that ϵ is a **path**[1]. If $\alpha_0 = \alpha_n$, i.e. ϵ starts and ends in the same node, we call ϵ an **n-cycle**. If the length of ϵ is not of importance, we simply say that ϵ is a cycle. The walk ϵ is directed if

$$\forall i \in \{1, \dots, n\} : \quad (\alpha_{i-1} \to \alpha_i) \in E(\mathcal{G}).$$

A directed cycle is a cycle in which all edges are directed. If there exists a directed path from a node, α, to another node, β, we will write $(\alpha \mapsto \beta)$. For a set of nodes, S, we say that the path ϵ **traverses** S if $S \cap \epsilon \neq \emptyset$. A path on the form $(\alpha \leftarrow \beta \to \gamma)$ is called a **fork**. If a path, ϵ, has a subwalk that is a fork, we say that ϵ contains a fork.

The **ancestors** of a node, α, are given as the set

$$\mathbf{AN}_{\mathcal{G}}(\alpha) := \left\{\beta \in V : \ (\beta \mapsto \alpha)\right\}.$$

[1] Note that all paths are by definition also walks.

For simplicity, we will usually write out walks by indicating both nodes and the direction of the edges connecting them. The **descendants** of α are given as the set

$$\mathbf{DE}_{\mathcal{G}}(\alpha) := \left\{\beta \in V : \ (\alpha \mapsto \beta)\right\}.$$

A node different from α that is not a descendant of α is called a **non-descendant** of α. The set of non-descendants of α is denoted $\mathbf{ND}_{\mathcal{G}}(\alpha)$. A **source node**, α, is a node that has no parents, i.e. $\mathbf{PA}_{\mathcal{G}}(\alpha) = \emptyset$. The **depth** of α, denoted $d(\alpha)$, is the length of the longest possible directed walk going from a source node into α. The depth of a source node is set to zero always.

A **section**, ω, of the walk ϵ is a maximal undirected subwalk, i.e. the longest subsequence $(\alpha_k, \dots, \alpha_j)$ such that all edges are undirected. In particular, if all edges in ϵ are directed, any single node is a section. The section ω is **collider** on ϵ if either 1) two arrow-heads of directed edges meet at ω or 2) there is both an undirected edge and a directed edge going into ω. That is, ω is a collider if ϵ takes one of the following forms: $(\dots \to \omega \leftarrow \dots)$, $(\dots \to \omega - \dots)$ or $(\dots - \omega \leftarrow \dots)$.

A directed graph that has no directed cycles, is called a directed acyclic graph (DAG). In the following, we restrict our attention to DAGs. An important concept in causal discovery is that of d-separation in DAGs:

Definition 1.1. Let $\mathcal{G}$ be a graph. A path, ϵ, between two nodes, α and β, is said to be d-separated by the set $S \subset V(\mathcal{G}) \backslash \{\alpha, \beta\}$, with S possibly being the empty set, if and only if

(i)) there exists a subwalk of ϵ that traverses S and has no colliders, or

(ii)) ϵ contains a collider, ω, such that $\omega \notin S$ and $\mathbf{DE}_{\mathcal{G}}(\omega) \cap S = \emptyset$.

A path that is not d-separated by S is called a d-connection relative to S. If $S = \emptyset$ we simply say that α and β are d-connected. If there does not exist a d-connected path between α and β relative to S, we say that α and β are d-separated by S and write

$$\alpha \perp_d \beta \mid S.$$

For disjoint sets, A, B and S, of $V(\mathcal{G})$, we say that A and B are d-separated by S, written $A \perp_d B \mid S$, if

$$\forall (\alpha, \beta) \in A \times B : \quad \alpha \perp_d \beta \mid S. \tag{1.1}$$

We end this section by remarking that the notion of d-separation is not restricted to directed graphs. If $\mathcal{G}$ is an undirected graph, it has no colliders. Thus, point (ii) of Definition 1.1 is never true and the definition reduces to whether a path goes through the conditioning set of nodes.

1.2 GRAPHS, INDEPENDENCE AND MARKOV PROPERTIES

We begin this section with a discussion on graphical representations of probabilistic models. To facilitate the discussion, we follow Lauritzen [2017] and define an *independence model*, $\perp_\sigma$ to be a ternary relation acting on subsets of a finite set – that is, a relation that takes a triple and returns a truth value. An instance of an independence model is the relation $\perp_d$ in equation (1.1). Similarly, for a sequence of labeled random variables, probabilistic conditional independence – i.e. the ternary relation $\perp\!\!\!\perp$ – is also an independence model over the set of labels. That is, they are both relations on a finite set. Generally speaking, it can be difficult to make statements about the conditional independence of random variables. In contrast, it is relatively easy to assess – both computationally and abstractly – whether nodes in a graph can be separated. As both independence models are relations on the same set of triples, it is possible to contain $\perp_d$ in $\perp\!\!\!\perp$, i.e. construct a graph in which d-separation of nodes imply independence of random variables. This leads to the following definition:

Definition 1.2. Let V be a finite set of labels and let $(X_\nu)_{\nu\in V}$ be a collection of random variables with labels in V. Denote by $\mathbb{P}$ the joint distribution of $(X_\nu)_{\nu\in V}$ and let $\mathcal{G}$ be a graph with node set equal to V. The distribution $\mathbb{P}$ is said to obey the **directed global Markov property** relative to $\mathcal{G}$, if for any disjoint sets $A, B, C \subseteq V$ it holds that

$$A \perp_d B \mid C \Rightarrow (X)_A \perp\!\!\!\perp (X)_B \mid (X)_C. \tag{DGM}$$

The notation $(X)_A$ is short-hand for $(X_a)_{a\in A}$. For short, we say that $\mathbb{P}$ is Markov relative to $\mathcal{G}$ whenever (DGM) is satisfied. Whenever $\mathbb{P}$ is Markov relative to $\mathcal{G}$, we call $\mathcal{G}$ a **graphical representation** of $\mathbb{P}$ (or $(X)_V$). If $\mathcal{G}$ is such that (DGM) becomes a bi-implication, we say that $\mathbb{P}$ is **faithful** to $\mathcal{G}$. In this setting, we call $\mathcal{G}$ a minimal representation.

The existence of a graphical representation is guaranteed; a fully connected graph – one in which every pair of nodes are directly connected – satisfies (DGM). Indeed, in the fully connected graph, the antecedent in (DGM) is always false and hence the statement in (DGM) becomes a tautology. However, there is nothing to be gained by representing $\mathbb{P}$ with a fully connected graph. Instead, we are interested in constructing a *sparse* representation of $\mathbb{P}$. A sparse graph is one in which the number of edges is close to as low as possible, while still satisfying that $\mathbb{P}$ is Markov relative to $\mathcal{G}$ or even faithful. Such graphical representations are useful for several reasons:

(i) It is sometimes much simpler to prove statements on separation in a graph than it is to prove statements on probabilistic independence.

(ii) They provide an economic way of describing joint distributions and conditional independence statements.

(iii) They provide a framework in which we can discuss the causal relations between random variables.

Throughout Chapter 3, point (i) above turns out to be a particularly attractive property which we rely heavily on in many of our proofs. We refrain from discussing point (iii) in this section but return to it in Section 1.4. To illustrate point (ii) above, suppose that we are given random variables

X_1 through X_5 and that we wish to write down every conditional independence statement about these. To do so, we first check if X_1 is independent of X_2 given X_3, and then if X_1 is independent of X_2 given X_3 *and* X_4 and so forth. For each pair of distinct variables, we must write down 2^3 independence statements. As there exist $\binom{5}{2}$ unordered pairs, this results in 80 statements. However, had we constructed a minimal representation of $X_1, \ldots, X_5$, the same information could have been stored in five nodes and at most $\frac{5 \cdot 4}{2}$ edges. If we instead had N random variables, the example generalizes to writing down $\binom{N}{2} 2^{N-2}$ independence statements versus constructing a graph with N nodes and at most $\frac{N \cdot (N-1)}{2}$ nodes.

We end this section by defining two properties that are similar to the global Markov property, and a result on when these properties are equivalent.

Definition 1.3. Let $(X_v)_{v \in V}$ be a collection of labelled random variables with joint distribution $\mathbb{P}$, and let $\mathcal{G}$ be a DAG such that $V(\mathcal{G}) = V$. The distribution $\mathbb{P}$ is said to obey the **directed local Markov property** relative to $\mathcal{G}$ if

$$\forall \alpha \in V : \quad X_\alpha \perp\!\!\!\perp \mathbf{ND}_{\mathcal{G}}(X_\alpha) \mid \mathbf{PA}_{\mathcal{G}}(X_\alpha). \tag{DLM}$$

Finally, we say that $\mathbb{P}$ obeys the **recursive factorization property** if $\mathbb{P}$ is dominated by some measure μ, and

$$\frac{\partial \mathbb{P}}{\partial \mu}(x) = \prod_{\alpha \in V} \frac{\partial \mathbb{P}}{\partial \mu}(x_\alpha \mid \mathbf{PA}_{\mathcal{G}}(X_\alpha)). \tag{DF}$$

If it is obvious which graph is being considered, we will imply say that $\mathbb{P}$ is locally Markov or that $\mathbb{P}$ factorizes. In the event that $\mathbb{P}$ has density with respect to a product measure we have the following result:

Theorem 1.4 (Lauritzen [2017, theorem 2.57])
Let $\mathcal{G}$ be a DAG and let $\mathbb{P}$ be a probability measure on a $|V(\mathcal{G})|$-dimensional measurable space. If $\mathbb{P}$ is dominated by a product measure, the following statements are equivalent:

1) $\mathbb{P}$ obeys the directed global Markov property.

2) $\mathbb{P}$ obeys the directed local Markov property.

3) $\mathbb{P}$ obeys the recursive factorization property.

1.3 STRUCTURAL CAUSAL MODELS

In this section, we introduce a class of probabilistic models that are of particular interest to us. These models can be thought of as a collection of random variables that have been constructed on the basis of a graph.

Definition 1.5. Let V denote a finite set of labels and let $(N_\nu)_{\nu\in V}$ be a collection of mutually independent random variables on a common probability space $(\mathbb{R}, \mathbb{B}, \mathbb{P})$. Let $\mathcal{G}$ be a DAG and let $\mathbf{S} = \{S_\nu\}_{\nu\in V}$ be the collection of assignments,

$$\forall \nu \in V : \quad S_\nu = \{X_\nu := f_\nu(\mathbf{PA}_\mathcal{G}(X_\nu), N_\nu)\}, \tag{1.2}$$

where every f_ν is some real-valued and measurable function with image in $\mathbb{R}$.

We call

$$\mathscr{C} := (\mathbf{S}, (N)_V)$$

a **Structural Causal Model** and say that it has graph $\mathcal{G}$. We call the joint distribution of $(X)_V$ the implied distribution of $\mathscr{C}$.

Remark 1.6. At a first glance, it may not be obvious that 'the implied distribution of $\mathscr{C}$' exists. However, the fact that we require $\mathcal{G}$ to be acyclic ensures that it exist. It can be constructed in the following manner: Order V according to node depth[2] and iterate through V, each time carrying out the assignment. That is, starting at a source node, ν_0, the random variable constructed by S_{ν_0} is the measurable function $X_{\nu_0} : \mathbb{R} \to \mathbb{R}$ given by $X_{\nu_0} = f_{\nu_0}(N_{v_0})$. This is well-defined as $\mathcal{G}$ is acyclic, and thus there must exist source nodes. Having done this for all source nodes, we move on to a node of depth one, ν_1, and carry out the assignment S_{ν_1}. The parent set of ν_1 is now a well-defined, finite set of random variables, and thus we can construct the random variable $X_{\nu_1}\mathbb{R}^{|\mathbf{PA}_\mathcal{G}(X_{\nu_1})|} \to \mathbb{R}$ by $X_{\nu_1} = f_{\nu_1}(\mathbf{PA}_\mathcal{G}(X_{\nu_1}), N_{\nu_1})$. This recursive construction of random variables is performed until we have constructed a random variable, X_ν, for each label, ν. Formally, the joint distribution of $(X)_V$ is then the image measure $\mathbb{P} \circ (X)_V^{-1}$.

Theorem 1.7 (Pearl [2009, theorem 1.4.1])
Let $\mathscr{C}$ be a structural causal model (SCM) with graph $\mathcal{G}$ and implied distribution $\mathbb{P}$. The distribution $\mathbb{P}$ is locally Markov with respect to $\mathcal{G}$.

Remark 1.8. If the measure $\mathbb{P}$ is dominated by a product measure – e.g. the Lebesgue measure on a Borel space of an appropriate dimension – it follows directly from Theorem 1.4 that $\mathbb{P}$ is also globally Markov.

Theorem 1.7 shows us, that if a set of random variables, $(X)_V$, is generated according to the scheme in (1.2), the graph $\mathcal{G}$ can be used to conclude at least some conditional independencies in $(X)_V$. If $\mathbb{P}$ is dominated by a product measure, it follows that $\mathcal{G}$ is even a graphical representation of $\mathbb{P}$. Having access to a graphical representation of $\mathbb{P}$ may seem redundant if we have constructed the SCM ourselves. Then we would necessarily know every assignment and thus not need $\mathcal{G}$ to find conditional independencies. In the remainder of this thesis we consider the converse situation: Suppose that we are given a joint distribution, $\mathbb{P}$, along with the information that $\mathbb{P}$ is the implied

[2] Such an ordering need not be unique.

distribution of some SCM which we do not know of. We then consider the problem of *recovering* the graph $\mathcal{G}$ which generated $\mathbb{P}$, by only using properties of the measure $\mathbb{P}$.

1.4 INTERVENTIONS AND CAUSAL EFFECTS

Before we begin our discussion on cause and effect, a good starting point is to conceptualize what cause and effect is. To do so, we must resolve a longstanding argument dating back to Ancient Greece, in which all the great philosophers have partaken, from Plato and Aristoteles to Hume, Kant and even Leibniz [Hulswit, 2004] – a problem that is slightly out of the scope of this thesis. Instead, we simply choose to think intuitively of cause and effect in terms of interventions; if changing X, but nothing else, changes the distribution of Y then X must be a cause of Y. Being mathematicians, we may also take the liberty of simply *defining* what a causal effect is within the confines of this thesis. The definition we make follows those of Pearl [2009] and Peters et al. [2017]. We start by introducing the notion of intervening in a structural model:

Definition 1.9 (Peters et al. [2017, Section 3.2])**.** Let $\mathscr{C} = (\mathbf{S}, (N)_V)$ be an SCM with DAG $\mathcal{G}$ and assignments as in (1.2). An **intervention** on $I \subseteq V$ is a collection of assignments

$$\mathbf{S'} = \{S'_\nu\}_{\nu \in I},$$

where

$$\forall \nu \in I: \qquad S'_\nu := \{X_\nu := \tilde{f}(\mathbf{PA}_{\tilde{\mathcal{G}}}(X_\nu), \tilde{N}_\nu)\},$$

such that $\tilde{\mathcal{G}}$ is a DAG satisfying

$$\forall \nu \in V \backslash I: \qquad \mathbf{PA}_{\tilde{\mathcal{G}}}(X_\nu) = \mathbf{PA}_{\mathcal{G}}(X_\nu).$$

The set $(\tilde{N}_\nu)_{\nu \in I}$ is a set of mutually independent noise variables, possibly differing from $(N)_I$. The distribution implied by the model

$$\tilde{\mathscr{C}} := (\mathbf{S}' \cup \{S_\nu\}_{\nu \in V \backslash I}, ((\tilde{N})_I, (N)_{V \backslash I}))$$

is called the **interventional distribution** and is denoted $\mathbb{P}^{\mathscr{C}, \mathrm{do}(\mathbf{S'})}$. If it is clear which model we are intervening in, we simply write $\mathbb{P}^{\mathrm{do}(\mathbf{S'})}$.

The notion of an intervention may seem complicated at a first glance. In most cases, we will intervene on a single variable, say X_κ, by setting it to a constant, c. In this case, the intervention amounts to replacing the assignment S_κ with $X_\kappa := c$ and deleting all edges going into κ in $\mathcal{G}$.

Definition 1.10 (See e.g. Pearl 2009, Peters et al. 2017)**.** Let $\mathscr{C}$ be an SCM with implied distribution $\mathbb{P}$. For any two labels $\alpha, \beta \in V(\mathcal{G})$, there is a **causal effect** from X_α to X_β if and only if there exists $c \in \mathbb{R}$, such that

$$\mathbb{P}^{\mathscr{C}, \mathrm{do}(X_\alpha := c)}_{X_\beta} \neq \mathbb{P}^{\mathscr{C}}_{X_\beta}.$$

That is, there is a causal effect from X_α to X_β if it is possible to affect the marginal distribution of X_β by intervening on X_α. If there is a causal effect from X_α to X_β, we write $X_\alpha \overset{c}{\to} X_\beta$. It is sometimes more convenient to use an equivalent definition of a causal effect

Remark 1.11 (Peters et al. [2017, Proposition 6.13])**.** There is a causal effect from X_α to X_β if and only if

$$X_\alpha \overset{c}{\to} X_\beta \Leftrightarrow \exists x, x' \in \mathbb{R}: \quad \mathbb{P}_{X_\beta}^{\mathscr{C},\mathrm{do}(X_\alpha:=x)} \neq \mathbb{P}_{X_\alpha}^{\mathscr{C},\mathrm{do}(X_\alpha:=x')}.$$

As we discussed in Section 1.2, conditional independence is closely related to separation in directed graphs. Hence, it is not surprising that we can relate the existence of causal effects to separation in $\mathcal{G}$.

Proposition 1.12 (Peters et al. [2017, proposition 6.14])
Let $\mathscr{C}$ be an SCM with graph $\mathcal{G}$ and let α and β be labels in V. If there is a causal effect from X_α to X_β, then there exists a directed path $\alpha \mapsto \beta$ in $\mathcal{G}$.

In other words, the existence of $\alpha \mapsto \beta$ is a necessary, but not sufficient, condition for there to exist a causal effect from X_α to X_β. The result is, in essence, just a reformulation of the global Markov property. The intervention do($X_\alpha := N$) deletes all parents of X_α in $\mathcal{G}$ and adds a new parent, N, which has no other connections. For α and β to be d-connected in the graph of $\tilde{\mathscr{C}}$, there would have to be a directed path $\alpha \mapsto \beta$ in $\tilde{\mathcal{G}}$. Provided that N is independent of all other noise variables in $\mathscr{C}$, and if N is such that $\mathbb{P}^{\mathscr{C},\mathrm{do}(X_\alpha:=N)}$ is again a dominated measure, Proposition 1.12 is equivalent with the directed global Markov property. Proposition 1.12 tells us, then, that given only a graph from an SCM, we are able to identify all *potential* causal relationships, but cannot be sure that there is one – unless the distribution is faithful. Conversely, we are able to identify some relations as non-causal. I two nodes are not connected by a directed walk, neither can be a cause of the other. The fact that nodes can be connected by a directed walk, but not have a causal relationship may seem counterintuitive. The following example illustrates a case in which this can happen.

Example 1.13. Consider an SCM, $\mathscr{C}$, with graph $\mathcal{G}$ as shown in figure 1.1. The numbers along the edges are taken to be coefficients in a linear mapping. For example, δ should be understood as the random variable $X_\delta = -X_\alpha + N_\delta$. Clearly $\alpha \mapsto \beta$. However, we claim that α is not a cause of β. Consider first the distribution function of X_β.

$$\begin{aligned}\mathbb{P}(X_\beta < x) =& \mathbb{P}(2X_\alpha + 2X_\delta + X_\gamma + N_\beta < x)\\ =& \mathbb{P}(2X_\alpha - 2X_\alpha + X_\gamma + N_\beta < x)\\ =& \mathbb{P}(X_\gamma + N_\beta < x).\end{aligned}$$

Consider now the intervention $\mathrm{do}(X_\alpha := c)$, for some $c \in \mathbb{R}$. We see that

$$\begin{aligned}\mathbb{P}^{\mathrm{do}(X_\alpha:=c)}(X_\beta < x) &= \mathbb{P}^{\mathrm{do}(X_\alpha:=c)}(2c + 2X_\delta + X_\gamma + N_\beta < x)\\ &= \mathbb{P}^{\mathrm{do}(X_\alpha:=c)}(2c - 2c + X_\gamma + N_\beta < x)\\ &= \mathbb{P}^{\mathrm{do}(X_\alpha:=c)}(X_\gamma + N_\beta < x).\end{aligned}$$

Thus,

$$\mathbb{P}_{X_\beta}^{\mathrm{do}(X_\alpha:=c)} = \mathbb{P}_{X_\beta}$$

for any choice of $c \in \mathbb{R}$. Formally, we conclude this by the uniqueness theorem for probability measures (see e.g. Hansen [2009, Theorem 3.7]), as the class of sets on the form $(-\infty, x)$ is an intersection-stable generator of $\mathbb{B}$. We conclude that there is no causal effect from X_α to X_β, even though $\alpha \mapsto \beta$.

As illustrated in Example 1.13, the situation in which two nodes are connected by a directed walk (and directly connected, even), but not causally related can arise when the SCM is such that two edges 'cancel out', so to speak. In Example 1.13 this happened because the edges $(\alpha \to \beta)$ and $(\delta \to \beta)$ cancelled out through the edge $(\alpha \to \delta)$. Yet, knowing the just structure of the graph in Example 1.13 provides us with information about the causal structure of the underlying mechanism that generated it. For instance, we can conclude that γ is a potential cause of β, but that β is certainly not a cause of γ. The caveat to this line of thinking is that in any practical setting, we will not know the structure of the underlying graph. So far, we have yet to discuss the matter of how we may learn the graph of a distribution. This, as it turns out, is no easy task. Estimating the graph of a distribution is an ambitious project on several accounts. For one, the computational complexity grows super-exponentially in the size of the vertex set, making it difficult to search the space of graphs [Bühlmann et al., 2014]. This is a focal point of this thesis and the primary subject matter of Chapter 3. Secondly, not all graphs *can* be learned from their distribution alone [Bühlmann et al., 2014]. When this is the case, we say the graph (or SCM) is not identifiable. Loosely speaking, a graph can be unidentifiable from its distribution, if the same distribution can be generated by multiple SCMs with different graphs. All of Chapter 2 is dedicated to the matter of identifiability and how it can be achieved. As it turns out, the task of finding restrictions on a model that ensures its identifiability becomes markedly easier when we consider a particular class of SCMs. Motivated by identifiability, we present here this smaller class of SCMs, as introduced in Peters et al. [2014], for which we can guarantee identifiability, by leaning on existing works by Zhang and Hyvarinen [2012] and Peters et al. [2014].

1.5 ADDITIVE NOISE MODELS

Definition 1.14 (See e.g. Peters et al. 2014, Hoyer et al. 2009a)**.** A Continuous Additive Noise Model (ANM) with graph $\mathcal{G}$ is an SCM in which every assignment is additive in the noise component and each parent, i.e. the assignments take the form

$$\forall \nu \in V(\mathcal{G}): \quad S_\nu = \left\{ X_\nu := \sum_{\gamma \in \mathbf{PA}_{\mathcal{G}}(X_\nu)} f_{\nu,\gamma}(X_\gamma) + N_\nu \right\}.$$

The individual functions $f_{\cdot,\cdot}$ need not be linear. Whenever ν is a source node, we use the convention that

$$\sum_{\gamma \in \mathbf{PA}_{\mathcal{G}}(X_\nu)} f_{\nu,\gamma}(X_\gamma) = 0.$$

The additive structure of the assignments in an additive noise model (ANM) is considerably simpler than that of a general SCM. In fact, identifiability of an ANM can be proven under relatively mild assumptions. To the knowledge of the author, such a result does not exist for general SCMs. The identifiability of ANMs is the key ingredient in causal discovery. If a model cannot be identified from its distribution, we have little to no possibility of drawing causal inference. Furthermore, in ANMs we can characterize a situation in which when the presence of a direct connection in a graph is equivalent with the presence of a causal effect.

Proposition 1.15
Let $\mathscr{C}$ be an ANM with implied distribution $\mathbb{P}$. Suppose there exists an edge $(\alpha \to \beta)$ in the graph, $\mathcal{G}$, of $\mathscr{C}$. If

$$X_\alpha \perp\!\!\!\perp \{\boldsymbol{PA}_{\mathcal{G}}(X_\beta), N_\beta\} \backslash \{X_\alpha\} \tag{1.3}$$

and

$$\exists x' \in \mathbb{R}: \quad \frac{d}{dx} f_{\beta,\alpha}(x) \bigg|_{x=x'} \neq 0, \tag{1.4}$$

then $X_\alpha \overset{c}{\to} X_\beta$.

Proof. By assumption, there exists c and c' in $\mathbb{R}$, with $c \neq c'$, such that

$$f_{\beta,\alpha}(c) \neq f_{\beta,\alpha}(c').$$

We then see that

$$\mathbb{P}^{\mathrm{do}(X_\alpha = c)}(X_\beta < x) = \mathbb{P}^{\mathrm{do}(X_\alpha = c)} \left(\sum_{m \in \mathbf{PA}_{\mathcal{G}}(X_\beta) \backslash \{X_\alpha\}} f_{\beta,m}(X_m) + N_\beta < x - f_{\beta,\alpha}(c) \right), \tag{1.5}$$

and

$$\mathbb{P}^{\mathrm{do}(X_\alpha = c')}(X_\beta < x) = \mathbb{P}^{\mathrm{do}(X_\alpha = c')} \left(\sum_{m \in \mathbf{PA}_{\mathcal{G}}(X_\beta) \backslash \{X_\alpha\}} f_{\beta,m}(X_m) + N_\beta < x - f_{\beta,\alpha}(c') \right). \tag{1.6}$$

By the assumption (1.3), this must imply that

$$\mathbb{P}_{X_\beta}^{\mathrm{do}(X_\alpha := c)} \neq \mathbb{P}_{X_\beta}^{\mathrm{do}(X_\alpha := c')},$$

as the terms in both (1.5) and (1.6) are distribution functions, and thus not constant everywhere. We conclude that $X_\alpha \overset{c}{\to} X_\beta$. □

Notice the importance of condition (1.3) in the proof of Proposition 1.15; (1.3) ensures that situations such as in Example 1.13 cannot happen. As X_α is independent of every other node that goes into X_β, we are guaranteed that intervening on X_α will not affect the distribution of $\mathbf{PA}_\mathcal{G}(X_\beta)\backslash\{X_\alpha\}$ and N_β, which allows for the conclusion. This is exactly what went wrong in Example 1.13: Here, X_α and X_γ were *not* independent. An immediate consequence, that is easily verifiable once given the graph of an ANM, is that if X_α is a single parent of X_β and $f_{\beta,\alpha}$ is not constant, then $X_\alpha \overset{c}{\to} X_\beta$ – no matter the underlying distributions. In Chapter 3 we will inadvertently restrict ourselves to a class of ANMs that will always satisfy conditions (1.3) and (1.4) which in turn implies that all direct connections will represent a causal relationship.

1.6 OUTLINE OF THESIS

The remainder of this thesis is structed into three chapters and a conclusion. Chapters 2 and 3 are theoretical and built with the end goal of proving that we can recover the graph of an ANM from its distribution by employing a score-based Greedy search algorithm. In Chapter 2, we collect existing results on identifiability to give sufficient conditions for an ANM to be identifiable from its joint distribution. In Chapter 3, we introduce the Entropy Score of graph and use identifiability to prove that it is maximized in the true graph of an ANM. We then decompose the entropy score into its summands to define what we call the marginal score. This is used to construct the Greedy entropy-search method for causal discovery. The remainder of Chapter 3 is spent on proving that the Greedy entropy-score can recover the graph of an ANM under regularity assumptions. In Chapter 4 we take a step back from the theory and discuss the implementation of the Greedy entropy-search. We then carry out simulation studies to assess the performance and applicability of the algorithm. Finally, we end the thesis with a conclusion and remarks on the future perspectives of the Greedy entropy-search method.

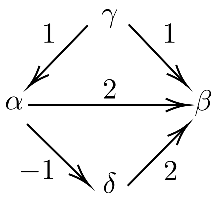

Figure 1.1: *The graph $\mathcal{G}$.*

CHAPTER 2

IDENTIFIABILITY OF ADDITIVE NOISE MODELS

In this short chapter, we discuss identifiability of Additive Noise Models. Its purpose is to formally introduce the notion of identifiable models. We collect existing works on identifiability and use these to state a list of conditions which are sufficient to ensure identifiability of an Additive Noise Model.

2.1 THE MODEL

Throughout this chapter, let $n \in \mathbb{N}$ be a natural number and let $\mathcal{I} := \{1, \dots, n\}$ denote an index set. We consider an additive noise model, $\mathscr{C}$, with DAG $\mathcal{G}$ and random variables, $(X)_{\mathcal{I}}$, assigned according to the following scheme

$$\forall i \in \mathcal{I} : \quad X_i := \sum_{m \in \mathbf{PA}_{\mathcal{G}}(X_i)} f_{i,m}(X_m) + N_i, \tag{2.1}$$

where the functions $f_{\cdot,\cdot}$ are elements of a function class, $\mathcal{F} \subseteq \{f : \mathbb{R} \to \mathbb{R}\}$, and where $(N)_{\mathcal{I}}$ is a sequence of independent random variables, each with density with respect the Lebesgue measure, m. Given a graph, $\mathcal{G}$, a family of functions, $\boldsymbol{f} \subseteq \mathcal{F}$, and a noise distribution, (2.1) implies a unique joint distribution of $(X)_{\mathcal{I}}$. Letting $\mathbb{Q}(\mathcal{G}, \boldsymbol{f})$ denote this distribution, we define

$$\mathcal{S}_{\mathcal{G}}^{\mathcal{F}} := \left\{ \mathbb{Q}(\mathcal{G}, \boldsymbol{f}) : \ \boldsymbol{f} \subset \mathcal{F} \right\}.$$

That is, $\mathcal{S}_{\mathcal{G}}^{\mathcal{F}}$ is the class of all distributions that can be implied by a graph $\mathcal{G}$ and (2.1) choosing different functions from $\mathcal{F}$. When it is obvious which graph is considered, we will simply write $\mathcal{S}^{\mathcal{F}}$.

2.2 IDENTIFIABILITY OF ADDITIVE NOISE MODELS

Consider the model described in Section 2.1, and denote this $\mathscr{C}$. Denote the graph of $\mathscr{C}$ by $\mathcal{G}^0$. We call $\mathscr{C}$ *identifiable*[1] if it holds that

$$\forall \mathcal{G} : \qquad \mathcal{S}_{\mathcal{G}^0}^{\mathcal{F}} \cap \mathcal{S}_{\mathcal{G}}^{\mathcal{F}} \neq \emptyset \Leftrightarrow \mathcal{G} = \mathcal{G}^0.$$

That is, we call an $\mathscr{C}$ identifiable if its implied distribution can *only* be implied by $\mathcal{G}^0$. Models that are not identifiable can instead be identified up to their Markov equivalence class – that is, the class of graphs for which the distribution is Markov – if their entailed distribution is faithful. The problem of identifiability can seem an odd one; even with full knowledge of knowledge of a distribution, it is potentially not possible to identify the graph that generated it. This phenomenon of non-identifiability is a consequence of the fact that two DAGs can entail the same set of *d*-separations. There exists well-documented methods for identifying the Markov equivalence graph of an SCM[2]. We are interested in cases where the exact graph of a model can be pinpointed, i.e. cases in which it is possible to obtain identifiability. In fact, we can make restrictions on ANMs that ensure identifiability. As it turns out, these restrictions are rooted in whether the functions and densities of an ANM can solve a particular third order differential equation. As a first step to proving identifiability, we define the triple

$$\mathcal{T}_{i,j}^{\mathscr{C}} := \left(f_{i,j}, \partial \mathbb{P}_{X_j}, \partial \mathbb{P}_{N_i} \right).$$

[1] Strictly speaking, it is $\mathcal{G}^0$ that is identifiable from the implied distribution of $\mathscr{C}$.

[2] The PC algorithm perhaps being the prime example.

In the above, $\partial\mathbb{P}_{X_j}$ and $\partial\mathbb{P}_{N_i}$ are short-hand for the Radon-Nikodym derivatives $\frac{\partial P_{X_j}}{\partial m}$ and $\frac{\partial P_{N_i}}{\partial m}$ – we use this notation throughout the thesis, whenever we find it convenient. To increase readability, we use the notation

$$\begin{aligned}\xi_i(x) &:= \log \partial\mathbb{P}_{X_i}(x)\\ \nu_i(x) &:= \log \partial\mathbb{P}_{N_i}(x).\end{aligned}$$

For convenience, we will drop the function arguments and simply write ξ_i and ν_i.

Lemma 2.1 (Zhang and Hyvarinen 2012, Theorem 1 and Theorem 8, Peters et al. 2014, Theorem 20)
Let $\mathscr{C}$ be an ANM generated as in (2.1)*, and assume that $n = 2$, i.e. the graph is $\mathcal{G} := (X_1 \to X_2)$. Assume that $f_{2,1} \in C^3$ and that all densities are three times differentiable and strictly positive on $\mathbb{R}$. If $\mathcal{G}$ is not identifiable, the triple $\mathcal{T}^{\mathscr{C}}_{2,1}$ will satisfy the differential equation:*

$$\xi_1''' - \frac{\xi_1'' f_{2,1}''}{f_{2,1}'} = f_{2,1}' f_{2,1}'' \left(\frac{\nu_2' \nu_2'''}{\nu_2''} - 2\nu_2'' \right) + \nu_2' \left(f_{2,1}''' - \frac{(f_{2,1}'')^2}{f_{2,1}'} \right) - f_{2,1}' \xi_1'' \frac{\nu_2'''}{\nu_2''}. \tag{2.2}$$

Furthermore, if the set of points satisfying $\nu_2'' f_{2,1}' = 0$ is discrete, i.e. it has Lebesgue measure zero, every case in which $\mathcal{T}^{\mathscr{C}}_{2,1}$ satisfies (2.2) *are given in Table 2.1, setting* $(i, j) = (2, 1)$*.*

Table 2.1: *Every case in which $\mathscr{C}$ is not identifiable. See Appendix B.1 for definitions.*

	$\exp(\nu_i)$	$\exp(\xi_i)$	$f_{i,j}$
i)	Gaussian	Gaussian	Linear
ii)	log-mix-lin-exp	Log-mix-lin-exp	Linear
iii)	As above	One-sided asymptotically exponential, not log-mix-lin-exp	Strictly monotonic and $f_{i,j}'$ vanishes at both ∞ and $-\infty$
iv)	As above	Mixture of two exponentials	As above
v)	Mixture of two exponentials	Two-sided asymptotically exponential	As above

Peters et al. [2014] consider the cases in which $\mathscr{C}$ is identifiable when $n \geq 2$. As it turns out, identifiability in the multivariate case is not that different from in the bivariate case. Once again, whether a model is identifiable is linked to its solvability of a third order differential equation. We introduce the modified triple

$$\mathcal{T}^{\mathcal{E}|S=s}_{i,j} := (f_{i,j}, \partial\mathbb{P}_{X_j|(X)_S=s}, \partial\mathbb{P}_{N_i}), \tag{2.3}$$

and the notation

$$\xi_{i|S} := \log \partial\mathbb{P}_{X_i|(X)_S=s}.$$

That is, $\xi_{i|S}$ is the conditional log-density of X_i, having observed $(X)_S = s \in \mathbb{R}^{|S|}$, for some set $S \subset V(\mathcal{G}^0)$. Identifiability of $\mathscr{C}$ will now be determined by whether the triples on the form in (2.3) can solve the differential equation

$$\xi'''_{j|S} - \frac{\xi''_{j|S} f''_{i,j}}{f'_{i,j}} = f'_{i,j} f''_{i,j} \left(\frac{\nu'_i \nu'''_i}{\nu''_i} - 2\nu''_i \right) + \nu'_i \left(f'''_{i,j} - \frac{(f''_{i,j})^2}{f'_{i,j}} \right) - f'_{i,j} \xi''_{j|S} \frac{\nu'''_i}{\nu''_i}. \tag{2.4}$$

We state the following proposition:

Proposition 2.2 (Peters et al. 2014, Theorem 28)
Let $\mathscr{C}$ be an ANM generated as in (2.1)*. Assume that $\mathcal{F} \subseteq C^3 \cap \mathcal{C}^{\complement}$, and that all densities are strictly positive on $\mathbb{R}$ and three times differentiable. Assume that the set*

$$\bigcup_{k \in V(\mathcal{G}^0)} \left\{ x \in \mathbb{R} : \; \frac{d^2}{dx^2} \log \partial \mathbb{P}_{N_k}(x) = 0 \right\}$$

is discrete. For every two distinct nodes $i, j \in V(\mathcal{G})$, let

$$\mathbb{S}_{i,j} := \left\{ S \subsetneq V(\mathcal{G}^0) : \; \boldsymbol{PA}_{\mathcal{G}}(X_i) \backslash X_j \subseteq S \subseteq \boldsymbol{ND}_{\mathcal{G}}(X_i) \backslash X_j \right\},$$

and let

$$D := \{ i \in V(\mathcal{G}^0) : \boldsymbol{PA}_{\mathcal{G}}(X_i) \neq \emptyset \}.$$

If

$$\forall i \in D \;\forall\; j \in \boldsymbol{PA}_{\mathcal{G}}(X_i) \;\forall S \in \mathbb{S}_{i,j} \;\exists s \in \mathbb{R}^{|S|} : \quad T^{\mathcal{E}|S}_{i,j} \text{ does not solve equation } (2.4) \tag{C1}$$

then $\mathscr{C}$ is identifiable.

Proposition 2.2 is the extension of Lemma 2.1 to the multivariate case – indeed, in the bivariate case, Proposition 2.2 is implied by Lemma 2.1. In both cases, identifiability is determined by solvability of a third-order differential equation. The condition (C1) can, however, be difficult to grasp, and thus the reader might benefit from an explanation in words: Pick a node, X_i, and a parent, X_j, of X_i. The condition now states that for every set, S, that covers at least $\mathbf{PA}_{\mathcal{G}}(X_i) \backslash X_j$ and at most $\mathbf{ND}_{\mathcal{G}}(X_i) \backslash X_j$, there must exist an s such that $\mathcal{T}^{\mathscr{C}|S=s}_{i,j}$ does not solve (2.4). See also figure 2.1 for an illustration. Given a triple, $\mathcal{T}^{\mathscr{C}|S=s}_{i,j}$, we can simply read from table 2.1 whether it satisfies condition (C1). These triples are, however, highly intangible. Alternatively, we can make restrictions on $\mathcal{F}$ and the family of noise distributions which we consider in order to ensure identifiability. Before we state Theorem 2.3, an identifiability result for ANMs generated as in (2.1), we list here a set of assumptions which are sufficient to ensure identifiability. Let the ANM be as described in Section 2.1.

(A1) The class $\mathcal{F}$ is closed under addition, and closed in $\mathcal{L}^2(\mathbb{R}, \mathbb{B}, \mathbb{P}_{X_i})$ for every $i \in \mathcal{I}$ and $\mathcal{F} \subseteq C^3$. Furthermore, every $f \in \mathcal{F}$ is satisfies that it is not constant on any interval, nor is it linear on any interval.

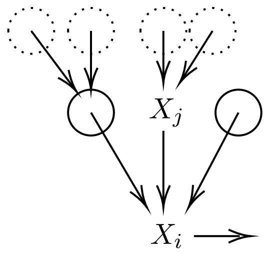

Figure 2.1: *An example of how S can be chosen in Proposition 2.2. Nodes with a solid circle are always conditioned on and nodes with dotted circles are conditioned on at some point. Nodes without circles are never conditioned on. Nodes different from X_j and X_i are left unlabeled, as their names are not relevant.*

(A2) For every $f \in \mathcal{F}$, *at least* one of of the following statements hold:

(A2.1) f is not strictly monotonic.

(A2.2) Either $\lim_{x \to -\infty} f'(x) \neq 0$ or $\lim_{x \to \infty} f'(x) \neq 0$.

(A3) The set

$$\bigcup_{k \in V(\mathcal{G}^0)} \left\{ x \in \mathbb{R} : \frac{d^2}{dx^2} \log \partial\mathbb{P}_{N_k}(x) = 0 \right\}$$

is discrete.

(A4) For every $i \in \mathcal{I}$ it holds that $\partial\mathbb{P}_{N_i}$ is three times differentiable and strictly positive on $\mathbb{R}$.

Whenever a model, $\mathscr{C}$, is such that the underlying functions and noise distributions satisfy an assumption above, we will simply say that $\mathscr{C}$ satisfies the assumption.

Theorem 2.3

Let $\mathscr{C}$ be an ANM generated as in (2.1) *and assume that $\mathscr{C}$ satisfies assumptions (A1) through (A4). Then $\mathscr{C}$ is identifiable.*

Proof. Fix $i \in V(\mathcal{G}^0)$, $j \in \mathbf{PA}_{\mathcal{G}}(X_i)$, $S \in \mathbb{S}_{i,j}$ and $s \in \mathbb{R}^{|S|}$. Every case in which $\mathcal{T}_{i,j}^{\mathscr{C}|S=s}$ solves (2.4) is given in Table 2.1 (replacing $\exp(\xi_i)$ with $\exp(\xi_{i|S})$). Cases (i) and (ii) are excluded by assumption (A1) and cases (iii) through (v) are excluded by assumption (A2). As i, j, S and s were arbitrary, it follows from Proposition 2.2 that $\mathscr{C}$ is identifiable. □

Theorem 2.3 is the foundation on which we build our algorithm for structural learning in Chapter 3. It should be remarked, however, that the conditions given for identifiability in Table 2.1 are sufficient but not necessary. Indeed, Peters and Bühlmann 2013b prove that an ANM with linear structural assignments and Gaussian noise can be identifiable, provided the noise variables all have equal variance. In fact, it remains identifiable even if the variances are not equal, but are all scalar products of the same unknown parameter [Peters and Bühlmann, 2013b, Theorem 1, Remark 4]. In other words, the implications stated in Lemma 2.1 and Proposition 2.2 are not bi-implications; it is possible that a triple as in (2.3) solves the differential equation (2.4) but that the underlying ANM is identifiable.

CHAPTER 3

A SCORE-BASED APPROACH FOR CAUSAL DISCOVERY

Having established when we can identify the graph of an ANM $\mathscr{C}$, we now turn our attention to *how* we can identify this graph. We introduce in Section 3.2 the notion of differential entropies, which are used to define the entropy score of a graph in Section 3.3. Motivated by the dimensional difficulties of scoring every single possible graph, we present in Section 3.4 the 'Greedy entropy-search' algorithm for causal discovery. In Section 3.5, we state without proof the main result of this thesis as well as the assumptions we make in order to prove it. We outline the proof details of Theorem 3.11 in Section 3.5. All of Section 3.6 is then spent on proving intermediary results that we will need to give a proof of Theorem 3.11. The proof is then given in 3.7. We then close the chapter with a few remarks on the Greedy entropy-search algorithm, as well as an example of an ANM on which it works.

3.1 ORIGINAL WORK AND CONTRIBUTIONS

The idea of performing a Greedy search by maximizing an entropy score originates from an unpublished manuscript, 'Estimating the Structure of Additive Noise Models with Greedy Search Algorithms', by Jonas Peters and Martin Wainwright. We cite this work as Peters and Wainwright [Unpublished]. All of Chapter 3 is based on this manuscript and contains both original work and results relayed directly from the manuscript. This section is intended to provide an overview of which results in chapter are new, which are adaptions or extensions, and which are simply relayed from Peters and Wainwright [Unpublished].

The definitions given in Definition 3.4 and Definition 3.8, as well as the proof of Theorem 3.7, which together forms the basis of this thesis, have been changed only in notation. Similarly, the proof of Lemma 3.17, which constitutes a key component in proving the main result of this thesis, is taken directly from Peters and Wainwright [Unpublished].

The proof idea of this thesis' main result, Theorem 3.11, was present in the manuscript and remains similar in overall structure, but has been altered in many ways internally. The main task of this thesis has been to fill in the gaps of this proof in order to give a complete proof.

Among the main original results we needed to give a complete proof, are Proposition 3.13, Proposition 3.15, Lemma 3.24 and Theorem 3.25. To the extent of our knowledge, both Lemma 3.24 and Theorem 3.25 are novel results. Furthermore, we give as a corollary an extension of Theorem 3.11 to a subclass of linear, non-Gaussian ANM. Other minor results are original work as well, but we do not list these here.

The algorithm proposed in Peters and Wainwright [Unpublished] was originally intended to work only for graphs which have no undirected cycles, i.e. polytrees. We extend this to a larger class of graphs, in which we allow some cycles. In particular, we extend the algorithm to work on undirected graphs in which all cycles have at least three colliders.

3.2 DIFFERENTIAL AND RELATIVE ENTROPY

Definition 3.1 (see e.g. Cover 2006, Chapter 8)**.** Let $\mathbb{P}$ and $\mathbb{Q}$ be measures on a measurable space, $(\mathcal{X}, \mathcal{A})$. Assume that $\mathbb{P}$ and $\mathbb{Q}$ are dominated by the same measure, μ, both with strictly positive densities on $\mathcal{X}$. The **differential entropy** of $\mathbb{P}$ is defined as

$$\mathbb{H}(\mathbb{P}) := -\int_{\mathcal{X}} \log \frac{\partial \mathbb{P}}{\partial \mu} \, d\mathbb{P}.$$

The **relative entropy** of $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$\mathrm{D}(\mathbb{P} \parallel \mathbb{Q}) := \int_{\mathbb{R}} \log \frac{\partial \mathbb{P}}{\partial \mathbb{Q}} \, d\mathbb{P} = -\mathbb{H}(\mathbb{P}) - \int_{\mathcal{X}} \log \frac{\partial \mathbb{Q}}{\partial \mu} \, d\mathbb{P}.$$

Whenever $\mathbb{P}$ and $\mathbb{Q}$ are probability measures and $(\mathcal{X}, \mathcal{A}) = (\mathbb{R}, \mathbb{B})$ and there are random variables, $X \sim \mathbb{P}$ and $Y \sim \mathbb{Q}$, we simply write $\mathbb{H}(X)$ and $\mathrm{D}(X \parallel Y)$.

Remark 3.2. The relative entropy is introduced here because of its close relation to differential entropy, but will not be used until the proof of Theorem 3.7.

The differential entropy originates from information theory, and can loosely be thought of as a measure of the amount of uncertainty in a stochastic system. Claude Shannon originally introduced the entropy only for discrete random variables. In his 1948 paper 'A Mathematical Theory of Communication' [Shannon, 2001], he showed that the entropy of a random variable is minimized and equal to 0 if the variable is degenerate, and maximized when the random variable is uniform on a finite set Shannon [2001, pp. 12–13]. These results resonate well with the notion of entropy measuring the uncertainty of a system. The differential entropy of a continuous random variable is not quite as well-behaved as in the discrete case. In particular, the differential entropy can be both negative and infinite. The case of infinite entropies is potentially troublesome when we start scoring graphs, as we will be calculating a sum of entropies. It is, however, possible to find explicit bounds on the differential entropy for variables with finite second moment.

Proposition 3.3 (Properties of differential entropy)
Let $X \sim f \cdot m$ be a random variable with finite second moment and assume that f is strictly positive on $\mathbb{R}$. The following statements hold:

i) For every $c \in \mathbb{R}$

$$\mathbb{H}(X + c) = \mathbb{H}(X).$$

ii) The entropy is bounded above by $\frac{1}{2}\log(2\pi e \mathbb{V}X)$, with equality if and only if X is normally distributed. The norm $\|\cdot\|_\infty$ denotes the $\mathcal{L}^\infty$ norm.

iii) The entropy is bounded below by $-\log(\|f\|_\infty)$.

iv) If $Y \sim g \cdot m$ is another random variable with second moment and independent from X, the entropy of $X + Y$ is bounded below by $-\log(\|f * g\|_\infty)$ and above by $\frac{1}{2}\log(2\pi e[\mathbb{V}X + \mathbb{V}Y + 2\operatorname{cov}(X,Y)])$. The symbol $*$ denotes convolution.

Proof. i) The result is immediate from translation invariance of the Lebesgue measure and given in most texts on information theory (see e.g. Cover 2006, Theorem 8.6.3).
ii) See Cover 2006, Theorem 8.6.5.
iii) Denote by $(f > 1)$ the set $\{x \in \mathbb{R} : f(x) > 1\}$ and suppose first that $m(f > 1) > 0$. Then

$$\mathbb{H}(X) = -\int_{\mathbb{R}} f \cdot \log(f)\ dm = \underbrace{\int_{(f\leq 1)} -f \cdot \log(f)\ dm}_{>0} + \underbrace{\int_{(f>1)} -f \cdot \log(f)\ dm}_{<0}.$$

In order for $\mathbb{H}(X) = -\infty$, we would need $\int_{(f>1)} f \cdot \log(f)\ dm = \infty$. However, as X has full support, f must be essentially bounded, i.e. $f \in \mathcal{L}^\infty$. Thus,

$$\int_{(f>1)} f \cdot \log(f)\ dm \leq \log(\|f\|_\infty) \int_{(f>1)} f\ dm < \log(\|f\|_\infty) < \infty.$$

Therefore, we conclude that

$$\mathbb{H}(X) > -\log(\|f\|_\infty) > -\infty.$$

If $m(f > 1) = 0$ we see that

$$\mathbb{H}(X) = -\int_{(f\leq 1)} f \log(f)\ dm \geq -\log(\|f\|_\infty) \int_{(f\leq 1)} f\ dm = -\log(\|f\|_\infty) > -\infty.$$

iv) Noting that $X + Y \sim f * g \cdot m$ (see e.g. Hansen 2009, Corollary 20.9) and that $f * g \in \mathcal{L}^\infty$, the proof is immediate from (ii) and (iii). This concludes the proof. □

3.3 THE ENTROPY SCORE

We are now ready to introduce the entropy score of a graph. Since this will require calculating entropies, we will need these to be well-defined and finite. On the basis of Proposition 3.3, we state two additional assumptions to ensure that all entropies are finite:

(A5) For every $i \in \mathcal{I}$ it holds that $\mathbb{E}N_i^2 < \infty$.

(A6) For every $f \in \mathcal{F}$ and every $i \in \mathcal{I}$ it holds that $\mathbb{E}f(N_i)^2 < \infty$.

Notation. In what remains of this thesis, we will routinely alternate between graphical arguments, i.e. ones that are pertaining to the graph $\mathcal{G}$, and probabilistic arguments, i.e. ones that are related to the random variables $(X)_{V(\mathcal{G})}$ in a model $\mathscr{C}$. To emphasize which are which – but also in an effort to increase readability – we strive to use Greek letters for the nodes whenever the argument is a graphical one, and write X with Greek subscripts whenever the argument is a probabilistic one. That is, we may, for example, take a sum over $\alpha \in \mathbf{PA}_{\mathcal{G}}(\beta)$, but we write the summands as $f_{\beta,\alpha}(X_\alpha)$.

Definition 3.4 (Peters and Wainwright [Unpublished])**.** Let $\mathscr{C}$ be an ANM generated as in (2.1) with DAG $\mathcal{G}$ and implied distribution $\mathbb{P}$. We define the entropy score of $\mathcal{G}$ to be

$$\ell(\mathcal{G}) := - \sum_{\nu \in V(\mathcal{G})} \mathbb{H}\left(X_\nu - \sum_{\gamma \in \mathbf{PA}_{\mathcal{G}}(\nu)} \hat{f}_{\nu,\gamma}(X_\gamma) \right),$$

where

$$\forall \nu \in V(\mathcal{G}) : \quad (\hat{f}_{\nu,\gamma})_{\gamma \in \mathbf{PA}_{\mathcal{G}}(\nu)} := \underset{(f)_{\mathbf{PA}_{\mathcal{G}}(\nu)} \in \mathcal{F} \cup \mathcal{C}}{\arg\min} \; \mathbb{E}\left(X_\nu - \sum_{\gamma \in \mathbf{PA}_{\mathcal{G}}(\nu)} f_{\nu,\gamma}(X_\gamma) \right)^2, \tag{3.1}$$

and $\mathcal{C}$ is the set of all constant functions with domain $\mathbb{R}$. In the case of source nodes, i.e. a node κ such that $\mathbf{PA}_{\mathcal{G}}(X_k) = \emptyset$, we use the convention that[1]

$$\sum_{\gamma \in \mathbf{PA}_{\mathcal{G}}(\kappa)} \hat{f}_{\kappa,\gamma}(X_\gamma) = 0.$$

We define the Gaussian score of $\mathcal{G}$ to be

$$\ell^g(\mathcal{G}) := - \sum_{\nu \in V(\mathcal{G})} \log\left(\mathbb{V}\left(X_\nu - \sum_{\gamma \in \mathbf{PA}_{\mathcal{G}}(\nu)} \hat{f}_{\nu,\gamma}(X_\gamma) \right) \right). \tag{3.2}$$

Finding the entropy score of a graph amounts to minimizing the $\mathcal{L}^2$-loss of each node given its parents, for example through regression, and then calculating the entropy of the resulting residuals, and similarly for the Gaussian score. Note also that the minimization problem in (3.1) is indeed well-defined, as we have assumed $\mathcal{F}$ to be closed in $\mathcal{L}^2$. We introduce the Gaussian score here but we will not use it until Section 3.4. However, the reader may observe that the Gaussian score is, in fact, a special case of the entropy score. Indeed, we saw in Proposition 3.3 point (ii), that the entropy score of a random variable, say X, can be bounded above by $\frac{1}{2}\log(2\pi e \mathbb{V} X)$ and that this bound is attained only when X is normally distributed. If we consider the difference in entropy between two normally distributed random variables the term $2\pi e$ cancels out and $\frac{1}{2}$ becomes a common scalar. Thus, we get a term as the one in equation (3.2). Therefore, we can think of the Gaussian score as a 'worst-case-scenario' entropy score. That is, the Gaussian score tells us what the entropy *would* have been, had every summand been normally distributed. As a consequence, when we let the noise variables in an ANM follow a normal distribution, the entropy score of the *true* residuals – i.e. when the regression is perfectly accurate – becomes equivalent with the Gaussian score.

Notation. In the discussion of minimizing $\mathcal{L}^2$-loss it is next to impossible not to mention conditional expectations. To avoid possible confusion, we stress here the notation which we use when writing conditional expectations. Let X and Y be random variables on the common probability

[1] In fact, we could set the sum to be any constant as both variances and entropies are translation invariant.

space $(\mathbb{R}, \mathbb{B}, \mathbb{P})$ and assume that $\mathbb{E}|X|< \infty$. The **conditional expectation** of X given the σ-algebra generated by Y is a $\sigma(Y)$-measurable random variable, written $\mathbb{E}[X \mid \sigma(Y)]$, which satisfies $\mathbb{E}|\mathbb{E}[X \mid \sigma(Y)]|< \infty$ and

$$\forall A \in \sigma(Y): \qquad \int_A \mathbb{E}[X \mid \sigma(Y)] \, d\mathbb{P} = \int_A X \, d\mathbb{P}.$$

For short, we write $\mathbb{E}[X \mid Y]$ in place of $\mathbb{E}[X \mid \sigma(Y)]$. By the Doob-Dynkin Lemma [Sokol and Rønn-Nielsen, 2016, Lemma 4.3.2], there exists a measurable mapping, $\phi_{X|Y}$, such that

$$\mathbb{E}[X \mid Y] = \phi_{X|Y} \circ Y.$$

This is often written as $\phi(y) = \mathbb{E}[X \mid Y = y]$. When we talk about conditional expectations as random variables, we write $\mathbb{E}[X \mid Y]$. When we talk about the underlying function, we write $\phi_{X|Y}$. If there is no danger of confusing which algebra we are conditioning on, we simply write ϕ in place of $\phi_{X|Y}$.

Proposition 3.5 (See e.g. Györfi et al. 2002, page 2)
Let X and Y be $\mathbb{R}$ and $\mathbb{R}^d$-valued random variables respectively. Assume that X and Y has finite second moment and let $\mathcal{M}^d$ be the space of measurable functions $f : \mathbb{R}^d \to \mathbb{R}$. It then holds that

$$\arg\min_{f \in \mathcal{M}^d} \mathbb{E}(X - f(Y))^2 = \phi_{X|Y}, \tag{3.3}$$

Proof. First note that $\mathcal{L}^2$ is a Hilbert space with inner product $\langle X, Y \rangle = \mathbb{E}(X \cdot Y)$ [Encyclopedia of Mathematics., a]. The orthogonal projection of X onto the space of $\sigma(Y)$-measurable functions is $\mathbb{E}[X \mid Y]$, as easily seen by

$$\langle X - \mathbb{E}[X \mid Y], Y \rangle = \mathbb{E}(X \cdot Y) - \mathbb{E}(Y \cdot \mathbb{E}[X \mid Y]) = \mathbb{E}(X \cdot Y) - \mathbb{E}(X \cdot Y) = 0.$$

It then follows from the Hilbert projection theorem (see e.g. Rudin 1987, Theorem 4.11) that the conditional expectation of X given Y exists and satisfies.

$$\min_{f \in \mathcal{M}^d} \mathbb{E}(X - f(Y))^2 = \mathbb{E}[X \mid Y]. \tag{3.4}$$

Thus, we conclude that the function $f \in \mathcal{M}^d$ that satisfies this must be $\phi_{X|Y}$. □

Proposition 3.5 is not a new result by any means – in fact, it is simply a characterization of the solution to any ordinary least squares minimization problem. However, we will make use of it so often throughout this thesis, that it is worth stating.

Remark 3.6. In the context of entropy scores, Proposition 3.5 can be used to characterize the solution to the minimization problem posed in Definition 3.4, whenever a node, say X, decomposes into a sum of terms that are either measurable in the algebra generated by its $\mathcal{G}$-parents or independent of them. Indeed, if the graph $\mathcal{G}$ is the true graph, then we can write

$$\begin{aligned}\mathbb{E}[X \mid \mathbf{PA}_{\mathcal{G}}(X)] &= \mathbb{E}\left[\sum_{\gamma \in \mathbf{PA}_{\mathcal{G}}(X)} f_\gamma(X_\gamma) + N \,\middle|\, \mathbf{PA}_{\mathcal{G}}(X)\right] \\ &= \sum_{\gamma \in \mathbf{PA}_{\mathcal{G}}(X)} f_\gamma(X_\gamma) + \mathbb{E}N.\end{aligned}$$

The function ϕ as described in Proposition 3.5 is then simply a linear combination of elements in $\mathcal{F}$ and therefore an element of $\mathcal{F}$ itself. That is, the minimum is attained in $\mathcal{F}$ and Proposition 3.5 then states that

$$\mathbb{E}\left(X - \sum_{\gamma \in \mathbf{PA}_{\mathcal{G}}(X)} \hat{f}_\gamma(X_\gamma)\right)^2 = \mathbb{E}\left(X - \mathbb{E}[X \mid \mathbf{PA}_{\mathcal{G}}(X)]\right).$$

A similar argument can be carried out when $\mathcal{G}$ is not the true graph, but a subgraph of it. However, this requires some assumptions on the graph that ensure a particular independence structure. We refrain from giving this argument for now but return to it in the proof of Proposition 3.13.

We now give a proof that the entropy score under an ANM that satisfies assumptions (A1) through (A6) will, in fact, attain its maximum in the true graph of the model.

Theorem 3.7 (Adapted from Peters and Wainwright [Unpublished])
Suppose that $\mathscr{C}$ is an ANM generated according to (2.1) *with graph $\mathcal{G}^0$ and functions $\boldsymbol{f}^0$ and let $\mathbb{P} := \mathbb{Q}(\mathcal{G}^0, \boldsymbol{f}^0)$. Assume that $\mathscr{C}$ satisfies assumptions (A1) through (A6). Then*

$$\arg\max_{\mathcal{G}} \ell(\mathcal{G}) = \mathcal{G}^0.$$

Proof. Let $V := V(\mathcal{G}^0)$ and let $\mathcal{G}$ be any DAG with $V(\mathcal{G}) = V$. Throughout this proof, all expectations are taken under $\mathbb{P}$. Observe that

$$\forall \nu \in V(\mathcal{G}): \qquad \left|\mathbb{H}\left(X_\nu - \sum_{\gamma \in \mathbf{PA}_{\mathcal{G}}(\nu)} f^0_{\nu,\gamma}(X_\nu)\right)\right| < \infty,$$

as shown in Proposition 3.3. To prove that the entropy score is maximized in $\mathcal{G}^0$, we consider the relative entropy between $\mathbb{P}$ and the $\mathbb{Q} \in \mathcal{S}^{\mathcal{F}}_{\mathcal{G}}$ that is 'closest' to $\mathbb{P}$ in relative entropy:[2]

$$\begin{aligned}-\inf_{\mathbb{Q} \in \mathcal{S}^{\mathcal{F}}_{\mathcal{G}}} \mathrm{D}(\mathbb{P} \parallel \mathbb{Q}) &= -\inf_{\mathbb{Q} \in \mathcal{S}^{\mathcal{F}}_{\mathcal{G}}} \mathbb{E}_{\mathbb{P}} \log\left(\frac{\partial \mathbb{P}((X)_V)}{\partial \mathbb{Q}((X)_V)}\right) \\ &= -\inf_{\mathbb{Q} \in \mathcal{S}^{\mathcal{F}}_{\mathcal{G}}} \left\{\mathbb{E}\log \partial\mathbb{P}((X)_V) - \mathbb{E}\log\partial\mathbb{Q}((X)_V)\right\} \\ &\overset{(a)}{=} -\mathbb{E}\log\partial\mathbb{P}((X)_V) - \inf_{\mathbb{Q} \in \mathcal{S}^{\mathcal{F}}_{\mathcal{G}}} \left\{-\mathbb{E}\log\partial\mathbb{Q}((X)_V)\right\}. \qquad (3.5)\end{aligned}$$

[2] The word 'closest' being in quotation marks because the relative entropy is not a metric.

Above, equality (a) follows from $\partial\mathbb{P}$ being constant over $\mathcal{S}_{\mathcal{G}}^{\mathcal{F}}$. The left-hand term in (3.5) can be rewritten as:

$$\begin{aligned}
-\mathbb{E}\log\partial\mathbb{P}((X)_V) &\overset{(b)}{=} -\mathbb{E}\log\left(\prod_{\nu\in V(\mathcal{G}^0)}\partial\mathbb{P}_{X_\nu}(X_\nu \mid \mathbf{PA}_{\mathcal{G}^0}(X_\nu))\right) \\
&= -\sum_{\nu\in V(\mathcal{G}^0)}\mathbb{E}\log\partial\mathbb{P}_{X_\nu}(X_\nu \mid \mathbf{PA}_{\mathcal{G}^0}(X_\nu)) \\
&\overset{(c)}{=} -\sum_{\nu\in V(\mathcal{G}^0)}\mathbb{E}\log\partial\mathbb{P}_{N_\nu}\left(X_\nu - \sum_{\gamma\in\mathbf{PA}_{\mathcal{G}^0}(\nu)} f^0_{\nu,\gamma}(X_\gamma)\right) \\
&= \sum_{\nu\in V(\mathcal{G}^0)}\mathbb{H}\left(X_\nu - \sum_{\gamma\in\mathbf{PA}_{\mathcal{G}^0}(X_\nu)} f^0_{\nu,\gamma}(X_\nu)\right) \\
&\overset{(d)}{=} \sum_{\nu\in V(\mathcal{G}^0)}\mathbb{H}\left(X_\nu - \sum_{\gamma\in\mathbf{PA}_{\mathcal{G}^0}(\nu)} \hat{f}_{\nu,\gamma}(X_\nu)\right) = -\ell(\mathcal{G}^0), \qquad (3.6)
\end{aligned}$$

where equality (b) follows from the Markov factorization property[3] and equality (c) follows from using transformation of densities (see e.g. Hansen 2009, Theorem 15.1). Equality (d) follows from Proposition 3.5 by applying the argument given in Remark 3.6 and from using Proposition 3.3 part (i). Next, we consider the right-hand term of equation (3.5). Analogous to the calculations leading to equation (3.6), we find that

$$\begin{aligned}
-\inf_{\mathbb{Q}\in\mathcal{S}_{\mathcal{G}}^{\mathcal{F}}}\left\{-\mathbb{E}\log\partial\mathbb{Q}(\mathbf{X})\right\} &= -\inf_{\mathbb{Q}\in\mathcal{S}_{\mathcal{G}}^{\mathcal{F}}}\left\{\sum_{\nu\in V(\mathcal{G}^0)}\mathbb{H}\left(X_\nu - \sum_{\gamma\in\mathbf{PA}_{\mathcal{G}}(\nu)} f^0_{\nu,\gamma}(X_\nu)\right)\right\} \\
&\geq -\sum_{\nu\in V(\mathcal{G}^0)}\mathbb{H}\left(X_\nu - \sum_{\gamma\in\mathbf{PA}_{\mathcal{G}}(\gamma)} \hat{f}_{\nu,\gamma}(X_\nu)\right) \\
&= \ell(\mathcal{G}). \qquad (3.7)
\end{aligned}$$

Observe that equations (3.6) and (3.7) together imply that the relative entropy is indeed well-defined and finite, as all the entropy terms are finite by assumption. Using the fact that relative entropies are weakly positive [Cover, 2006, theorem 8.6.1] and by combining equations (3.5), (3.6) and (3.7) we obtain that

$$0 \geq -\mathrm{D}(\mathbb{P} \parallel \mathbb{Q}) \geq \ell(\mathcal{G}) - \ell(\mathcal{G}^0).$$

Then

$$\ell(\mathcal{G}^0) \geq \ell(\mathcal{G}),$$

[3] Recall from Chapter 1 that this holds in any SCM which has density with respect to a product of Lebesgue measures.

for any graph $\mathcal{G}$. The relative entropy is zero if and only if $\mathbb{P} = \mathbb{Q}$ Cover [2006, theorem 8.6.1]. By Theorem 2.3 it then follows that there is equality only when $\mathcal{G} = \mathcal{G}^0$.

□

The above result shows that, given increasingly large samples of data from an ANM, if we can consistently estimate graph scores, then we can consistently recover the graph that generated the data. However, its practical applications are limited, as the cardinality of the space of graphs that must be searched grows super-exponentially in the number of vertices. In fact, the number of DAGs with n vertices can be shown to satisfy the recurrence relation

$$R_n = \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} 2^{k(n-k)} R_{n-k},$$

with $R_0 = 1$ [McKay et al., 2003, pp. 3]. Even if we were able to score graphs at a rate of one billion per second, a graph with only 12 vertices would still take 16.5 billion years to score – a mere three billion years more than the currently estimated age of the universe. In other words; scoring every single possible graph is not feasible.

3.4 GREEDILY SEARCHING THE GRAPH SPACE

Instead of searching the entire space of graphs, we make use of the fact that we have defined the score to be additive in each noise term to construct a Greedy search method – That is, one in which a graph is constructed in a stepwise manner, each time adding edges that are locally optimal – to reduce the computational complexity. Denote by $\mathcal{G}_{\alpha\to\beta}$ a graph with $V(\mathcal{G}_{\alpha\to\beta}) = V(\mathcal{G})$ and $E(\mathcal{G}_{\alpha\to\beta}) = E(\mathcal{G}) \cup (\alpha \to \beta)$, and define

$$\Delta\ell\left(\mathcal{G}, \alpha \to \beta\right) := \ell(\mathcal{G}_{\alpha\to\beta}) - \ell(\mathcal{G}) \tag{3.8}$$

and

$$\Delta\ell^{\mathrm{g}}\left(\mathcal{G}, \alpha \to \beta\right) := \ell^{g}(\mathcal{G}_{\alpha\to\beta}) - \ell^{g}(\mathcal{G}). \tag{3.9}$$

Writing out the terms of $\Delta\ell$ and $\Delta\ell^{g}$, we obtain

$$\Delta\ell\left(\mathcal{G}, \alpha \to \beta\right) = \mathbb{H}\left(X_\beta - \sum_{\gamma\in\mathbf{PA}_{\mathcal{G}}(\beta)} \hat{f}_{\alpha,\gamma}(X_\gamma)\right) - \mathbb{H}\left(X_\beta - \sum_{\gamma\in\mathbf{PA}_{\mathcal{G}}(\beta)} \hat{f}_{\alpha,\gamma}(X_\gamma) - \hat{f}_{\alpha,\beta}(X_\beta)\right),$$

$$\Delta\ell^{\mathrm{g}}\left(\mathcal{G}, \alpha \to \beta\right) = \log\mathbb{V}\left(X_\beta - \sum_{\gamma\in\mathbf{PA}_{\mathcal{G}}(\beta)} \hat{f}_{\alpha,\gamma}(X_\gamma)\right) - \log\mathbb{V}\left(X_\beta - \sum_{\gamma\in\mathbf{PA}_{\mathcal{G}}(\beta)} \hat{f}_{\alpha,\gamma}(X_\gamma) - \hat{f}_{\alpha,\beta}(X_\beta)\right).$$

We refer to $\Delta\ell$ and $\Delta\ell^{g}$ as the *marginal* entropy and Gaussian scores, respectively. Finally, we let

$$R(\mathcal{G}) := \left\{\alpha, \beta \in V(\mathcal{G}) : \ (\alpha \to \beta) \notin \mathcal{G} \ \wedge \ \mathcal{G}_{\alpha\to\beta} \text{ is a DAG}\right\}. \tag{3.10}$$

That is, $R(\mathcal{G})$ is the set of edges that can be added to $\mathcal{G}$ without introducing any directed cycles. We then define the **Greedy entropy-search** for graph learning:

Definition 3.8 (Peters and Wainwright [Unpublished])**.** Let $\mathscr{C}$ be an ANM with graph $\mathcal{G}^0$ and implied distribution $\mathbb{P}$. Let $\mathcal{G}$ be any subgraph of $\mathcal{G}^0$. The Greedy Entropy-Search algorithm updates the graph $\mathcal{G}$ iteratively as described in Algorithm 1 below. We use the notation $\mathcal{G}^{s=k}$ to denote the graph at step number k in the Greedy entropy-search.

Algorithm 1: The Greedy Entropy-Search

```
Input: $\mathcal{G}$, $\mathbb{P}$
$k := 0$
$\mathcal{G}^{s=k} := \mathcal{G}$
$(\alpha \to \beta) := \underset{(\kappa \to \delta) \in R(\mathcal{G}^{s=k})}{\arg\max} \Delta\ell^{\mathrm{g}} \left(\mathcal{G}^{s=k}, \kappa \to \delta\right)$
while $\Delta\ell^{g} \left(\mathcal{G}^{s=k}, \alpha \to \beta\right) > 0$ do
    $(\tilde{\alpha} \to \tilde{\beta}) := \underset{(\kappa \to \delta) \in \{(\alpha \to \beta), (\beta \to \alpha)\}}{\arg\max} \Delta\ell \left(\mathcal{G}^{s=k}, \kappa \to \delta\right)$
    $\mathcal{G}^{s=k+1} := \mathcal{G}^{s=k}_{\tilde{\alpha} \to \tilde{\beta}}$
    $k := k + 1$
    $(\alpha \to \beta) := \underset{(\kappa \to \delta) \in R(\mathcal{G}^{s=k})}{\arg\max} \Delta\ell^{\mathrm{g}} \left(\mathcal{G}^{s=k}, \kappa \to \delta\right)$
return $\mathcal{G}^{s=k}$
```

In words, the Greedy entropy-search takes a graph and a joint distribution as input. We first identify the edge that maximizes the marginal Gaussian score, but direct it by maximizing the marginal entropy score. That is, we use the Gaussian score for identification of edges, but not for directing them. This process is repeated until the Gaussian score cannot increase any further. At this point, the reader might wonder why we use both the entropy score and the Gaussian score instead of simply using the entropy score; we have already proven that the entropy score is maximized by the true graph and so it might seem more natural to use only the entropy score. In fact, the reason that we use the Gaussian score for identifying edges is motivated purely by aspirations of proving that the Greedy entropy-search can recover the graph of an ANM. On the path to proving this result, we need to make use of Jensen's inequality and since we do not have a similar inequality available for entropies, we make use of the Gaussian score instead. Conversely, it does not suffice to use *only* the Gaussian scores when greedily searching the graph space. Indeed, Peters and Wainwright [Unpublished] construct an example in which the Gaussian score cannot distinguish the correct direction of edges, while the entropy score can. This pragmatic approach does, however, carry a benefit with it when we move away from the population case and into a finite-sample setting. Generally speaking, it is very difficult to estimate entropies, but relatively easy to estimate variances and so the burden of estimation lessens.

The Greedy entropy-search acts in a locally optimal manner, in the sense the we always add the edge which provides the largest increase in score at a given step. The goal of this chapter is to prove Theorem 3.11, which states that the Greedy entropy-search is – under additional assumptions

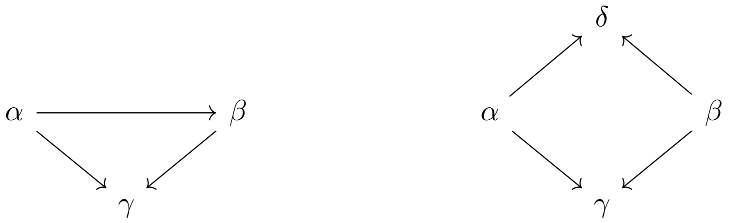

Figure 3.1: *Examples of graphs that we do not consider in our Greedy entropy-search. Left: A triangular structure. Right: A diamond structure.*

– *globally* optimal, in the sense that it recovers the true graph of an ANM. To prove Theorem 3.11 we need the independence structure of the random variables in the model we are considering to be 'sufficiently nice'. Loosely speaking, we can only prove optimality of the Greedy entropy-search when the model is such that the parents of every node are mutually independent. That is, we do not allow for two causes of a variable to be effects of the same cause, nor do we allow them to be causes of one another. Informally, this excludes graphs which have a triangular structure[4]. Furthermore, we do not allow for two nodes to both be direct causes of the same two effects. This restricts the class of graphs to not have a diamond-like structure. See Figure 3.1 for an example of such structures. We introduce first this new class of graphs and prove an alternative characterization of them after which we state Theorem 3.11 and give an outline of its proof. We then state and prove every intermediary result which will be used in the proof of Theorem 3.11.

Unrelated graphs

Definition 3.9 (Unrelated graphs)**.** Let $\mathcal{G}$ be a DAG. We say that $\mathcal{G}$ has *unrelated parents* – or simply that $\mathcal{G}$ is unrelated – if the following two statements hold for every node $\kappa \in V(\mathcal{G})$ that has at least two parents:

(i) Every disjoint partition of κ's parents are d-separated. That is

$$\forall \Omega \subsetneq \mathbf{PA}_{\mathcal{G}}(\kappa): \qquad \Omega \perp_d \mathbf{PA}_{\mathcal{G}}(\kappa) \backslash \Omega.$$

(ii) For every pair of distinct nodes $\alpha, \beta \in \mathbf{PA}_{\mathcal{G}}(\kappa)$, the subgraph induced by $\mathbf{DE}_{\mathcal{G}}(\alpha) \cup \mathbf{DE}_{\mathcal{G}}(\beta)$, denoted $\mathcal{G}_{\mathbf{DE}_{\mathcal{G}}(\alpha) \cup \mathbf{DE}_{\mathcal{G}}(\beta)}$, has no undirected cycles in its skeleton.

The name 'unrelated' is inspired by the fact that no two parents have any common relation apart from their common child. Before moving on, we give an alternative characterization of unrelated graphs, which we rely heavily on when we prove optimality of the Greedy entropy-search.

Lemma 3.10
Let $\mathcal{G}$ be a DAG and let κ be an arbitrary node in $\mathcal{G}$. The following statements are equivalent:

[4] In graph terminology this amounts to excluding *chordal* graphs.

(i) $\mathcal{G}$ is unrelated.

(ii) Every cycle in $\mathcal{G}$ has at least three colliders.

Proof. $(i) \Rightarrow (ii)$: Suppose $\mathcal{G}$ is unrelated and fix $\omega, \theta \in \mathbf{PA}_{\mathcal{G}}(\kappa)$ such that $\omega \neq \theta$. Let $\tilde{\mathcal{G}} := \mathcal{G}_{\mathbf{DE}_{\mathcal{G}}(\omega) \cup \mathbf{DE}_{\mathcal{G}}(\theta)}$. By assumption, $\omega \perp_d \theta$ in $\mathcal{G}$ and thus every path between ω and θ must be blocked by a collider. Furthermore, there are no colliders in the skeleton of $\tilde{\mathcal{G}}$. If there is a cycle through ω and θ that traverses their children, this cycle must then traverse at least three colliders. Suppose then for contradiction the existence of a cycle through ω and θ that only has two colliders. One of these colliders is κ, the common child of ω and θ. Let δ be the other collider on this cycle. The node δ must then have parents, ρ_1 and ρ_2 that are ancestors of ω and θ respectively. But then the subgraph consisting of $\mathbf{DE}_{\mathcal{G}}(\rho_1)$ and $\mathbf{DE}_{\mathcal{G}}(\rho_2)$ must contain a cycle

$$(\delta \leftarrow \rho_1 \rightarrow \ldots \rightarrow \omega \rightarrow \kappa \leftarrow \theta \leftarrow \ldots \leftarrow \rho_2 \rightarrow \delta)$$

and so there is an undirected cycle in the skeleton of this graph, which is a contradiction. We then conclude that every cycle in $\mathcal{G}$ must contain at least three colliders.

$(ii) \Rightarrow (i)$: Let $\Omega \subsetneq \mathbf{PA}_{\mathcal{G}}(\kappa)$ and $\Theta := \mathbf{PA}_{\mathcal{G}}(\kappa) \backslash \Omega$ and fix $\theta \in \Theta$ and $\omega \in \Omega$. As θ and ω are both parents of k, they cannot be directly connected, as this would introduce a cycle $(\omega \rightarrow \kappa \leftarrow \theta \rightarrow \omega)$ with only one collider. We consider then any path $\epsilon := (\omega, \ldots, \theta)$ between θ and ω that does not traverse κ. If no such path exists, we have $\theta \perp_d \omega$ as κ is a collider. If such a path exists, it follows that the extended path, $\tilde{\epsilon} := (\epsilon \rightarrow \kappa \leftarrow \omega)$ is a cycle. By assumption, $\tilde{\epsilon}$ must traverse at least three colliders and as one such collider is κ, it follows that ϵ must traverse at least one collider and thus ϵ is blocked. As we chose ϵ arbitrarily, it follows that $\theta \perp_d \omega$. As θ and ω were chosen arbitrarily, it follows that

$$\Theta \perp_d \Omega. \tag{3.11}$$

Choose now any two distinct nodes in $\mathcal{G}$ and denote these[5] by α and β. Let $\tilde{\mathcal{G}}$ be the subgraph consisting of all descendants of α and β. If α and β have no common descendants in $\mathcal{G}$, then $\tilde{\mathcal{G}}$ cannot have any undirected cycles in its skeleton. If it did, either α or β would be a descendant of itself, which would imply the existence of a directed cycle in $\mathcal{G}$. By the same argument, if β is a descendant of α or vice versa, there cannot be an undirected cycle in $\text{ske}(\tilde{\mathcal{G}})$, as this would imply a directed cycle in $\mathcal{G}$. Suppose then that α and β share a common descendant, δ_1, and assume for contradiction that there is an undirected cycle in $\text{ske}(\tilde{\mathcal{G}})$. This implies that they also share a second descendant, δ_2, that is not a descendant of δ_1. We can then find a cycle with only two colliders

$$(\alpha \rightarrow \ldots \rightarrow \delta_1 \leftarrow \ldots \leftarrow \beta \rightarrow \ldots \rightarrow \delta_2 \leftarrow \ldots \leftarrow \alpha),$$

which contradicts the assumption that every cycle in $\mathcal{G}$ has at least three colliders. As α and β were chosen arbitrarily, we conclude that (i) follows from (ii).

□

[5] We denote these differently than θ and ω as they are allowed to be different.

3.5 OPTIMALITY OF THE GREEDY-ENTROPY SEARCH – WITHOUT PROOF

Before we state this chapters main result, Theorem 3.11, we list here sufficient conditions for it to hold. These are largely similar to the assumptions made earlier, (A1) through (A6). Assumptions (B1), (B2), (B5) and (B6) are identical to their respective A-versions. Assumption (B3) is a stronger version of (A4), which implies both assumption (A4) and (A3). Assumptions (B4) and (B7) are new. The reason for making further assumptions are discussed in Section 3.5 and Section 3.6. We suggest that the reader does delve in too deeply in the added assumptions just yet, as we will attempt to make them clear in Section 3.6.

For an ANM $\mathscr{C}$ with graph $\mathcal{G}^0$, we assume that:

(B1) The class $\mathcal{F}$ is closed under addition, closed in $\mathcal{L}^2(\mathbb{R}, \mathbb{B}, \mathbb{P}_{X_\nu})$ for every $\nu \in V(\mathcal{G})$, and that $\mathcal{F} \subseteq C^3$. Furthermore, every $f \in \mathcal{F}$ satisfies that it is not constant on any interval, nor is it linear on any interval.

(B2) For every $f \in \mathcal{F}$, *at least* one of of the following statements hold true:

(B2.1) f is not strictly monotonic.

(B2.2) Either $\lim_{x \to -\infty} f'(x) \neq 0$ or $\lim_{x \to \infty} f'(x) \neq 0$.

(B3) For every $\nu \in V(\mathcal{G})$ it holds that $\partial \mathbb{P}_{N_\nu} \in C^\omega_+$, where C^ω_+ is the class of all strictly positive real analytic functions.

(B4) For every $f \in \mathcal{F}$, there exists a countable partition of $\mathbb{R}$, i.e. a pairwise disjoint sequence of sets that covers $\mathbb{R}$, such that f is locally bijective on each set.

(B5) For every $\nu \in V(\mathcal{G})$ it holds that $\mathbb{E}N_\nu^2 < \infty$.

(B6) For every $f \in \mathcal{F}$ and every $\nu \in V(\mathcal{G})$ it holds that $\mathbb{E}f(N_\nu)^2 < \infty$.

(B7) For every $f, g \in \mathcal{F}$ and every $\nu \in V(\mathcal{G})$, the mapping $x \mapsto \mathbb{E}f(g(x) + N_\nu)$ belongs to $\mathcal{F}$.

Finally, we redefine $R(\mathcal{G})$ from (3.10) to reflect that we will now only consider unrelated graphs:

$$R(\mathcal{G}) := \left\{\alpha, \beta \in V(\mathcal{G}) : \ (\alpha \to \beta) \notin \mathcal{G} \ \wedge \ \mathcal{G}_{\alpha \to \beta} \text{ is an unrelated DAG}\right\}.$$

We then state the main theorem of this thesis:

Theorem 3.11 (Adapted from Peters and Wainwright [Unpublished])
Let $\mathscr{C}$ be an ANM with graph $\mathcal{G}^0$ and distribution $\mathbb{P}$. Assume that $\mathcal{G}^0$ is unrelated and assume that $\mathscr{C}$ satisfies assumptions (B1) through (B7). Denote by $\mathcal{G}^{s=k}$ the graph of the Greedy entropy-search at step number k and let $\mathcal{G}^{s=0}$ be a graph with $V(\mathcal{G}^{s=0}) = V(\mathcal{G}^0)$ and $E(\mathcal{G}^{s=0}) = \emptyset$, and let $d := |E(\mathcal{G}^0)|$. Performing a Greedy-entropy search as described in Definition 3.8*, it holds that*

$$\mathcal{G}^{s=d} = \mathcal{G}^0.$$

That is, the true graph $\mathcal{G}^0$ is recovered in d steps.

Outline of proof

We start the proof by observing that it suffices to show for a fixed $k < d$ that

$$\mathcal{G}^{s=k+1} \subseteq \mathcal{G}^0$$

and that the Gaussian score cannot be increased further at $s = d$. This is proven in Corollary 3.16. In the following, let $(\alpha - \beta)$ be the undirected edge that maximizes $\Delta\ell^g$, and let $(\alpha \to \beta)$ be the direction of $(\alpha - \beta)$ that maximizes $\Delta\ell$. To increase readability, we write $\mathcal{G}^s$ in place of $\mathcal{G}^{s=k}$ The proof Theorem 3.11 is done in two parts. In part one, we show that the edge $(\alpha - \beta)$ must lie in $\text{ske}(\mathcal{G}^0)$. That is, either $(\alpha \to \beta) \in E(\mathcal{G}^0)$ or $(\beta \to \alpha) \in E(\mathcal{G}^0)$. We will assume for contradiction that $(\alpha - \beta) \notin \text{ske}(\mathcal{G}^0)$ and show that this contradicts the fact that $\Delta\ell^g$ was maximized in $(\alpha - \beta)$. To show this, we rely on three separate results:

(i) The marginal Gaussian score is strictly positive in correct[6] edges.

(ii) The marginal Gaussian score is identically zero in the edge $(\gamma \to \delta)$ if γ and δ can be d-separated in $\mathcal{G}^s$ by the parents of δ.

(iii) Whenever there are paths of the form $(\gamma \leftarrow \alpha \to \beta)$ in $\mathcal{G}^0$, the marginal Gaussian score is necessarily higher in $(\alpha \to \beta)$ than in $(\gamma \to \beta)$. Informally, this amounts to saying that it always better to add an edge from α to β than it is to add an edge from a noisy version of α into β.

The above three results are proved in Proposition 3.13, Proposition 3.15 and Lemma 3.17 respectively, and are sufficient to prove part one.

In part two we show that the chosen direction $(\alpha \to \beta)$ is correct. This is, again, shown by contradiction. The proof is carried out by considering a subgraph of $\mathcal{G}^s$, denoted $\tilde{\mathcal{G}}$, and showing that this is again the graph of an identifiable ANM, $\tilde{\mathscr{C}}$, only with a different noise distribution. In particular, the noise variable takes the form

$$\tilde{N}_\nu := N_\nu + \sum_{\gamma \in \mathbf{PA}_{\mathcal{G}^0}(\nu) \setminus \mathbf{PA}_{\mathcal{G}^s}(\nu)} f^0_{\nu,\gamma}(X_\gamma) \tag{3.12}$$

for some $\nu \in V(\mathcal{G}^0)$. We can then use the fact that the entropy score is maximized by $\tilde{\mathcal{G}}$ – as shown in Theorem 3.7 – to reach a contradiction. To show that $\tilde{\mathscr{C}}$ is identifiable, we need to show that the model $\tilde{\mathscr{C}}$ still satisfies assumptions (A1) through (A6). Assumptions (A1) and (A2) are restrictions of the function class, $\mathcal{F}$, and are thus trivially satisfied. Similarly, assumptions (A5) and (A6) are also trivially satisfied, as (A6) ensures that all variables of the type in equation (3.12) have finite second moment. It remains, then, to prove that the noise distributions in $\mathscr{C}$ satisfies assumptions (A3) and (A4). That is, for every $\nu \in V(\tilde{\mathcal{G}})$ we need to show that the following holds:

[6] Correct in the sense that they are in fact present in $\mathcal{G}^0$

(i) The density $\partial\mathbb{P}_{\tilde{N}_\nu}$ is strictly positive and three times continuously differentiable.

(ii) The differential equation $(\log \partial\mathbb{P}_{\tilde{N}_\nu})'' = 0$ has only discretely many solutions.

We prove in Lemma 3.20 that $\partial\mathbb{P}_{\tilde{N}_\nu}$ is strictly positive on $\mathbb{R}$ whenever $\mathcal{F}$ is such that transformations in $\mathcal{F}$ have density. We then put the remainder of point (i) on hold and move on to point (ii). We do this because our solution to point (ii) turns out to imply that $\partial\mathbb{P}_{\tilde{N}_\nu}$ is three times continuously differentiable – and smooth, even. The main difficulty lies in proving that variables of the type in (3.12) satisfy assumption (A3) and most of Section 3.6 is spent on this. By further restricting the original family of noise distributions to have real analytic[7] densities, we are able to prove through Lemma 3.24, Theorem 3.25 and Corollary 3.26 that the random variable $\tilde{N}_\nu$ has a density that is in C^∞ and satisfies point (ii) above.

Section 3.6 is dedicated to presenting and proving the results that we will make use of in the proof of Theorem 3.11.

3.6 PREREQUISITES FOR PROVING THEOREM 3.11

In the following section, let $\mathcal{G}^s$ and $\mathcal{G}^0$ be graphs as described in Theorem 3.11.

Part one

We begin this section by proving that the Gaussian score is unchanged when adding edges between conditionally independent nodes. First, however, this requires another result:

Proposition 3.12 (Doob's conditional independence statement)
Let $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ be sigma algebras. It then holds that $\mathcal{A}$ and $\mathcal{B}$ are independent given $\mathcal{C}$, if and only if

$$\forall A \in \mathcal{A}: \quad \mathbb{P}(A \mid \mathcal{B},\mathcal{C}) = \mathbb{P}(A \mid \mathcal{C}). \tag{3.13}$$

In particular, if X, Y and Z are random variables with finite expectations such that $X \perp\!\!\!\perp Y \mid Z$, it holds that

$$\mathbb{E}(X \mid Y, Z) = \mathbb{E}(X \mid Z). \tag{3.14}$$

Proof. A proof of (3.13) is given in Kallenberg [2002, proposition 6.6]. The statement in (3.14) is immediate from (3.13). □

Proposition 3.13
Let $\mathscr{C}$ be an ANM satisfying assumptions (A1) through (A6). Denote by $\mathcal{G}^0$ the graph of $\mathscr{C}$ and

[7] Readers not familiar with real analytic functions may simply think of these as smooth functions for now. A precise definition is given in Definition 3.23.

assume that $\mathcal{G}^0$ is unrelated. Let $\mathcal{G}$ be a subgraph of $\mathcal{G}^0$. Let α and β be distinct nodes in $V(\mathcal{G}^0)$ such that α and β are not directly connected in $\mathcal{G}^0$. If α and β are d-separated in $\mathcal{G}$ by α's parents, i.e. if

$$X_\alpha \perp\!\!\!\perp X_\beta \mid \boldsymbol{PA}_{\mathcal{G}}(X_\alpha) \tag{3.15}$$

then

$$\Delta\ell^g(\mathcal{G}, \beta \to \alpha) = 0.$$

Proof. Let

$$\Omega := \mathbf{PA}_{\mathcal{G}}(\alpha) \qquad \text{and} \qquad \Theta := \mathbf{PA}_{\mathcal{G}^0}(\alpha) \backslash \Omega.$$

That is, Ω is the set of $\mathcal{G}$-parents of α and Θ is the set of parents of α that are not in $\mathcal{G}$. If $\Theta = \emptyset$ or $\Omega = \emptyset$, we use the convention that summing over an empty set is identically zero. Let ϕ be the function satisfying $\phi \circ (X)_\Omega = \mathbb{E}[X \mid (X)_\Omega]$. We start by writing out the following:

$$\begin{aligned}
\mathbb{E}[X_\alpha \mid (X)_\Omega] &= \mathbb{E}\left[N_\alpha + \sum_{\omega \in \Omega} f^0_{\alpha,\omega}(X_\omega) + \sum_{\theta \in \Theta} f^0_{\alpha,\theta}(X_\theta) \;\middle|\; (X)_\Omega \right] \\
&= \mathbb{E}[N_\alpha \mid (X)_\Omega] + \mathbb{E}\left[\sum_{\omega \in \Omega} f^0_{\alpha,\omega}(X_\omega) \;\middle|\; (X)_\Omega \right] + \mathbb{E}\left[\sum_{\theta \in \Theta} f^0_{\alpha,\theta}(X_\theta) \;\middle|\; (X)_\Omega \right] \\
&\overset{(a)}{=} \mathbb{E}N_\alpha + \sum_{\omega \in \Omega} f^0_{\alpha,\omega}(X_\omega) + \mathbb{E}\left(\sum_{\theta \in \Theta} f^0_{\alpha,\theta}(X_\theta) \right) \\
&=: k + \sum_{\omega \in \Omega} f^0_{\alpha,\omega}(X_\omega).
\end{aligned} \tag{3.16}$$

To get equality (a) in the above, we use that

$$(N_\alpha, (X)_\Theta) \perp\!\!\!\perp (X)_\Omega,$$

as $\mathcal{G}^0$ is unrelated by assumption. The rewrite in (3.16) is to emphasize that the unconditional expectations are constants. By (3.16) we see that

$$\phi((x)_\Omega) = \sum_{\omega \in \Omega}^{d} f^0_{\alpha,\omega}(x_\omega) + k$$

almost surely. This means that $\phi \in \mathcal{F} \cap \mathcal{M}$ and so

$$\begin{aligned}
\min_{(f)_\Omega \in \mathcal{F} \cup \mathcal{C}} \mathbb{E}\left(X_\alpha - \sum_{\omega \in \Omega} f_\omega(X_\omega) \right)^2 &= \min_{(f)_\Omega \in \mathcal{M}} \mathbb{E}\left(X_\alpha - \sum_{\omega \in \Omega} f_\omega(X_\omega) \right)^2 \\
&= \mathbb{E}\left(X_\alpha - \mathbb{E}[X_\alpha \mid (X)_\Omega] \right)^2,
\end{aligned} \tag{3.17}$$

where the last equality follows from Proposition 3.5 and the fact that linear combinations of measurable functions are measurable. We then use equation (3.20) to get

$$\mathbb{V}\left(X_\alpha - \sum_{\omega \in \Omega} \hat{f}_\omega(X_\omega) - \hat{f}_\beta(X_\beta) \right) = \min_{(f)_\Omega, f_\beta \in \mathcal{F} \cup \mathcal{C}} \mathbb{V}\left(X_\alpha - \sum_{\omega \in \Omega} f_\omega(X_\omega) - f_\beta(X_\beta) \right)$$

$$
\begin{aligned}
&\overset{(b)}{\geq} \min_{(f)_\Omega, f_\beta \in \mathcal{M}} \mathbb{V}\left(X_\alpha - \sum_{\omega\in\Omega} f_\omega(X_\omega) - f_\beta(X_\beta)\right) \\
&= \mathbb{V}\left(X_\alpha - \mathbb{E}[X_\alpha \mid (X)_\Omega, X_\beta]\right) \\
&\overset{(c)}{=} \mathbb{V}\left(X_\alpha - \mathbb{E}[X_\alpha \mid (X)_\Omega]\right) \\
&= \min_{(f)_\Omega \in \mathcal{F}\cup\mathcal{C}} \mathbb{V}\left(X_\alpha - \sum_{\omega\in\Omega} f_\omega(X_\omega)\right) \\
&= \mathbb{V}\left(X_\alpha - \sum_{\omega\in\Omega} \hat{f}_\omega(X_\omega)\right).
\end{aligned}
$$

Above, inequality (b) follows from the fact that $\mathcal{F}\cup\mathcal{C} \subseteq \mathcal{M}$; by minimizing over a larger class of functions, we can weakly improve the minimization. To get equality (c) we use Proposition 3.12 and the assumption (3.15). In summary, we have shown that

$$\Delta\ell^{\mathrm{g}}\left(\mathcal{G}, \beta \to \alpha\right) = \log \mathbb{V}\left(X_\alpha - \sum_{\omega\in\Omega} \hat{f}_\omega(X_\omega)\right) - \mathbb{V}\left(X_\alpha - \sum_{\omega\in\Omega} \hat{f}_\omega(X_\omega) - \hat{f}_\beta(X_\beta)\right) \leq 0.$$

By construction, we allow for constant functions in the $\mathcal{L}^2$ minimization and thus the inequality above tightens to an equality by choosing $\hat{f}_\beta = 0$ identically. This concludes the proof. □

Remark 3.14. Observe that the calculations made in the proof of Proposition 3.13 leading to equation (3.16) and the subsequent conclusion in (3.17) hold true whenever the set Ω is a subset[8] of the true parents and independent of the parents that are not included. Indeed, we never used the structure of Ω, apart from the fact that all nodes in Ω were parents in the true graphs as well. When we assume the graph of the ANM to be unrelated, the independence of $(X)_\Omega$ and $(X)_\Theta$ is implied.

Next, we will in Proposition 3.15 prove that $\Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \alpha \to \beta\right) > 0$ whenever the edge $(\alpha \to \beta)$ is in $\mathcal{G}^0$, but not in $\mathcal{G}^s$.

Proposition 3.15

Let $\mathscr{C}$ be an ANM satisfying assumptions (A1) through (A6). Let $\mathcal{G}^0$ be the graph of $\mathscr{C}$ and assume that it is unrelated. Let $\mathcal{G}$ be a true subgraph of $\mathcal{G}^0$. If the edge $(\alpha \to \beta)$ is in $\mathcal{G}^0$ but not in $\mathcal{G}$ it follows that

$$\Delta\ell^{g}\left(\mathcal{G}, X_\alpha \to X_\beta\right) > 0.$$

Proof. Let $\Omega := \mathbf{PA}_{\mathcal{G}}(\beta)$ and let $\Theta := \mathbf{PA}_{\mathcal{G}^0}(\beta)\backslash\Omega$. To prove the statement, we prove the equivalent statement that

$$\mathbb{V}\left(X_\beta - \sum_{\omega\in\Omega} \hat{f}_\omega(X_\omega) - \hat{f}_\alpha(X_\alpha)\right) < \mathbb{V}\left(X_\beta - \sum_{\omega\in\Omega} \hat{f}_\omega(X_\omega)\right). \tag{3.18}$$

[8] Possibly with equality

By Remark 3.14 we conclude that

$$\mathbb{V}\left(X_\beta - \sum_{\omega\in\Omega} \hat{f}_\omega(X_\omega)\right) = \mathbb{V}\left(X_\beta - \mathbb{E}[X \mid (X)_\Omega]\right)$$

and

$$\mathbb{V}\left(X_\beta - \sum_{\omega\in\Omega} \hat{f}_\omega(X_\omega) - \hat{f}_\alpha(X_\alpha)\right) = \mathbb{V}\left(X_\beta - \mathbb{E}[X \mid (X)_\Omega, X_\alpha]\right),$$

as $\Omega \cup \{\alpha\} \subseteq \mathbf{PA}_{\mathcal{G}^0}(\beta)$ by assumption. As $\mathcal{G}^0$ is unrelated we have

$$(X)_\Omega \perp\!\!\!\perp (X)_\Theta$$

which implies that

$$\begin{aligned}
\mathbb{V}(X_\beta - \mathbb{E}[X \mid (X)_\Omega]) =& \mathbb{V}\left(N_\beta + \sum_{\theta\in\Theta} f^0_{\beta,\theta}(X_\theta) + \sum_{\omega\in\Omega} f^0_{\beta,\omega}(X_\omega)\right.\\
&\left. - \mathbb{E}\left[N_\beta + \sum_{\theta\in\Theta} f^0_{\beta,\theta}(X_\theta) + \sum_{\omega\in\Omega} f^0_{\beta,\omega}(X_\omega) \,\middle|\, (X)_\Omega\right]\right)\\
=& \mathbb{V}\left(N_\beta - \mathbb{E}N_\beta + \sum_{\theta\in\Theta} f^0_{\beta,\theta}(X_\theta) - \mathbb{E}\left(\sum_{\theta\in\Theta} f^0_{\beta,\theta}(X_\theta)\right)\right)\\
=& \mathbb{V}N_\beta + \sum_{\theta\in\Theta} \mathbb{V}f^0_{\beta,\theta}(X_\theta)\\
=& \mathbb{V}N_\beta + \sum_{\theta\in\Theta\setminus\{\alpha\}} \mathbb{V}f^0_{\beta,\theta}(X_\theta) + \mathbb{V}f^0_{\beta,\alpha}(X_\alpha). \qquad (3.19)
\end{aligned}$$

Similarly, we find that

$$\begin{aligned}
&\mathbb{V}(X_\beta - \mathbb{E}[X \mid (X)_\Omega, X_\alpha])\\
&\quad= \mathbb{V}\Bigg(N_\beta + \sum_{\theta\in\Theta\setminus\{\alpha\}} f^0_{\beta,\theta}(X_\theta) + \sum_{\omega\in\Omega} f^0_{\beta,\omega}(X_\omega) + f^0_{\beta,\alpha}(X_\alpha)\\
&\quad - \mathbb{E}\left[N_\beta + \sum_{\theta\in\Theta\setminus\{\alpha\}} f^0_{\beta,\theta}(X_\theta) + \sum_{\omega\in\Omega} f^0_{\beta,\omega}(X_\omega) + f^0_{\beta,\alpha}(X_\alpha) \,\middle|\, (X)_\Omega, X_\alpha\right]\Bigg)\\
&\quad= \mathbb{V}\left(N_\beta - \mathbb{E}N_\beta + \sum_{\theta\in\Theta\setminus\{\alpha\}} f^0_{\beta,\theta}(X_\theta) - \mathbb{E}\left(\sum_{\theta\in\Theta\setminus\{\alpha\}} f^0_{\beta,\theta}(X_\theta)\right)\right)\\
&\quad= \mathbb{V}N_\beta + \sum_{\theta\in\Theta\setminus\{\alpha\}} \mathbb{V}f^0_{\beta,\theta}(X_\theta). \qquad (3.20)
\end{aligned}$$

Comparing equations (3.19) and (3.20) we see that

$$\mathbb{V}(X_\beta - \mathbb{E}[X \mid (X)_\Omega, X_\alpha]) < \mathbb{V}(X_\beta - \mathbb{E}[X \mid (X)_\Omega]).$$

Using Remark 3.14 again, this implies

$$\mathbb{V}\left(X_\beta - \sum_{\omega\in\Omega} \hat{f}_\omega(X_\omega) - \hat{f}_\alpha(X_\alpha)\right) < \mathbb{V}\left(X_\beta - \sum_{\omega\in\Omega} \hat{f}_\omega(X_\omega)\right)$$

which in turn yields

$$\Delta\ell^{\mathrm{g}}\left(\mathcal{G}, X_\alpha \to X_\beta\right) > 0,$$

which concludes the proof. □

Proposition 3.13 and Proposition 3.15 in combination imply two things about the Greedy entropy-search: 1) As long as we can still find edges that increase the Gaussian score, we have not yet recovered all of $\mathcal{G}^0$, and 2) when we have recovered every edge in $\mathcal{G}^0$, *every* other candidate edge will have a marginal Gaussian score of exactly zero. For referential purposes, we state the second point as a corollary:

Corollary 3.16
Let $\mathscr{C}$ be an ANM with graph $\mathcal{G}^0$. For every two distinct nodes α and β in $\mathcal{G}^0$, that are not directly connected in $\mathcal{G}^0$ exactly one of the following statements will be true:

(i) $\Delta\ell^{g}\left(\mathcal{G}^0, \alpha \to \beta\right) = 0.$

(ii) Adding $(\alpha \to \beta)$ to $\mathcal{G}^0$ induces a cycle.

Proof. Fix α and β such that they are not directly connected in $\mathcal{G}^0$. If β is a descendant of α, then α is a non-descendant of β. By the Directed Local Markov property,[9] it follows that $X_\beta \perp\!\!\!\perp X_\alpha \mid \mathbf{PA}_{\mathcal{G}^0}(\beta)$, which implies that

$$\Delta\ell^{\mathrm{g}}\left(\mathcal{G}^0, \alpha \to \beta\right) = 0$$

by Proposition 3.13. If β is not a descendant of α there are four possibilities: Either 1) α is a descendant of β, 2) every path between α and β traverses a collider, 3) there are no paths between α and β or 4) every path between α and β traverses a fork. In the first case, adding $(\alpha \to \beta)$ induces a directed cycle. In the second and third case, we have $X_\beta \perp\!\!\!\perp X_\alpha$ unconditionally and thus $\Delta\ell^{\mathrm{g}}\left(\mathcal{G}^0, \alpha \to \beta\right) = 0$ by Proposition 3.13. In the fourth case, we use the directed Markov property to conclude that the marginal score is zero. □

We conclude this subsection with one more lemma, Lemma 3.17, and an example. Lemma 3.17 is a technical one, and its relation to building graphs from a Greedy search is not immediate.

[9] Recall equation (DLM) in Definition 1.3.

Example 3.19 is intended to illustrate a situation in which Lemma 3.17 can be used to correctly identify edges in a Greedy search. The reader can potentially benefit from skipping the proof of Lemma 3.17 and returning after reading Example 3.19.

Lemma 3.17
Let $(X,Y) \sim \mathbb{P}_{(X,Y)}$ and $N \sim \mathbb{P}_N$ be random variables such that $N \perp\!\!\!\perp (X,Y)$. Let g be a function belonging to $\mathcal{F}$. For all $f \in \mathcal{F}$ assume that the mapping $x \mapsto \mathbb{E}_{\mathbb{P}_N} f(g(x)+N)$ is again in $\mathcal{F}$. It then holds that

$$\min_{f\in\mathcal{F}} \mathbb{E}_{\mathbb{P}_{(X,Y)}} \big(Y - f(X)\big)^2 < \min_{f\in\mathcal{F}} \mathbb{E}_{\mathbb{P}_{(X,Y)}\otimes\mathbb{P}_N} \big(Y - f(g(X)+N)\big)^2 .$$

Proof. Take any $f \in \mathcal{F}$. Since the mapping $x \mapsto (y-x)^2$ is strictly convex on $\mathbb{R}$ for any choice of y it follows from Jensen's inequality (see e.g Hansen 2009, theorem 16.31) that

$$(Y - \mathbb{E}_{\mathbb{P}_N} f(g(X)+N))^2 < \mathbb{E}_{\mathbb{P}_N}(Y - f(g(X)+N))^2 \tag{3.21}$$

Taking expectations under $\mathbb{P}_{(X,Y)}$ on both sides and using that the mapping $x \mapsto (y-x)^2$ is a positive, (Borel) measurable[10] function and using monotonicity of integrals we find that

$$\mathbb{E}_{\mathbb{P}_{(X,Y)}} \big(Y - \mathbb{E}_{\mathbb{P}_N} f(g(X)+N)\big)^2 < \mathbb{E}_{\mathbb{P}_{(X,Y)}\otimes\mathbb{P}_N} \big(Y - f(g(X)+N)\big)^2 \tag{3.22}$$

We have used Tonelli's theorem (see e.g. Hansen 2009, theorem 9.10) to collect the right-hand side expectation under a product measure – this holds as probability spaces are σ-finite trivially. As (3.22) holds for any $f \in \mathcal{F}$, it follows that

$$\min_{f\in\mathcal{F}} \mathbb{E}_{\mathbb{P}_{(X,Y)}} \big(Y - f(X)\big)^2 < \min_{f\in\mathcal{F}} \mathbb{E}_{\mathbb{P}_{(X,Y)}\otimes\mathbb{P}_N} \big(Y - f(g(X)+N)\big)^2 .$$

The change from $\mathbb{E}_{\mathbb{P}_N} f(g(X)+N)$ to $f(X)$ when taking the minimum on the left-hand side above, holds as both are elements in $\mathcal{F}$ by assumption. This concludes the proof. □

Remark 3.18. Be sure to notice how delicate the application of Jensen's inequality in equation (3.21) is; if Y is measurable with respect to $\sigma(g(X)+N)$, this does not hold in general. Indeed, a simple example is when $Y := g(X)+N$. Then $(Y - \mathbb{E}_N \mathbb{E}[Y \mid g(X)+N])^2 = (N - \mathbb{E}_N N)^2$, but $\mathbb{E}_N(Y - \mathbb{E}[Y \mid g(X)+N])^2 = \mathbb{E}_N 0 = 0$. Observe also that Lemma 3.17 implies

$$\min_{f\in\mathcal{F}} \mathbb{E}_{\mathbb{P}_{(X,Y)}} \big(Y - f_1(Z) - f_2(X)\big)^2 < \min_{f\in\mathcal{F}} \mathbb{E}_{\mathbb{P}_{(X,Y)}\otimes\mathbb{P}_N} \big(Y - f_1(Z) - f_2(g(X)+N)\big)^2 ,$$

for some random variable Z, such that $N \perp\!\!\!\perp Z$ as well. The proof is exactly the same, one simply swaps Y for $Y - f_1(Z)$ all the way throughout.

In essence, Lemma 3.17 means that the $\mathcal{L}^2$ loss of regressing Y onto X is always lower than regressing Y onto a noisier version of X. Its relation to graph scoring can be illustrated by the following example:

[10] As both X and Y are random variables, they are measurable mappings simply by construction.

Example 3.19. Consider an ANM $\mathscr{C}$, with graph $\mathcal{G}$ as illustrated in Figure 3.2 (left). Suppose we know the distribution implied by $\mathscr{C}$ – but not the graph – and we are doing a Greedy search to recover $\mathcal{G}$. In the first step of the Greedy search we have correctly added the edge $(\alpha \to \gamma)$ to obtain the graph $\mathcal{G}^{s=1}$, which has no other edges. Suppose now that we are doing a second step of the Greedy search, and that we have narrowed down the candidate edges to two choices: 1) Either we should add $(\alpha \to \beta)$ to $\mathcal{G}^{s=1}$, or 2) we should add $(\gamma \to \beta)$. By construction, the noise term in X_α, i.e. N_α, is independent of X_β. Lemma 3.17 now tells us that it must then be better to add the edge $(\gamma \to \beta)$. This is essentially because α can be thought of as a noisier version of γ, and thus there is more information about β contained in γ. More formally, we know that

$$X_\alpha = f_\gamma(X_\gamma) + N_\alpha,$$

and thus

$$\min_{f\in\mathcal{F}} \mathbb{E}(X_\beta - f(X_\gamma))^2 < \min_{f\in\mathcal{F}} \mathbb{E}(X_\beta - f(f_\gamma(X_\gamma) + N_\gamma))^2 = \min_{f\in\mathcal{F}} \mathbb{E}(X_\beta - f(X_\alpha))^2. \tag{3.23}$$

Because we allow for constant functions in the determination the $\hat{f}$-functions, the residuals will always be unbiased, which makes the statement in (3.23) equivalent to

$$\mathbb{V}\left((X_\beta - \hat{f}(X_\gamma)\right) < \mathbb{V}\left((X_\beta - \hat{f}(X_\alpha)\right).$$

This in turn implies that $\Delta\ell^{\mathrm{g}}\left(\mathcal{G}^1, \gamma \to \beta\right) > \Delta\ell^{\mathrm{g}}\left(\mathcal{G}^1, \alpha \to \beta\right)$.

Part two

We now turn our attention to part two of the proof of Theorem 3.11. Recall that this involves constructing a sub-model, $\tilde{\mathscr{C}}$, which we require to also be an identifiable ANM. This in turn requires $\tilde{\mathscr{C}}$ to still satisfy assumptions (A3) and (A4). We begin by introducing some notation, after which we move on to prove Lemma 3.20 which provides sufficient conditions under which the convolution of two random variables has strictly positive density on $\mathbb{R}$.

Notation. Let $n \in \mathbb{N} \cup \{\infty\}$. A sequence of intervals, $\{\mathcal{X}_i\}_{i=1}^n$, that are pairwise disjoint and such that

$$\bigcup_{i=1}^n \mathcal{X}_i = \mathbb{R},$$

is called a partition of $\mathbb{R}$.
The restriction of a function, f, to $\mathcal{X}_i$ is the mapping $f|_{\mathcal{X}_i} \colon \mathcal{X}_i \to f(\mathcal{X}_i)$ given by

$$\forall x \in \mathcal{X}_i : \qquad f|_{\mathcal{X}_i}(x) = f(x).$$

Observe that when $n = 1$, we simply have $\mathcal{X}_1 = \mathbb{R}$ and $f|_{\mathcal{X}_1} = f$.
The convolution of two integrable functions, f and g, is denoted by $f * g$.

Lemma 3.20

Let N and X be random variables with strictly positive density on $\mathbb{R}$, such that $N \perp\!\!\!\perp X$. Let f be a function in $\mathcal{F}$ and define $Y := f(X) + N$. Assume that there exists a countable partition of $\mathbb{R}$, denoted $\{\mathcal{X}_i\}_{i=1}^n$, such that every $f|_{\mathcal{X}_i}$ is a bijection. Then Y has strictly positive density on $\mathbb{R}$.

Proof. By assumption, f is continuously differentiable, constant nowhere and invertible on every $\mathcal{X}_i$. It then follows from the Inverse Function Theorem (see e.g. Rudin 1976, Theorem 9.24) that

$$\forall i \leq n: \qquad f|_{\mathcal{X}_i}^{-1} \in C^1(f(\mathcal{X}_i)).$$

That is, the inversion of every restriction is itself continuously differentiable on an interval. It then follows that $f(X)$ has density with respect to the Lebesgue measure (see Hansen 2009, Theorem 12.9). Now, let

$$\mathcal{M} := \{x \in \mathbb{R} : \ \partial\mathbb{P}_{f(X)}(x) = 0\}$$

be the set on which $f(X)$ has density equal to zero[11]. It must hold that $m(\mathbb{R}\backslash\mathcal{M}) > 0$, as $\partial\mathbb{P}_{f(X)}$ is a density. It is then trivial that

$$\begin{aligned}\partial\mathbb{P}_Y(y) = (\partial\mathbb{P}_N * \partial\mathbb{P}_{f(X)})(y) &= \int_{\mathbb{R}} \partial\mathbb{P}_N(y-t) \cdot \partial\mathbb{P}_{f(X)}(t)\ dm(t) \\ &= \int_{\mathbb{R}\backslash\mathcal{M}} \partial\mathbb{P}_N(y-t) \cdot \partial\mathbb{P}_{f(X)}(t)\ dm(t) \\ &> \int_{\mathbb{R}} 0\ dm(t) = 0,\end{aligned}$$

by the convolution formula (see e.g. Hansen 2009, corollary 20.9) and monotonicity of integrals. □

As seen by Lemma 3.20, for a convolution of the type $f(X) + N$ to have strictly positive density, it is sufficient that X and N have strictly positive density, and that $f \in \mathcal{F}$ can be inverted on a partition. This a mild assumption, as we have already assumed every $f \in \mathcal{F}$ to be in C^3 and be nowhere constant. In particular, by restricting $\mathcal{F}$ to the class of functions that are invertible on a partition, Lemma 3.20 implies that every convolution in an ANM has strictly positive density under assumptions (A1) and (A4); by assumption, source nodes have strictly positive densities, and nodes of depth one are exactly of the type in Lemma 3.20 and will therefore also have strictly positive density. This, in turn, implies that nodes of depth two are also convolutions of the type considered in Lemma 3.20 and so forth. When f can be partitioned as in Lemma 3.20 we say that f is a density preserving transformation.

The remainder of this section is concerned with assumption (A3). That is, we study the solutions of the differential equation $(\log f)'' = 0$. More generally, we aim to find conditions on density functions f and g such that, whenever $(\log f)'' = 0$ only has discretely many solutions, it should hold that $(\log f * g)'' = 0$ has at most discretely many solutions. In particular, this means that we

[11] This is possibly the empty set.

are looking for conditions which ensure that there does not exist an interval on which $(\log f * g)'' = 0$. Throughout this section, let

$$\mathcal{E}(K) := \big\{ f : \mathbb{R} \to \mathbb{R} \mid \forall x \in K : \ (\log f(x))'' = 0 \big\}$$

be the family of all functions that solve the differential equation $(\log f)'' = 0$ on the set K. Integrating twice and taking exponentials, it is easily seen that

$$\mathcal{E}(K) := \Big\{ (c_1, c_2) \in \mathbb{R}^2 \mid \forall x \in K : \ f(x) = \exp\left(c_1 \cdot x + c_2\right) \Big\}.$$

That is, $\mathcal{E}(K)$ is the family of all exponential functions, which have a linear exponential. Observe that the family of all functions that are constant on K is contained in $\mathcal{E}(K)$, as c_1 is potentially zero. A function belonging to $\mathcal{E}((a,b))$, for some interval (a,b), is called a **log-linear** function. Similarly, a function that is not log-linear is called a **non-log-linear** function. That is, a non-log-linear function is one that satisfies assumption (A4). Intuitively, we might expect that the convolution of f and g is log-linear if and only if both f and g are. This is not an unreasonable thought; clearly $\mathcal{E}$ is closed under products, and thus $f * g$ becomes an integral of a function that is log-linear somewhere. As exponential functions are eigenfunctions of the integral operator, it is natural to think that $\mathcal{E}$ is closed under convolutions. That is, however, not always the case. To motivate the problem of convolutions, we give two examples below. Example 3.21 illustrates a case in which two non-log-linear density functions convolve into a log-linear function. In contrast to this, Example 3.22 illustrates the opposite case, in which two log-linear functions convolve into a non-log-linear function.

Example 3.21. let $N \sim \mathcal{N}(0,1)$ and $M \sim f \cdot m$ be random variables, where

$$f(x) \propto \exp(-x^2) \cdot \mathbb{1}_{(-\infty,0)\cup[\xi,\infty)}(x) + (1-x) \cdot \mathbb{1}_{[0,1/2)}(x) + x \cdot \mathbb{1}_{[1/2,1)} + (2-x) \cdot \mathbb{1}_{[1,\xi)}, \tag{3.24}$$

and ξ is the solution of the equation $2 - x = \exp(-x^2)$. That is, M is a random variable that behaves as a normal distribution on $(-\infty, 0) \cup [\xi, \infty)$, but not in between those intervals. It is easy to see that $f(x)$ is indeed a strictly positive density function, when (3.24) is properly normalized. The normalization constant is, however, not of interest, so we simply disregard it. Next we note that $f(x)$ is non-log-linear everywhere. Define $\Psi : \mathbb{R} \to (0,1)$ by $\Psi(x) = \mathbb{P}(N \leq x)$ and construct $X := \Psi(N) + M$. It is trivial that $\Psi(N) \sim U(0,1)$. Denote by g the density of $\Psi(N)$. By the convolution formula, X has density f_X given by

$$f_X(t) = (f * g)(t) = \int_{\mathbb{R}} f(x) \cdot g(t-x)\ dm(x) = \int_{(t-1,t)} f(x)\ dm(x).$$

At $t = 1$, $f * g$ is exactly the area of the box $(0,1)^2$ minus the area of the triangle

$$A := \{(x,y) \in \mathbb{R}^2 : \ x \in (0,1),\ y \in (f(x), 1)\}$$

which is seen to be

$$(f * g)(1) = 1 - \frac{1}{4} = \frac{3}{4}.$$

For any $\epsilon \in (0, 1/2)$, however, it is clear that $(f * g)(1 + \epsilon)$ is the area of the box $(\epsilon, 1 + \epsilon)^2$ minus the area of

$$A' := \{(x, y) \in \mathbb{R}^2 : \ x \in (\epsilon, 1 + \epsilon), \ y \in (f(x), 1)\}.$$

The latter set no longer spans a triangle in $\mathbb{R}^2$. However, one can note that on the interval $[0, 1/2]$, $f(x)$ is 1-periodic and thus $m_2(A) = m_2(A')$ by translation invariance of the Lebesgue measure. Similarly $m_2((0, 1)^2) = m_2((\epsilon, 1 + \epsilon)^2)$. See Figure 3.4 for an illustration. In conclusion, we have found that

$$\forall t \in [1, 3/2] : \quad (f * g)(t) = \frac{3}{4},$$

and thus $f * g \in \mathcal{E}([1, 3/2])$. That is, the convolved variable X will not fulfill assumption (A3), even though both N and M do.

Example 3.21 shows us, that while it may be difficult to find examples of convolutions that do not satisfy assumption (A3), it is not impossible. The trick in the above example was to construct f in a way such it was periodic on an interval. In fact, the curvature of f on this interval does not matter[12]; if there is $[a, b] \subset \mathbb{R}$ such that f is T-periodic on $[a, b]$, the convolution of f with the density of a $U(0, T)$-distributed random variable will be constant on $[a + T, b + T]$ by applying the same argument as in Example 3.21.

The next example goes in the opposite direction of Example 3.21 to show that the convolution of two log-linear functions need not itself be log-linear:

Example 3.22 ($\mathcal{E}$ is not closed under $*$)**.** Let N be a Laplace$(0, 1)$-distributed random variable. The variable then N has density

$$f(x) \propto \exp(-|x|).$$

Let $U \sim U(-1/2, 1/2)$ be a uniformly distributed random variable on $(-1/2, 1/2)$, such that $U \perp\!\!\!\perp N$. Let $\Psi : (-1/2, 1/2) \to \mathbb{R}$ be given by

$$\Psi(x) = -\operatorname{sign}(U) \cdot \log(1 - 2 \cdot |U|).$$

By standard transformation results, we find that the variable $M := \Psi(U)$ is also Laplace$(0, 1)$-distributed and thus has density f. It is easy to see that $f \in \mathcal{E}(\mathbb{R} \backslash \{0\})$. However, given $t > 0$, we see that

$$\forall x \in (0, t) : \quad f(x) \cdot f(t - x) \propto \exp(-x) \cdot \exp(-(t - x)) = \exp(-t).$$

This then means that

$$(f * f)(t) \propto 2 \int_{\mathbb{R}_+} \exp(-2x - t) \ dm(x) + m_2((0, t) \times (0, \exp(-t))) = \exp(-t) \cdot (1 + t), \tag{3.25}$$

for $t > 0$. Similarly, for $t \leq 0$ we have

$$(f * f)(t) \propto 2 \int_{\mathbb{R}_+} \exp(-2x + t) \ dm(x) + m_2((t, 0) \times (0, \exp(-t))) = \exp(t) \cdot (1 - t). \tag{3.26}$$

[12] provided that f is continuous.

The terms in equations (3.25) and (3.26) are both non-log-linear, and we conclude that $f * f$ is non-log-linear. That is, even though both N and U were log-linear functions, the convolution $Y := \Psi(U) + N$ is non-log-linear.

For the sake of good order, it should be noted that we cheated a little bit in Example 3.22; in the setting of ANMs, we require by assumption (A4), that the noise distributions are supported on $\mathbb{R}$ – which U is not in Example 3.22. However, we can reach the same conclusion by considering four i.i.d. standard normal random variables, $U_1, \ldots, U_4$, and defining $\Psi : \mathbb{R}^4 \to \mathbb{R}$ by $\Psi(u_1, \ldots, u_4) = u_1 \cdot u_2 - u_3 \cdot u_4$. Then $M := \Psi(U_1, \ldots, U_4) \sim \text{Laplace}(0, 1)$ (see e.g. Ding and Blitzstein 2018) and the argument becomes identical to that in Example 3.22.

The purpose of Example 3.21 and Example 3.22 is to show that it can be surprisingly difficult to make restrictions on the noise variables in an ANM to ensure that their convolution is non-log-linear. The reason that Example 3.21 wound up producing a log-linear density, was that the density f was, in a sense, not sufficiently smooth. Indeed, the piecewise definition of f was what induced the periodicity of f. It is appealing to try and exclude such examples by simply requiring that the densities of the noise variables are not defined in a piece-wise manner. However, *any* function can trivially be defined in a piecewise manner. Instead, it seems that the requirement should be that the densities of the noise variables are sufficiently smooth. It is, however, not clear just when a function is 'sufficiently smooth', and should be formalized. In Lemma 3.24 and Theorem 3.25 below, we provide a solution to this problem by using real analytic functions. We begin with a definition:

Definition 3.23 (Krantz and Parks 2002)**.** A function, $f : \mathbb{R} \to \mathbb{R}$, is said to be real analytic on an interval, I, if and only if $f \in C^\infty(I)$ and for every $y \in I$ there exists an open neighborhood of y, $I' \subseteq I$, such that f can be represented as a convergent power series on I', i.e.

$$\forall x \in I' : \quad f(x) = \sum_{k=0}^{\infty} a_k (x-y)^k,$$

with $a_k \in \mathbb{R}$ for all $k \in \mathbb{N}_0$.

If f is real analytic on I, we write $f \in C^\omega(I)$ and if f is strictly positive, we write $f \in C^\omega_+(I)$. Whenever $I = \mathbb{R}$, we simply write C^ω and C^ω_+.

Lemma 3.24

Let $f \in C^\omega$. Assume that there exists an interval $[a, b]$ such that $f \in \mathcal{E}([a, b])$. Then $f \in \mathcal{E}(\mathbb{R})$.

Proof. Assume for contradiction that $f \notin \mathcal{E}(\mathbb{R} \backslash [a, b])$. As f is log-linear and smooth in a it follows that

$$\forall n \in \mathbb{N}_0: \qquad f^{(n)}(a) = k \cdot c^n \cdot \exp(c \cdot a) =: \tilde{k} \cdot c^n,$$

for real constants $\tilde{k}$ and c. Because f is real analytic everywhere, f can be represented as a Taylor series on an open interval, J, around a (see Krantz and Parks 2002, Corollary 1.1.16):

$$\forall x \in J: \qquad f(x) = \sum_{i=0}^{\infty} \frac{f^{(i)}}{i!}(x-a)^i = \tilde{k} \sum_{i=0}^{\infty} \frac{c^i}{i!}(x-a)^i = \tilde{k} \exp(c \cdot (x-a)).$$

But then $f \in \mathcal{E}(J)$. Since a is centered in J, this is a contradiction as it implies that $f \in \mathcal{E}((a - m(J)/2, a))$. We then conclude that $f \in \mathcal{E}(\mathbb{R} \backslash [a,b])$ and thus also $f \in \mathcal{E}(\mathbb{R})$. □

Lemma 3.24 shows that a real analytic function is either log-linear everywhere – i.e. in $\mathcal{E}(\mathbb{R})$ – or not log-linear anywhere. This may not be a surprising fact, as we argued earlier that the situation in Example 3.21 was caused by a lack of smoothness.

Theorem 3.25
Let $f \in C^\omega \cap \mathcal{L}^\infty$ and $g \in \mathcal{L}^1$ be functions. Then $f * g \in C^\omega$ and

$$\forall n \in \mathbb{N}_0: \qquad (f*g)^{(n)} = f^{(n)} * g.$$

Proof. Fix $n \in \mathbb{N}_0$ and define

$$f_n(x,t,\epsilon) = \frac{f^{(n)}(x+\epsilon-t) - f^{(n)}(x-t)}{\epsilon} g(t).$$

By the Mean Value Theorem (see e.g. Rudin 1976, Theorem 5.10) and Krantz and Parks [2002, Corollary 1.2.9], there exists real constants, $C > 0$ and $R > 0$, such that

$$\forall (t,x,\epsilon) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+ : \quad |f_n(x,t,\epsilon)| \leq C \cdot \frac{(n+1)!}{R^{n+1}} \cdot g(t).$$

That is, f_n is dominated by an integrable function. By definition, $f_n(x,t,\epsilon) \to f^{(n+1)}(x-t) \cdot g(t)$ as $\epsilon \to 0$. By the Dominated Convergence Theorem (see e.g. Schilling 2005, Theorem 11.2) we find that

$$\begin{aligned}
\frac{d}{dx}(f^{(n)} * g)(x) &= \lim_{\epsilon \to 0} \frac{(f^{(n)} * g)(x+\epsilon) - (f^{(n)} * g)(x)}{\epsilon} \\
&= \lim_{\epsilon \to 0} \int_{\mathbb{R}} \frac{1}{\epsilon} \left[f^{(n)}(x+\epsilon-t) - f^{(n)}(x-t) \right] \cdot g(t) \; dm(t) \\
&= \lim_{\epsilon \to 0} \int_{\mathbb{R}} f_n(x,t,\epsilon) \; dm(t) \\
&= \int_{\mathbb{R}} \lim_{\epsilon \to 0} f_n(x,t,\epsilon) \; dm(t) \\
&= \int_{\mathbb{R}} f^{(n+1)}(x-t) \cdot g(t) \; dm(t) \\
&= (f^{(n+1)} * g)(x),
\end{aligned}$$

for arbitrary $n \in \mathbb{N}_0$ and $x \in \mathbb{R}$. It then follows that $f^{(n+1)}$ is bounded and continuous [Schilling, 2005, Theorem 14.8]. As n was arbitrary, we conclude that $f * g \in C^\infty$. Furthermore, since $f \in C^\omega_+$ and $g \in \mathcal{L}^1$, it must hold that there exists constants such that

$$\forall n \in \mathbb{N}_0: \qquad f^{(n)} * g \leq C_n \frac{n!}{R_n^n} \int_{\mathbb{R}} g \; dm < \infty,$$

i.e. every derivative of the convolution is bounded. It then follows that $f * g$ is real analytic [Krantz and Parks, 2002, Lemma 1.2.10]. □

Corollary 3.26

Let $f \in C^{\omega}_{+}$ and $g \in \mathcal{L}^{\infty}$ be density functions. For any $a, b \in \bar{\mathbb{R}}$ with $a < b$ it holds that $f * g \notin \mathcal{E}([a,b])$.

Proof. Since f and g are both non-negative and measurable functions, it follows from Tonelli's Theorem (see e.g. Hansen 2009, Theorem 9.4) that

$$\begin{aligned}\int_{\mathbb{R}} (f * g)(t)\ dm(t) &= \int_{\mathbb{R}} \int_{\mathbb{R}} f(t-x) \cdot g(x)\ dm(x)\ dm(t) \\ &= \int_{\mathbb{R}} g(x) \int_{\mathbb{R}} f(t-x)\ dm(t)\ dm(x) \\ &= \int_{\mathbb{R}} g(x)\ dm(x) = 1 < \infty,\end{aligned}$$

i.e. $f * g \in \mathcal{L}^1(\mathbb{R})$. Using Theorem 3.25 we conclude that $f * g \in C^{\infty}(\mathbb{R}) \cap \mathcal{L}^1$. It is trivial that

$$\mathcal{E}(\mathbb{R}) \cap \mathcal{L}^1 = \left\{ (c_1, c_2) \in (0,\infty) \times \mathbb{R} :\ f(x) = \exp(-c_1 \cdot |x| + c_2) \right\}.$$

But for any f on this form, f' is not continuous in 0, since

$$\lim_{x \to 0^-} f'(x) = c_1 \cdot \exp(c_2) \neq -c_1 \cdot \exp(c_2) = \lim_{x \to 0^+} f'(x).$$

Thus, we must have $f * g \notin \mathcal{E}(\mathbb{R})$. Lemma 3.24 then implies that $f * g \notin \mathcal{E}([a,b])$ for any $a, b \in \mathbb{R}$ with $a < b$. □

Corollary 3.26 tells us, that if we restrict our attention to noise variables that have sufficiently well-behaved densities and transformations – i.e. real analytic densities and density-preserving transformations – any convolution will be non-log-linear. In light of Corollary 3.26, we see that Example 3.21 failed[13] because the noise density, f, is not analytic in any of the points in $\{0, 1/2, 1, \xi\}$ - in fact, it is not even differentiable in these points. Thus, f does not have a convergent power series representation in these points, which causes Theorem 3.25 to fail. Observe also that using real analytic density functions resolves the outstanding matter of differentiability in assumption (A4), which we ignored previously; Theorem 3.25 shows that the convolved density is not only C^3, but in fact C^{∞}.

It is important to be aware that Corollary 3.26 provides sufficient conditions, but that these conditions are not necessary. Indeed, this is shown in Example 3.22, where the chosen noise densities were not real analytic, but their convolution was non-log-linear. In other words, restricting ourselves to real analytic densities provides *a* solution – not *the* solution. It is likely that a version of Corollary 3.26 can be obtained with much weaker conditions. In fact, we only used analyticity twice: Once to bound the derivatives in Theorem 3.25 and once in Lemma 3.24 to Taylor-expand the

[13] Strictly speaking, it was *possible* for it to fail

function f. If a proof of Lemma 3.24 can be found that does not require analyticity, Theorem 3.25 can be proven by simply assuming that f has bounded derivatives up to at least the fourth order. Still, it is not unreasonable to assume that the noise variables in an ANM have real analytic densities. As the following example will show, the class C^{ω}_{+} is non-empty and in fact contains the density of any normal random variable:

Example 3.27. Let $X \sim \mathcal{N}(\mu, \sigma^2) = f \cdot m$. It is trivial that

$$\forall n \in \mathbb{N}: \qquad f^{(n)}(x) = P_n(x) \cdot \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right),$$

where P_n is a polynomial of degree n. Thus $f \in C^{\infty}$ and we find that

$$\forall n \in \mathbb{N}: \qquad \lim_{x \to \pm\infty} f^{(n)}(x) = 0.$$

This implies that

$$\forall n \in \mathbb{N}: \qquad \|f^{(n)}\|_{\infty} < \infty.$$

Since $f > 0$ we conclude that $f \in C^{\omega}_{+}$.

In summary, we have shown during this section, that by introducing two new assumptions – namely that all noise densities are real and analytic, and that every function f belonging to $\mathcal{F}$ is density-preserving – we ensure that a random variable on constructed by

$$\tilde{N}_i := N_i + \sum_{m \in \mathbf{PA}_{\mathcal{G}^0}(X_i) \setminus \mathbf{PA}_{\mathcal{G}^s}(X_i)} f^0_{i,j}(X_m)$$

in an ANM will satisfy assumptions (A3) and (A4).

3.7 PROOF OF OPTIMALITY OF THE GREEDY-ENTROPY SEARCH

We are now ready to prove Theorem 3.11. We start by recalling the statement of Theorem 3.11 here.

Theorem (Theorem 3.11 restated)
Let $\mathscr{C}$ be an ANM with graph $\mathcal{G}^0$ and distribution $\mathbb{P}$. Assume that $\mathcal{G}^0$ is unrelated and assume that $\mathscr{C}$ satisfies assumptions (B1) through (B7). Denote by $\mathcal{G}^{s=k}$ the graph of the Greedy entropy-search at step number k and let $\mathcal{G}^{s=0}$ be a graph with $V(\mathcal{G}^{s=0}) = V(\mathcal{G}^0)$ and $E(\mathcal{G}^{s=0}) = \emptyset$, and let $d := |E(\mathcal{G}^0)|$. Performing a Greedy-entropy search as described in Definition 3.8, *it holds that*

$$\mathcal{G}^{s=d} = \mathcal{G}^0.$$

That is, the true graph $\mathcal{G}^0$ is recovered in d steps.

Proof. Observe that it suffices to prove that a single step of the Greedy-entropy search is correct, as the marginal Gaussian score of every candidate edge must be exactly zero when $\mathcal{G}^{s=d} = \mathcal{G}^0$, by Corollary 3.16.

Fix $k \in \{0, \dots, d-1\}$. To increase readability in the following, we write $\mathcal{G}^s$ in place of $\mathcal{G}^{s=k}$. Assume without loss of generality that the marginal Gaussian score is maximized in either $(\alpha \to \beta)$ or in $(\beta \to \alpha)$, and that the marginal entropy score is maximized in $(\alpha \to \beta)$. That is, assume

$$(\alpha - \beta) = \underset{(\kappa \to \mu) \in R(\mathcal{G}^0)}{\arg\max} \; \Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \kappa \to \mu\right) \tag{3.27}$$

and that

$$\Delta\ell\left(\mathcal{G}^s, \alpha \to \beta\right) > \Delta\ell\left(\mathcal{G}^s, \beta \to \alpha\right). \tag{3.28}$$

To increase readability, we will in the following use the notation

$$\Omega_\alpha := \mathbf{PA}_{\mathcal{G}^s}(\alpha), \qquad \Theta_\alpha := \mathbf{PA}_{\mathcal{G}^0}(\alpha) \backslash \Omega \tag{3.29}$$

to indicate the parents of a node α included in $\mathcal{G}^s$ and the parents which are in $\mathcal{G}^0$ but not yet included in $\mathcal{G}^s$, respectively. Be aware that this notation is local to this proof.

We show that $(\alpha \to \beta)$ lies in $\mathcal{G}^0$ by showing two separate things. In part one we show that the undirected edge $(\alpha - \beta)$ must be in the skeleton of $\mathcal{G}^0$ and in part two that $(\beta \to \alpha)$ cannot be an edge in $\mathcal{G}^0$. Combined, these two yield that $(\alpha \to \beta)$ lies in $\mathcal{G}^0$.

Part 1:
To show that $(\alpha - \beta)$ lies in the skeleton of $\mathcal{G}^0$, we assume for contradiction that $(\alpha - \beta) \notin \mathrm{ske}(\mathcal{G}^0)$. We will consider three exhaustive cases and reach a contradiction in each of them.

Case 1 *- There are no d-connecting paths between α and β in $\mathcal{G}^0$:*
If there are no paths d-connecting α and β in $\mathcal{G}^0$ it follows that $X_\alpha \perp\!\!\!\perp X_\beta$. By Proposition 3.13, this implies that

$$\Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \alpha \to \beta\right) = \Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \beta \to \alpha\right) = 0.$$

As $\mathcal{G}^s$ does not yet contain every edge of $\mathcal{G}^0$, Proposition 3.15 states that there must exist another edge with positive marginal Gaussian score. This contradicts that the marginal Gaussian score was maximized in $(\alpha - \beta)$.

Case 2 *- There is a d-connection between α and β in $\mathcal{G}^0$ through $\boldsymbol{PA}_{\mathcal{G}^0}(\alpha)$:*
Assume throughout all of Case 2, that there exists a d-connection between α and β in $\mathcal{G}^0$ through a true parent of α. Observe first, that there cannot be any d-connections between the two that go through a child of α – if there was, we could combine the two to create a cycle with only one collider. In fact, there must be *exactly* one d-connecting path between α and β. If there were two d-connecting paths between α and β, we could combine them to construct a cycle with at most two

colliders. By Lemma 3.10, an unrelated graph has at least three colliders in every cycle, so this cannot be the case. Denote by ϵ the path that d-connects α and β and let π be the parent of α that lies on ϵ. We will reach a contradiction by showing that we can always find other edges that have a higher marginal Gaussian score than $(\alpha \to \beta)$ and $(\beta \to \alpha)$. To do this, we consider four exhaustive sub-cases. See Figure 3.5 located on Page 59 for example graphs of each subcase.

Case 2.1 - ϵ traverses $\mathbf{PA}_{\mathcal{G}^0}(\beta)$ and $\pi \in \Omega_\alpha$:
Suppose that the d-connecting path ϵ traverses a parent of β and that π belongs to Ω_α. Denote by ρ the parent of β on ϵ. As π is a $\mathcal{G}^s$-parent of α on the only path that d-connects α and β, it must hold that π d-separates α and β in $\mathcal{G}^s$. By the local Markov property, this implies that

$$X_\alpha \perp\!\!\!\perp X_\beta \mid X_\pi. \tag{3.30}$$

Proposition 3.13 then implies that

$$\Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \beta \to \alpha\right) = 0.$$

By the same argument, if $\rho \in \Omega_\beta$ then $\Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \alpha \to \beta\right) = 0$, which would be a contradiction. Suppose then that ρ is *not* a $\mathcal{G}^s$ parent of β. We will show that $\Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \pi \to \beta\right) > \Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \alpha \to \beta\right)$, i.e. that it is in fact better to add the edge $(\pi \to \beta)$ than $(\alpha \to \beta)$. We start by rewriting X_α as

$$X_\alpha = f^0_{\alpha,\pi}(X_\pi) + \tilde{N}_\alpha,$$

where we have set

$$\tilde{N}_\alpha := \sum_{\gamma \in \mathbf{PA}_{\mathcal{G}^0}(\alpha)\setminus\{\pi\}} f^0_{\alpha,\gamma}(X_\gamma) + N_\alpha.$$

Since $\mathcal{G}^0$ is an unrelated graph, it holds that

$$(X_\beta, (X)_{\Omega_\beta}) \perp\!\!\!\perp \tilde{N}_\alpha. \tag{3.31}$$

Thus, we may apply Lemma 3.17 to get

$$\begin{aligned}\mathbb{E}\left(X_\beta + \sum_{\omega\in\Omega_\beta} \hat{f}_\omega(X_\omega) - \hat{f}_\pi(X_\pi)\right)^2 &< \mathbb{E}\left(X_\beta + \sum_{\omega\in\Omega_\beta} \hat{f}_\omega(X_\omega) - \hat{f}_\pi(f^0_{\alpha,\pi}(X_\pi) + \tilde{N}_\alpha)\right)^2 \\ &= \mathbb{E}\left(X_\beta + \sum_{\omega\in\Omega_\beta} \hat{f}_\omega(X_\omega) - \hat{f}_\alpha(X_\alpha)\right)^2. \end{aligned} \tag{3.32}$$

Since we allow for constant functions in the determination of the $\hat{f}$ functions, the residual will always be unbiased. Equation (3.32) then becomes equivalent with

$$\mathbb{V}\left(X_\beta + \sum_{\omega\in\Omega_\beta} \hat{f}_\omega(X_\omega) - \hat{f}_\pi(X_\pi)\right) < \mathbb{V}\left(X_\beta + \sum_{\omega\in\Omega_\beta} \hat{f}_\omega(X_\omega) - \hat{f}_\pi(f^0_{\alpha,\pi}(X_\pi) + \tilde{N}_\alpha)\right)$$

$$= \mathbb{V}\left(X_\beta + \sum_{\omega\in\Omega_\beta} \hat{f}_\omega(X_\omega) - \hat{f}_\alpha(X_\alpha)\right).$$

Hence, we conclude that

$$\Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \pi \to \beta\right) > \Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \alpha \to \beta\right)$$

and that we have reached a contradiction.

Case 2.2 - ϵ traverses $\mathbf{PA}_{\mathcal{G}^0}(\beta)$ and $\pi \in \Theta_\alpha$:
Once again, let ρ be the $\mathcal{G}^0$ parent of β that lies on ϵ. Notice first that equation (3.32) is unchanged when π is not yet included in $\mathcal{G}^s$; the application of Lemma 3.17 hinged on the independence statement in equation (3.30), which remains true when $\pi \in \Theta_\alpha$. We then proceed as in Case 2.1 and use Proposition 3.13 and Lemma 3.17 to conclude that

$$\Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \pi \to \beta\right) \geq (1 - \mathbb{1}_{\rho\in\Omega_\beta}) \cdot \Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \alpha \to \beta\right),$$

with equality if and only if $\rho \in \Omega_\beta$. It then remains to find an edge with a higher score than $(\beta \to \alpha)$. However, we remark that

$$(X_\alpha, (X)_{\Omega_\alpha}) \perp\!\!\!\perp \tilde{N}_\beta,$$

where we have defined

$$\tilde{N}_\beta := \sum_{\gamma\in\mathbf{PA}_{\mathcal{G}^0}(\beta)\setminus\{\rho\}} f^0_{\beta,\gamma}(X_\gamma) + N_\beta. \tag{3.33}$$

This is then exactly symmetric to the statement in equation (3.30). By symmetry, we conclude from Lemma 3.17 that

$$\Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \rho \to \alpha\right) > \Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \beta \to \alpha\right),$$

which is a contradiction.

Case 2.3 - ϵ traverses $\mathbf{CH}_{\mathcal{G}^0}(\beta)$ and $\pi \in \Omega_\alpha$:
Denote by ρ now the $\mathcal{G}^0$-child of β on the path ϵ. When ϵ traverses a child of β, we note that ϵ is a directed path; otherwise there would be a collider on ϵ, and so ϵ would not be a d-connection. As π is included in $\mathcal{G}^s$ it follows that $\Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \beta \to \alpha\right) = 0$. We then need to find an edge that improves the Gaussian score more that $(\alpha \to \beta)$ does. If every edge along ϵ is already included in $\mathcal{G}^s$, then adding $(\alpha \to \beta)$ induces a directed cycle and so $(\alpha \to \beta)$ would not be a candidate edge. Suppose then that there is an edge along ϵ that is not yet included in $\mathcal{G}^s$. We will show that the marginal Gaussian score of adding $(\pi \to \beta)$ must be higher than the marginal score of adding $(\alpha \to \beta)$. By applying the same rewrite as in equation (3.33), we find again that

$$(X_\beta, X_\pi) \perp\!\!\!\perp \tilde{N}_\alpha,$$

as $\mathcal{G}^0$ is unrelated by assumption. The argument is then identical to that of Case 2.1 and we conclude by Lemma 3.17 that

$$\mathbb{V}\left(X_\beta - \sum_{\omega\in\Omega_\beta} \hat{f}_\omega(X_\omega) - \hat{f}_\pi(X_\pi)\right) < \mathbb{V}\left(X_\beta - \sum_{\omega\in\Omega_\beta} \hat{f}_\omega(X_\omega) - \hat{f}_\alpha(X_\alpha)\right).$$

Then

$$\Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \pi \to \beta\right) > \Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \alpha \to \beta\right)$$

and so we have reached a contradiction.

Case 2.4 - ϵ traverses $\mathbf{CH}_{\mathcal{G}^0}(\beta)$ and $\pi \in \Theta_\alpha$:
The argument for the edge $(\alpha \to \beta)$ not maximizing the marginal score is identical to the arguments given in Case 2.1 and 2.3, and so we conclude that

$$\Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \pi \to \beta\right) > \Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \alpha \to \beta\right).$$

We then need to find an edge that has a higher marginal score that $(\beta \to \alpha)$. Here we cannot apply Lemma 3.17, because there is a directed path from β to α in $\mathcal{G}^0$, which implies that $\tilde{N}_\alpha \not\perp\!\!\!\perp X_\beta$.[14] Instead, we exploit the independence structure of $\mathscr{C}$ to get

$$\begin{aligned}
&\mathbb{V}\left(X_\alpha - \sum_{\omega\in\Omega_\alpha} \hat{f}_\omega(X_\omega) \; - \hat{f}_\beta(X_\beta)\right) \\
&\qquad\qquad = \mathbb{V}\left(N_\alpha + \sum_{\gamma\in\mathbf{PA}_{\mathcal{G}^0}(\alpha)\setminus\{\pi\}} f^0_{\alpha,\gamma}(X_\theta) + f^0_{\alpha,\pi}(X_\pi) - \sum_{\omega\in\Omega_\alpha} \hat{f}_\omega(X_\omega) - \hat{f}_\beta(X_\beta)\right) \\
&\qquad\qquad \overset{(a)}{=} \mathbb{V}\left(N_\alpha + \sum_{\gamma\in\mathbf{PA}_{\mathcal{G}^0}(\alpha)\setminus\{\pi\}} f^0_{\alpha,\gamma} - \sum_{\omega\in\Omega_\alpha} \hat{f}_\omega(X_\omega)\right) + \mathbb{V}\left(f^0_{\alpha,\pi}(X_\pi) - \hat{f}_\beta(X_\beta)\right) \\
&\qquad\qquad > \mathbb{V}\left(N_\alpha + \sum_{\gamma\in\mathbf{PA}_{\mathcal{G}^0}(\alpha)\setminus\{\pi\}} f^0_{\alpha,\gamma}(X_\gamma) - \sum_{\omega\in\Omega_\alpha} \hat{f}_\omega(X_\omega)\right) \\
&\qquad\qquad = \mathbb{V}\left(X_\alpha - f^0_{\alpha,\pi}(X_\pi) - \sum_{\omega\in\Omega_\alpha} \hat{f}_\omega(X_\omega)\right) \\
&\qquad\qquad \geq \mathbb{V}\left(X_\alpha - \hat{f}_\pi(X_\pi) - \sum_{\omega\in\Omega_\alpha} \hat{f}_\omega(X_\omega)\right).
\end{aligned} \tag{3.34}$$

To get equality (a), we use that $\mathcal{G}^0$ is unrelated and that α must be a collider on any path between β and a node in $\mathbf{PA}_{\mathcal{G}^0}(\alpha)\setminus\{\pi\}$. That is,

$$((X)_{\mathbf{PA}_{\mathcal{G}^0}(\alpha)\setminus\{\pi\}}, N_\alpha) \perp\!\!\!\perp (X_\pi, X_\beta),$$

[14] With $\tilde{N}_\alpha$ defined as in (3.33)

which allows us to split the variance into a sum of variances. We conclude by equation (3.34) that

$$\Delta\ell^{\mathrm{g}}(\mathcal{G}^s, \pi \to \alpha) > \Delta\ell^{\mathrm{g}}(\mathcal{G}^s, \beta \to \alpha)$$

and so we have reached a contradiction.

Case 3 *- There is a d-connection between α and β through $\boldsymbol{CH}_{\mathcal{G}^0}(\alpha)$:*
Assume now the existence of a path ϵ that d-connects α and β in $\mathcal{G}^0$ through a true child of α. As in Case 2, ϵ must be the *only* d-connecting path between α and β, as $\mathcal{G}^0$ is an unrelated graph. Because ϵ traverses a child of β, the path must in fact be directed and therefore traverse a true parent of β. Case 3 is then symmetric to Cases 2.3 and 2.4, and we conclude that $\Delta\ell^g$ cannot be maximized in $(\alpha \to \beta)$ nor in $(\beta \to \alpha)$.

In summary, we have throughout Cases 1, 2 and 3 shown that if α and β are d-connected in $\mathcal{G}^0$ but $(\alpha - \beta)$ does not lie in ske$(\mathcal{G}^0)$, then $\Delta\ell^g$ cannot be maximized in $(\alpha \to \beta)$ nor in $(\beta \to \alpha)$. This is a contradiction to the assumption that the edge $(\alpha - \beta)$ was selected. Thus, we conclude that the undirected edge $(\alpha - \beta)$ must lie in the skeleton of $\mathcal{G}^0$.

Part 2 - show that $(\alpha \to \beta) \in \mathcal{G}^0$:
Assume for contradiction that $(\alpha \to \beta)$ is *not* an edge in $\mathcal{G}^0$. As we have shown that $(\alpha - \beta) \in$ ske$(\mathcal{G}^0)$, this must imply that $(\beta \to \alpha) \in E(\mathcal{G}^0)$. Consider then the subgraph $\tilde{\mathcal{G}} := (\tilde{V}, \tilde{E})$ of $\mathcal{G}^s$ in which we set

$$\begin{aligned}
\tilde{V} &= \{\alpha\} \cup \{\beta\} \cup \Omega_\alpha \cup \Omega_\beta \\
\tilde{E} &= (\beta \to \alpha) \cup \bigcup_{\gamma \in \mathbf{PA}_{\mathcal{G}^s}(\alpha)} (\gamma \to \alpha) \cup \bigcup_{\delta \in \mathbf{PA}_{\mathcal{G}^s}(\beta)} (\delta \to \beta).
\end{aligned}$$

That is, $\tilde{\mathcal{G}}$ is the graph consisting of α, β and their $\mathcal{G}^s$-parents, with an edge added from β into α. This subgraph can be thought of as the graph of an ANM, $\tilde{\mathscr{C}}$, with assignments

$$\forall \nu \in V(\tilde{\mathcal{G}}) : \qquad X_\nu := \sum_{\gamma \in \mathbf{PA}_{\tilde{\mathcal{G}}}(\nu)} f^0_{\nu,\gamma}(X_\gamma) + \tilde{N}_\nu,$$

where

$$\tilde{N}_\nu = \sum_{\gamma \in \mathbf{PA}_{\mathcal{G}^0}(\nu) \setminus \mathbf{PA}_{\tilde{\mathcal{G}}}(X_\nu)} f^0_{\nu,\gamma}(X_\gamma) + N_\nu. \tag{3.35}$$

As $\mathcal{G}^0$ is unrelated, so is $\tilde{\mathcal{G}}$ which means that

$$\forall \nu \in V(\tilde{\mathcal{G}}) : \qquad \mathbf{PA}_{\mathcal{G}^0}(\nu) \setminus \mathbf{PA}_{\tilde{\mathcal{G}}}(\nu) \perp_d \mathbf{PA}_{\tilde{\mathcal{G}}}(\nu).$$

That is, $(\tilde{N})_{V(\tilde{\mathcal{G}})}$ is a sequence of mutually independent random variables, and so $\tilde{\mathscr{C}}$ is indeed an ANM. Furthermore, it follows from Lemma 3.20 and Corollary 3.26 that each $\tilde{N}$ has a density that is strictly positive and real analytic on $\mathbb{R}$ and therefore also in C^3. Then $\tilde{\mathscr{C}}$ satisfies assumptions

(A1) through (A6), and $\tilde{\mathscr{C}}$ must therefore be identifiable, according to Theorem 2.3. Consider then the graph $\tilde{\mathcal{G}}_{\alpha\to\beta}$, in which we flip the direction of the edge $(\beta \to \alpha)$. By Theorem 3.7 we must have

$$\ell(\tilde{\mathcal{G}}) - \ell(\tilde{\mathcal{G}}_{\alpha\to\beta}) > 0. \tag{3.36}$$

The two graph scores in equation (3.36) only differ in α and β, as the two graphs are equal everywhere else. Expanding the difference, we find that

$$\begin{aligned}
&-\mathbb{H}\Big(X_\alpha - \sum_{\omega\in\Omega_\alpha} \hat{f}_\omega(X_\omega) - \hat{f}_\beta(X_\beta)\Big) - \mathbb{H}\Big(X_\beta - \sum_{\omega\in\Omega_\beta} \hat{f}_\omega(X_\omega)\Big) \\
&\qquad + \mathbb{H}\Big(X_\beta - \sum_{\omega\in\Omega_\beta} \hat{f}_\omega(X_\omega) - \hat{f}_\alpha(X_\alpha)\Big) + \mathbb{H}\Big(X_\alpha - \sum_{\omega\in\Omega_\alpha} \hat{f}_\omega(X_\omega)\Big) > 0 \\
&\Leftrightarrow \Big[\mathbb{H}\Big(X_\alpha - \sum_{\omega\in\Omega_\alpha} \hat{f}_\omega(X_\omega)\Big) - \mathbb{H}\Big(X_\alpha - \sum_{\omega\in\Omega_\alpha} \hat{f}_\omega(X_\omega) - \hat{f}_\beta(X_\beta))\Big] \\
&\qquad - \Big[\mathbb{H}\Big(X_\beta - \sum_{\omega\in\Omega_\beta} \hat{f}_\omega(X_\omega)\Big) - \mathbb{H}\Big(X_\beta - \sum_{\omega\in\Omega_\beta} \hat{f}_\omega(X_\omega) - \hat{f}_\alpha(X_\alpha)\Big)\Big] > 0.
\end{aligned}$$

But this exactly means that

$$\Delta\ell\left(\mathcal{G}^s, \beta \to \alpha\right) - \Delta\ell\left(\mathcal{G}^s, \alpha \to \beta\right) > 0$$

which contradicts the assumption that $\Delta\ell\left(\mathcal{G}^s, \alpha \to \beta\right) > \Delta\ell\left(\mathcal{G}^s, \beta \to \alpha\right)$. Thus, we must have $(\alpha \to \beta) \in E(\mathcal{G}^s)$, which concludes the proof. □

Remark 3.28. In Case 2.1 of the proof of Theorem 3.11 we relied on Lemma 3.17 to reach a contradiction. It is possible to reach a contradiction here without using Lemma 3.17 and instead exploiting the assumed independence structure of the model. In fact, because of the symmetry of the cases we only need to apply Lemma 3.17 in Cases 2.3 and 2.4 to show that the edge $(\pi \to \beta)$ has a higher score than $(\alpha \to \beta)$. We remark this because the application of Lemma 3.17 requires assumption (B7), which is not used anywhere else. Thus, if a different method of showing that $(\pi \to \beta)$ scores higher than $(\alpha \to \beta)$ in Cases 2.3 and 2.4 is found, assumption (B7) can be omitted. However, we do not have a different method of reaching this contradiction, and so we stick to applying Lemma 3.17 in all cases. We provide in Appendix B.2 the alternative way of proving Case 2.1.

The reader may have observed that the assumption of $\mathcal{F}$ being a class of non-linear functions was made only with the goal of having identifiability of the model. However, as we discussed during Chapter 2, non-linearity is not the only way to obtain identifiability. In fact, we can prove Theorem 3.11 when the function class consists only of linear functions by imposing one additional assumption.

Corollary 3.29
Let $\mathscr{C}$ be an ANM with graph $\mathcal{G}^0$ and assignments on the form

$$\forall \nu \in V(\mathcal{G}^0): \qquad X_\nu := \sum_{\gamma \in \boldsymbol{PA}_{\mathcal{G}^0}(\nu)} (a_{\nu,\gamma} + b_{\nu,\gamma} \cdot X_\gamma) + N_\nu,$$

for real coefficients a and real, non-zero coefficients b. Assume that $\mathscr{C}$ satisfies assumptions (B3) and (B5) of Section 3.5. Assume furthermore that for every $\nu \in V(\mathcal{G}^0)$, the random variable N_ν is not normally distributed, nor does it follow a log-mix-lin-exp distribution. If it holds true that every linear combination of noise variables in $\mathscr{C}$ is again neither normally nor log-mix-lin-exp distributed, then $\mathcal{G}^0$ can be recovered by the Greedy entropy-search.

Proof. Let $\mathcal{F}'$ be the class of linear functions with non-vanishing slopes. Clearly $\mathcal{F}' \subseteq C^3$ and for every $f \in \mathcal{F}'$ it holds true that f' does not vanish in the limits. Identifiability of $\mathscr{C}$ and every sub-model of $\mathscr{C}$ as constructed in part two of the proof of Theorem 3.11 is then ensured by Proposition 2.2. Assumptions (B4), (B5) and (B7) are all trivially satisfied when all assignments are linear. The statement then follows by re-doing the proof of Theorem 3.11; part one is identical and in part two we get identifiability of the subgraph by using Corollary 3.26 and the assumptions on the family of noise distribution. □

The assumption in Corollary 3.29 that no two noise variables convolve into a normal distribution or a log-mix-lin-exp distribution may seem rather strict. Given a family of noise distributions, however, it is quite easy to check whether this conditions holds true by using characteristic functions. Let ρ_Z denote the characteristic function of any random variable, Z, and consider a fixed node, ν. We can rewrite X_ν as

$$\begin{aligned}
X_\nu &= \sum_{\gamma \in \mathbf{PA}_{\mathcal{G}^0}(\nu)} (a_{\nu,\gamma} + b_{\nu,\gamma} \cdot X_\gamma) + N_\nu \\
&= \sum_{\gamma \in \mathbf{PA}_{\mathcal{G}^0}(\nu)} \left(a_{\nu,\gamma} + b_{\nu,\gamma} \cdot \sum_{\mu \in \mathbf{PA}_{\mathcal{G}^0}(\gamma)} (a_{\gamma,\mu} + b_{\gamma,\mu} \cdot X_\mu) \right) + N_\nu \\
&= \dots \\
&= \sum_{\gamma \in \mathbf{AN}_{\mathcal{G}^0}(\nu)} (\tilde{a}_{\nu,\gamma} + \tilde{b}_{\nu,\gamma} \cdot N_\gamma) + N_\nu,
\end{aligned}$$

for some coefficients $\tilde{a}$ and $\tilde{b}$. By properties of characteristic functions (see Sokol and Rønn-Nielsen 2016, Lemma 3.4.9 and Lemma 3.4.15) we get

$$\rho_{X_\nu}(s) = \prod_{\gamma \in \mathbf{AN}_{\mathcal{G}^0}(\nu)} \exp(i \cdot \tilde{a}_{\nu,\gamma} \cdot s) \cdot \rho_{N_\gamma}(\tilde{b}_{\nu,\gamma} \cdot s) \cdot \rho_{N_\nu}(s). \tag{3.37}$$

If we know the characteristic function of every noise variable, we can then check whether every product on the form (3.37) (for differing candidate graphs) equals the characteristic function of a normal distribution or a log-mix-lin-exp distribution, both of which have easily computable closed-form expressions. In particular, the characteristic function of normally distributed variable, $Z \sim \mathcal{N}(\mu, \sigma^2)$, is given by

$$\rho_Z(s) = \exp(i \cdot \mu \cdot s - \frac{1}{2}\sigma^2 \cdot s^2). \tag{3.38}$$

Similarly, if W is log-mix-lin-exp distributed, then

$$\rho_W(s) = \cdot \exp(c_4) \frac{(-c_1)^{(i \cdot s - c_3)/c_2} \cdot \Gamma\left(\frac{i \cdot s - c_3}{c_2}\right)}{|c_2|}, \tag{3.39}$$

where $c_1 < 0$, $c_2 \cdot c_3 > 0$ and $c_4 \in \mathbb{R}$ are constants. These expressions can be readily computed in any language that can compute Fourier transforms. Then, if we are given an ANM with a fixed noise distribution and linear assignments we can check whether whether the graph is recoverable by our Greedy entropy-search. Formally, this relies on uniqueness of characteristic functions – see Sokol and Rønn-Nielsen [2016, Theorem 3.4.19] for a reference.

3.8 REMARKS ON THE GREEDY ENTROPY-SEARCH METHOD

We close this chapter with a few remarks about using the Greedy entropy-search and provide an example of an ANM in which conditions (B1) through (B7) are satisfied.

In the statement of Theorem 3.11 we assumed that we start the Greedy entropy-search from an empty graph. However, we can just as well begin the search from any other true subgraph of $\mathcal{G}^0$ as every step is identical in the proof. We highlight this property because it essentially allows for incorporation of prior knowledge in learning the structure of a causal system. In many settings, this is a desirable property as subject matter experts will often have some prior knowledge of the system – e.g. on the basis of randomized trials. The possibility of including prior knowledge becomes relevant in a finite-data setting. When we only observe part of the models distribution, we have to estimate the marginal scores when we perform the Greedy entropy-search. Every estimation comes with an error probability that is propagated through every step of the search. If we can include prior knowledge[15] – i.e. start from a graph that is closer to the true graph – the burden of estimation lessens and the probability of recovering the true graph increases.

We claimed in Chapter 1 after proving Proposition 1.15 that it would turn out that every direct connection in the graph of an ANM that is recoverable by a Greedy entropy-search is causal. Two key ingredients went into the proof of Proposition 1.15; the structural equation should not be constant and an independence relationship between parents needed to be satisfied. In proving optimality of the Greedy entropy-search we have, almost inadvertently, assumed both of these conditions be to true. Indeed, we state the following theorem:

[15] Provided that this prior knowledge is in fact correct.

Theorem 3.30
Let $\mathscr{C}$ be an ANM satisfying assumptions (B1) through (B7). Let $\mathcal{G}^0$ be the graph of $\mathscr{C}$ and assume this to be unrelated. For all α and β in $V(\mathcal{G}^0)$ it holds that

$$(\alpha \to \beta) \in E(\mathcal{G}^0) \Rightarrow X_\alpha \overset{c}{\to} X_\beta.$$

Proof. Fix α and β in $V(\mathcal{G}^0)$ such that $(\alpha \to \beta)$ is an edge in $\mathcal{G}^0$. By assumption (B1), the structural function $f^0_{\beta,\alpha}$ is constant nowhere on the real line. Furthermore, the graph $\mathcal{G}^0$ is unrelated by assumption which means that

$$X_\alpha \perp\!\!\!\perp \mathbf{PA}_{\mathcal{G}^0}(X_\beta)\backslash\{X_\alpha\}.$$

It then follows from Proposition 1.15 that $X_\alpha \overset{c}{\to} X_\beta$. □

The implication of Theorem 3.30 is, that if we are given a distribution from an ANM for which we can recover the graph, then every direct connection in the recovered graph must in fact represent a causal relationship. That is, if we are given a distribution which we truly believe is generated by an ANM that satisfies assumptions (B1) through (B7), then we may draw causal inference from the recovered graph. However, it still remains to show that such a distribution exists.

It is not difficult to think of distributions that have full support, real analytic densities and finite second moment. Below, we list a few such distributions in Table 3.1. We showed in Example 3.27 that the normal distribution satisfies assumptions (B3) and (B5). The Gumbel distribution has a strictly positive, non-log-linear density and has second moment [Weisstein, b]. Furthermore, we see that its density function is a product of compositions of C^ω_+-functions, which implies that the distribution function is itself in C^ω_+ [Encyclopedia of Mathematics., b]. Thus, the Gumbel distribution satisfies assumptions (B3) and (B5). Note, however, that the Gumbel distribution is a special case of a log-mix-lin-exp distribution and so an ANM with Gumbel-distributed noise and linear assignments is not identifiable[16]. The logistic distribution has finite second moment [Weisstein, c] and its density is clearly non-log-linear. The density of the logistic distribution is a product of C^ω_+-functions and therefore in C^ω_+ itself [Encyclopedia of Mathematics., b]. Finally, the hyperbolic secant distribution has finite second moment (see Ding 2014) and the hyperbolic functions are all analytic by construction. This list of distributions is by no means exhaustive; it is merely included here to stress that the assumption that all noise variables have an analytic density is not overly restrictive. Observe also also that the Hyperbolic secant distribution has characteristic function $\operatorname{sech}(s)$. The product of finitely many functions on the form $s \mapsto \operatorname{sech}(b \cdot s)$ cannot take the form of either equation (3.38) or (3.39). This implies that an ANM with linear assignments and Hyperbolic secant-noise can be recovered by the Greedy entropy-search.

Similarly, it is not difficult to find examples of function classes which satisfy assumptions (B1), (B2) and (B4). The difficult part is to show that assumptions (B6) and (B7) are satisfied for a

[16] Recall Proposition 2.2.

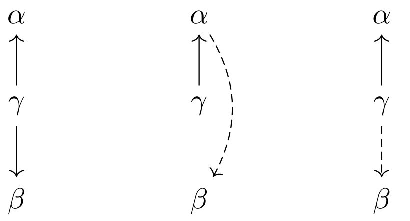

Figure 3.2: *Left: The graph $\mathcal{G}$. Middle: The graph $\mathcal{G}^{s=1}_{\alpha\to\beta}$. Right: The graph $\mathcal{G}^{s=1}_{\gamma\to\beta}$.*

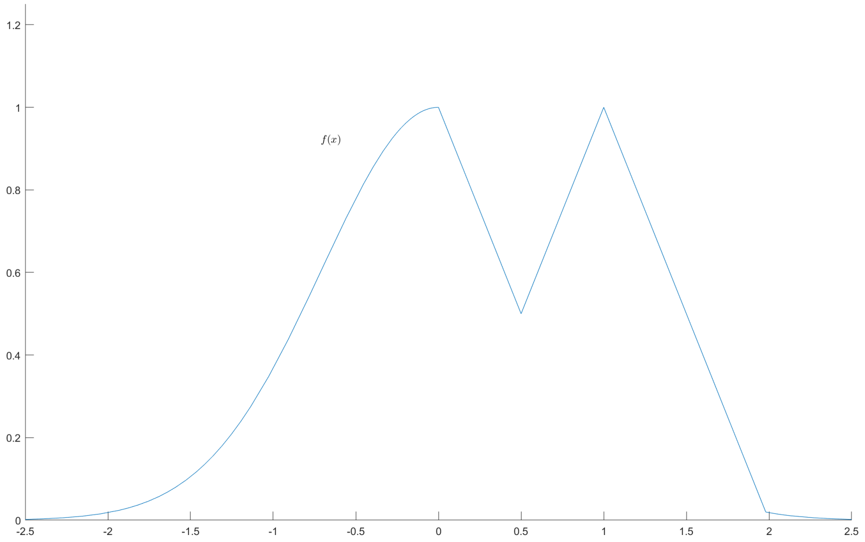

Figure 3.3: *The unnormalized density of M, $f(x)$.*

Table 3.1: *A selection of distributions which satisfy assumptions (B3) and (B5)*

Distribution	**Density function**	**Parameters**
Normal	$f(x) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$\mu \in \mathbb{R}, \sigma > 0$
Gumbel	$f(x) = \frac{1}{\beta} \cdot \exp\left(\frac{x-\alpha}{\beta} - \exp\left(\frac{x-\alpha}{\beta}\right)\right)$	$\alpha \in \mathbb{R}, \beta > 0$
Logistic	$f(x) = \frac{\exp(-(x-m)/b)}{b \cdot \left(1+\exp(-(x-m)/b)\right)^2}$	$m \in \mathbb{R}, b > 0$
Hyperbolic secant	$f(x) = \frac{1}{2}\ \mathrm{sech} \cdot \left(\frac{\pi}{2}\, x\right)$	

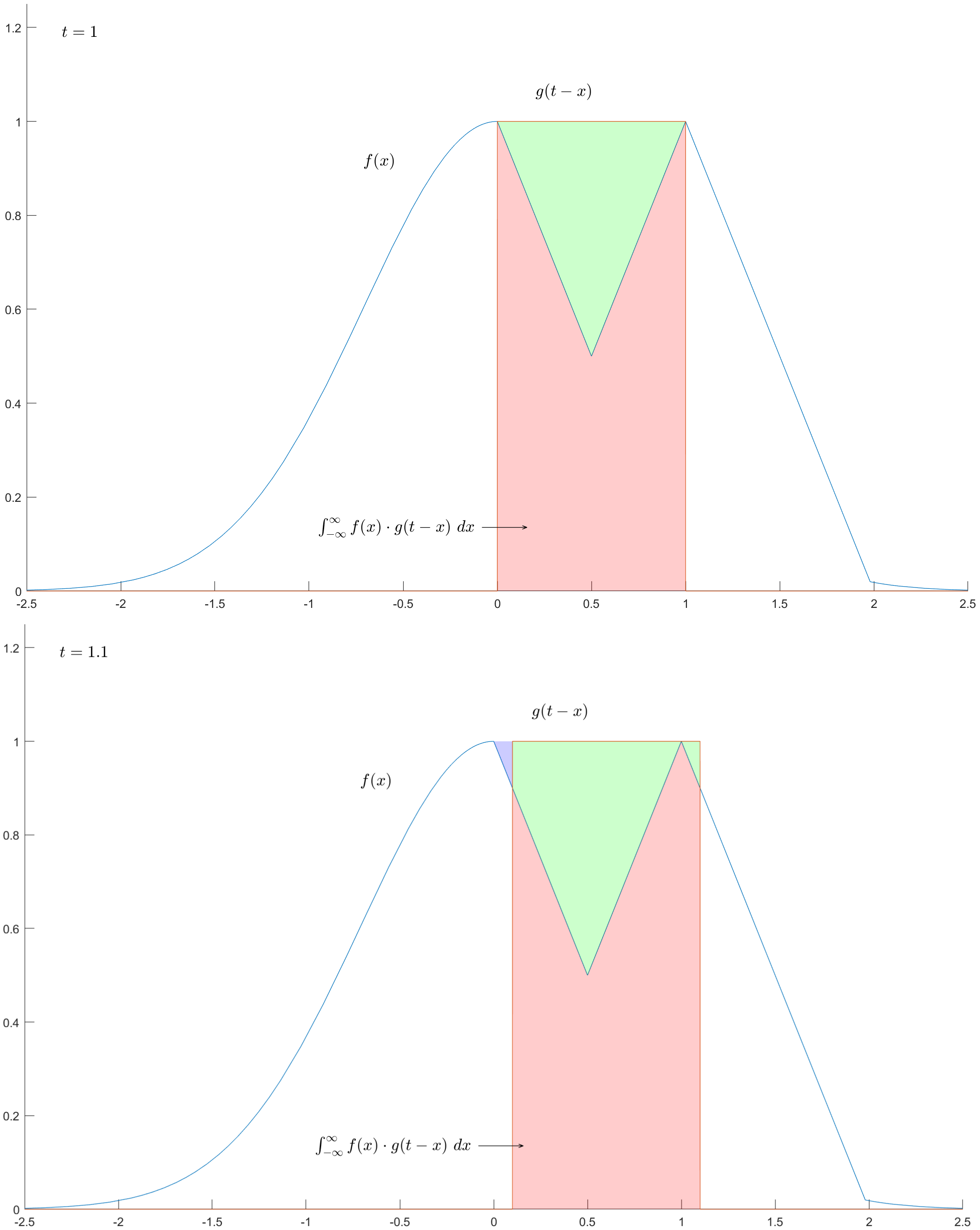

Figure 3.4: *An illustration of the convolution $(f * g)(t)$ in Example 3.21 for $t = 1$ and $t = 1.1$. Top: Blue curve: $f(x)$. Red curve: $g(t-x)$. Red area: $(f * g)(t)$. Green area: The set A. Bottom: Blue and red curves as above and red area as above. Green area: A'. By shifting the right-most green triangle along the x-axis by -1 into the blue area, the set A is obtained.*

specific combination of a function class and a family of noise distributions. Still, we can construct simple examples in which all assumptions are satisfied. The following provides an example of a model with non-linear functions that satisfies the assumptions of Theorem 3.11.

Example 3.31. Let $\mathscr{C}$ be an ANM generated as in (2.1). Let $\mathcal{F}$ be the class of all polynomials that have finite degree of at least two. Assume that every noise component, N_ν, follows a normal distribution with any mean and variance. We do not assume that every noise component has the same mean and variance. We argue that $\mathscr{C}$ satisfies the assumptions (B1) through (B7). Assumptions (B1) and (B2) hold trivially. Assumption (B3) holds by Example 3.27. To show (B4), take f in $\mathcal{F}$ and let $(r_i)_{i=1}^N$ denote the roots of f'. As f has degree at least two, there exists at least one such root and as f has a finite degree there can only be finitely many such roots. Let $(\tilde{r}_i)_{i=1}^n$ be an ordering of $(r_i)_{i=1}^N$ such that $\tilde{r}_1 < \cdots < \tilde{r}_N$. We can then construct a partition of $\mathbb{R}$ as follows[17]: Let

$$\mathcal{X}_1 \coloneqq (-\infty, \tilde{r}_1),$$

and for all i between 2 and N

$$\mathcal{X}_i \coloneqq [\tilde{r}_{i-1}, r_i)$$

and lastly

$$\mathcal{X}_{N+1} \coloneqq [\tilde{r}_N, \infty).$$

Then $(\mathcal{X}_i)_{i=1}^{N+1}$ is a partition of $\mathbb{R}$ and f must be strictly monotonic on every $\mathcal{X}_i$ and thus also invertible on $\mathcal{X}_i$. This shows that assumption (B4) holds. Assumptions (B5) and (B6) are satisfied as a normal distribution has moments of all orders. Take now $f, g \in \mathcal{F}$ and a random variable N that is normally distributed[18]. Let n be the degree of f and let m be the degree of g. By the Binomial Theorem [Weisstein, a] the composite function $f(g(x) + N)$ must then be a polynomial of degree $n \cdot m$, which implies that $\mathbb{E}f(g(x) + N)$ is a polynomial of degree $n \cdot m$ in which the coefficients are allowed to depend on the moments of N up to order n. That is, $\mathbb{E}f(g(x)+N)$ is itself an element of $\mathcal{F}$, and so assumption (B7) is satisfied. We conclude that an ANM with normally distributed noise, polynomial functions and an unrelated graph is recoverable by the Greedy entropy-search.

We close this chapter by remarking that Example 3.31 holds true for any choice of noise variables that have C_+^ω-density and moments of all orders.

[17] If $N = 1$ we simply omit the middle step.

[18] We can choose such an N trivially, because any linear combination of normal random variables is normally distributed.

$$\begin{tikzcd}
\Omega_\beta \arrow[d] & & & \Omega_\alpha\backslash\{\pi\} \arrow[d] \\
\beta & \rho \arrow[l, dashed] \arrow[r, leftrightsquigarrow, "d"] & \pi \arrow[r] & \alpha \\
\Theta_\beta\backslash\{\rho\} \arrow[u, dotted] & & & \Theta_\alpha \arrow[u, dotted]
\end{tikzcd}$$

Case 2.1

$$\begin{tikzcd}
\Omega_\beta \arrow[d] & & & \Omega_\alpha\backslash\{\pi\} \arrow[d] \\
\beta & \rho \arrow[l, dashed] \arrow[r, leftrightsquigarrow, "d"] & \pi \arrow[r, dotted] & \alpha \\
\Theta_\beta\backslash\{\rho\} \arrow[u, dotted] & & & \Theta_\alpha \arrow[u, dotted]
\end{tikzcd}$$

Case 2.2

$$\begin{tikzcd}
\Omega_\beta \arrow[d] & & & \Omega_\alpha\backslash\{\pi\} \arrow[d] \\
\beta \arrow[r, dashed] & \rho \arrow[r, dashed, mapsto] & \pi \arrow[r] & \alpha \\
\Theta_\beta \arrow[u, dotted] & & & \Theta_\alpha \arrow[u, dotted]
\end{tikzcd}$$

Case 2.3

$$\begin{tikzcd}
\Omega_\beta \arrow[d] & & & \Omega_\alpha\backslash\{\pi\} \arrow[d] \\
\beta \arrow[r, dashed] & \rho \arrow[r, dashed, mapsto] & \pi \arrow[r, dotted] & \alpha \\
\Theta_\beta \arrow[u, dotted] & & & \Theta_\alpha \arrow[u, dotted]
\end{tikzcd}$$

Case 2.4

Figure 3.5: *Example graphs of cases 2.1 through 2.4. Dotted lines indicate edges that are in $\mathcal{G}^0$ but not $\mathcal{G}^s$ and full lines indicate an edge that is in $\mathcal{G}^s$. Double-sided squiggly lines are d-connections of indeterminate length. Dashed lines indicate that the edge may be in $\mathcal{G}^s$ but not necessarily.*

CHAPTER 4

IMPLEMENTING THE GREEDY ENTROPY-SEARCH

The previous chapter was concerned with recovering the graph of an ANM, given full knowledge of its implied distribution. However, it is never realistic to assume that we know the full distribution of anything in a real-life setting. In this chapter we discuss the transition from theory to practice when using the Greedy entropy-search and highlight some of the difficulties in doing so. We do not provide a proof of consistency, but give a brief discussion on the subject and why it is difficult to prove consistency. Instead, we implement a version of the Greedy entropy-search in R, disregarding any potential consistency issues, and give an empirical account of its performance based on simulations. In the process of implementation, we slightly modify the Greedy entropy-search algorithm to account for the uncertainties associated with inference on finite samples.

NOTATION

To facilitate the discussion of implementation and estimation, we first adapt the notation that was used in the previous chapters. Because our discussion of estimation will be a conceptual one, we are less rigorous with the theory than we were in the previous chapters, and so we relax the notation accordingly. Given an ANM $\mathscr{C}$ and a graph $\mathcal{G}$, we define $P_\nu^{\mathcal{G}} := \mathbf{PA}_{\mathcal{G}}(X_\nu)$ to be the vector of X_ν's $\mathcal{G}$ parents. If X_ν has no parents, the convention is that $P_\nu^{\mathcal{G}} = 0$ and $\hat{f}(P_\nu^{\mathcal{G}}) = 0$ for any function. Given a sample from an ANM, $((X)^i_{V(\mathcal{G})})_{i=1}^n$, we use bold symbols to indicate the vector (or matrix) of samples. That is, we write $\boldsymbol{X}_\nu$ to denote the n samples of X_ν and so forth. As in Chapter 3, we use ϕ to denote conditional expectations. In particular, we write $\phi_{X_\nu|P_\nu^{\mathcal{G}}}$ to denote the mapping $z \mapsto \mathbb{E}[X_\nu \mid P_\nu^{\mathcal{G}} = z]$. We let $\hat{\phi}$ be the output function of a generic regression technique. We define the residuals from regressing X_ν onto its $\mathcal{G}$-parents by

$$\hat{\boldsymbol{N}}_\nu^{\mathcal{G}} := \boldsymbol{X}_\nu - \hat{\phi}_{X_\nu|P_\nu^{\mathcal{G}}}(\boldsymbol{P}_\nu^{\mathcal{G}}).$$

When there is no danger of confusion, we omit the superscript $\mathcal{G}$ from $\hat{\boldsymbol{N}}_\nu^{\mathcal{G}}$.

4.1 DIFFICULTIES IN PROVING CONSISTENCY

Suppose we are given a finite, i.i.d. sample of data, say of size n. We do not know the distribution, $\mathbb{P}$, that generated this data, but we are told that it comes from an ANM, $\mathscr{C}$, that satisfies the assumptions of Theorem 3.11. We then know that it is possible to recover the graph of $\mathscr{C}$ using the Greedy entropy-search. However, as we only observed a finite part of $\mathbb{P}$, we cannot explicitly compute the marginal scores under $\mathbb{P}$. Instead, we must turn to estimation. The first step to using the Greedy entropy-search for graph recovery in a finite-data setting is then to construct estimators of the marginal scores. Suppose that we are given consistent estimators of variance and entropy, respectively, and a consistent regression technique[1]. Denote these by $\hat{\mathbb{H}}$, $\hat{\mathbb{V}}$ and $\hat{\phi}$ respectively. The natural estimators of the marginal scores are then

$$\widehat{\Delta\ell}(\mathcal{G}, \alpha \to \beta) := \hat{\mathbb{H}}(\hat{\boldsymbol{N}}_\nu^{\mathcal{G}}) - \hat{\mathbb{H}}(\hat{\boldsymbol{N}}_\nu^{\mathcal{G}_{\alpha\to\beta}})$$

and

$$\widehat{\Delta\ell}^g(\mathcal{G}, \alpha \to \beta) := \log \hat{\mathbb{V}}(\hat{\boldsymbol{N}}_\nu^{\mathcal{G}}) - \log \hat{\mathbb{V}}(\hat{\boldsymbol{N}}_\nu^{\mathcal{G}_{\alpha\to\beta}}).$$

Since all three estimators are consistent, we could hope that $\widehat{\Delta\ell}$ and $\widehat{\Delta\ell}^g$ are as well. Unfortunately, we cannot guarantee that they will be. Informally speaking, the consistency of $\hat{\mathbb{H}}$ and $\hat{\mathbb{V}}$ may be 'lost' in a sense, when acting on the estimated residuals, which are *close* to the true residuals, but never equal. To get a sense of why this may be the case, we can think about an analogy from real analysis:

[1] In the confines of this discussion, it does not matter whether we take consistency to mean almost sure convergence or convergence in probability.

Given two sequences of functions, $(f_n)_{n\in\mathbb{N}}$ and $(g_n)_{n\in\mathbb{N}}$, both of which converge uniformly with limits f and g respectively, we may ask ourselves if $f_n \circ g_n$ converges uniformly – or even just pointwise. In general, the answer to this question is no[2]. To get convergence of the composite function, we need to impose further restrictions on the sequence of functions (here by assuming uniform continuity). In the same vein, consistency of each estimator is not enough to guarantee consistency of the plug-in estimators [Kpotufe et al., 2013]. A similar case to ours is considered in Kpotufe et al. [2013], in which a proof of consistency for the composite estimator $\hat{\mathbb{H}}(\hat{\boldsymbol{N}}_\nu^{\mathcal{G}})$ is given by assuming boundedness of the structural assignments and that the densities obey a tail power-law. Unfortunately, their result does not apply here, since they also assume a bounded support of one of the data-generating processes. In contrast, we have assumed that all noise distributions have full support[3]. However, it is unclear how vital the assumption of boundedness is throughout the proof and if the proof can be modified to still hold true without this assumption. We do not attempt this, but simply remark that it does not seem out of the realm of possibility, given that their remaining assumptions are all rather mild. Perhaps the most important takeaway from Kpotufe et al. [2013] is that they employ sample splitting in order to achieve consistency of the entropy estimator. That is, they split their sample into two parts. The first one is used to estimate the regression function. From this, they *predict* the residuals on the other sample which they then use to estimate the entropy. Performing the estimation on split samples has the advantage that the resulting sequence of estimated residuals is once again an i.i.d. sequence of random variables, making it easier to achieve consistency – or at least *prove* consistency. On the other hand, if the chosen regression technique has poor predictive power, the estimation procedure can potentially perform worse in a finite-sample setting, when the sample is split into two. Furthermore, by employing sample splitting we are introducing the undesirable property of seed-dependence into the estimation procedure. That is, running the same estimation on the same dataset twice can potentially yield two different conclusions, depending on how the sample gets split. Thus, the necessity of sample splitting is unclear, and may even vary from case to case. Kpotufe et al. [2013] also give a proof of consistency without sample splitting for a specific estimator by imposing stronger conditions on the data-generating process. Thus, it is possible to achieve consistency without sample splitting – but undoubtedly more difficult. We leave the discussion of consistency here and instead implement a version of our Greedy entropy-search. We then assess its performance with and without employing sample splitting in a simulation scenario.

4.2 CONSIDERATIONS IN IMPLEMENTING THE GREEDY ENTROPY-SEARCH

In this section we revisit the Greedy entropy-search as described in Algorithm 1. As we leave the population case and move into a finite-sample setting, we need to adjust our algorithm accordingly. In particular, we need to adjust to the exit condition; in the proof of Theorem 3.11 we used

[2] Pick for example $f_n(x) = \mathbb{1}_{(0,\infty)}(x)$ and $g_n(x) = x/n$.

[3] Recall Assumptions (A4) and/or (B3)

Proposition 3.13 [4] to conclude that the Greedy entropy-search would indeed stop running when it had reached the true graph. In a finite-sample setting, it is unlikely that we will ever estimate the marginal Gaussian score to be *exactly* zero. Even if we are adding an edge between nodes that are independent – not just conditionally – it is possible that the regression improves ever so slightly, simply by chance. The marginal Gaussian score would then be estimated to something very close to zero, but not exactly zero. By Theorem 3.7, we could also use the marginal entropy score being zero or negative as a stopping criterion, though this suffers from the same issues as the Gaussian score. Thus, using the marginal score as a stopping criterion would likely result in the algorithm adding more and more edges until the graph is either cyclic or fully connected. Thus, we need to modify the algorithm to account for this phenomenon. We do this here by implementing a significance test. That is, we set a confidence level and perform a test of whether the marginal score – either Gaussian or entropy – is significantly different from zero. The test should be one-sided, as we do not wish to add an edge that significantly lowers the marginal score. How this test is constructed depends on the choice of estimators and their distribution.

Choice of estimators

Before we can implement the Greedy entropy-search, we need to choose a regression technique and estimators of entropy and variance. We do not discuss the choice of variance estimators since most regression techniques have built-in estimation of residual variance. When we employ sample splitting, we simply use the standard sum of residual squares estimator. Furthermore, we only consider estimators which are readily available in R. We have two criteria when choosing a regression. First, it should be non-parametric as we do not wish to specify *a priori* the function class $\mathcal{F}$. Popular choices of non-parametric regression techniques in R include the mgcv package [Wood, 2003], which implements generalized additive models, local polynomial regression which comes with base R in form of the function loess, and finally the mboost package [Hofner et al., 2015]. Secondly, the regression technique should run reasonably fast. We require this because the number of regressions to be performed in a Greedy entropy-search grows quadratically in the number of observed variables[5]. In our experience, the gam function from mgcv for fitting generalized additive models ran efficiently when the covariate terms were smoothed using cubic splines, but slightly slower when using the standard thin-plate splines. The runtime of the function gamboost from the mboost package was comparable to using gam with thin-plate splines. Using loess was inefficient for large datasets. Motivated by achieving efficient runtime, we opted to use gam with cubic spline smoothing for regressions.

The problem of estimating differential entropy is difficult and still being studied, though it is possible to construct consistent entropy estimators. Most of these are either plug-in estimators that require a density estimate, or they are based on k-nearest-neighbor (KNN) techniques. Beirlant et al.

[4] Recall that this stated that $\Delta\ell^g = 0$ when the candidate nodes can be d-separated by their parent sets.

[5] For p variables, $p \cdot (p-1)$ regressions are performed in the initial step – one regression per possible edge. At every step thereafter, the previous models can be recycled and only the node with new parents need to be re-estimated.

[1997] provide an overview of entropy estimators. An altogether different issue in entropy estimation is that very few entropy estimators are implemented in R. A KNN-based estimator is available in the package `FNN` [Beygelzimer et al., 2019] and a modified version based on a recent paper by Berrett et al. [2019] is available in the package `IndepTest` [Berrett et al., 2018]. This estimator is itself a modified version of the Kozachenko-Leonenko estimator as proposed by Kozachenko and Leonenko [1987].[6] We attempted to use the estimator from Berrett et al. [2019], which we call the KL-estimator, and while it performed well on some large, simulated datasets, we found it to behave unexpectedly in settings with lower sample sizes. Furthermore, the KL-estimator only works when there are no duplicates present in the data. If there is even a single duplicate in a single variable it estimates the entropy to minus infinity – this is most likely due to the estimator taking the logarithm of the KNN-distances in. While not an issue for simulated data, it is a hindrance in real-life datasets, where duplicates often occur in data – e.g. through rounding during data collection. Instead, we construct a resubstitution estimator of entropy, which we define as [Beirlant et al., 1997, Equation (8)]

$$\hat{\mathbb{H}}_n(\hat{\boldsymbol{N}}_\nu) := -\frac{1}{n}\sum_{i=1}^{n} \log \hat{p}_{\hat{N}_\nu}(\hat{N}_\nu^i), \tag{4.1}$$

where $\hat{p}$ is a density estimate of the random variable $\hat{N}_\nu$. To estimate the density, we used the `logspline` package in R [Kooperberg, 2019]. In our experience, the resubstitution estimator performed markedly better than the KL-estimator in small samples. In larger samples, the two estimators had similar performance – see Figure 4.1 for a comparison. As briefly mentioned in the beginning of Section 4.2, we wish to construct a test to determine when edges significantly improve the marginal score. In order to be able to construct this, we need to know something about the distribution of either the variance estimator or the entropy estimator. As a proxy of $\mathbb{V}\hat{\mathbb{H}}_n$, we use the sample variance

$$\hat{\mathbb{V}}_n\hat{\mathbb{H}}_n(\hat{\boldsymbol{N}}_\nu) := \frac{1}{n}\sum_{i=1}^{n}\Big(-\log \hat{p}_{\hat{N}_\nu}(\hat{N}_\nu^i) - \hat{\mathbb{H}}_n(\hat{\boldsymbol{N}}_\nu)\Big)^2. \tag{4.2}$$

If we estimate $\hat{p}$ on half the sample and use the remaining half to estimate the residuals, the expression in equation (4.2) is the asymptotic variance of $\hat{\mathbb{H}}_n$, by the Central Limit Theorem (CLT). This holds because the act of sample splitting ensures that $\hat{\mathbb{H}}_n$ is simply the sample mean of an i.i.d. sequence. If we do not split the sample in this manner, we do not have any guarantees that $\hat{\mathbb{H}}_n$ is asymptotically normal. However, we can assess the distribution of $\hat{\mathbb{H}}_n$ through simulation. In Figure 4.2 we have performed 1000 simulations of sample size 1000 for four different distributions and each time we have calculated $\hat{\mathbb{H}}_v$ without employing sample splitting. From these we deem it likely that the estimator given by (4.1) is in fact asymptotically normal, with asymptotic variance given by (4.2). A version of Figure 4.2 with the sample size lowered to 250 is shown in Appendix A.1, which yields largely the same conclusion. Using $\hat{\mathbb{V}}_n\hat{\mathbb{H}}_n$ to estimate the asymptotic variance,

[6] The cited paper here is the original paper in Russian. To the knowledge of the author, no English translation of it exists.

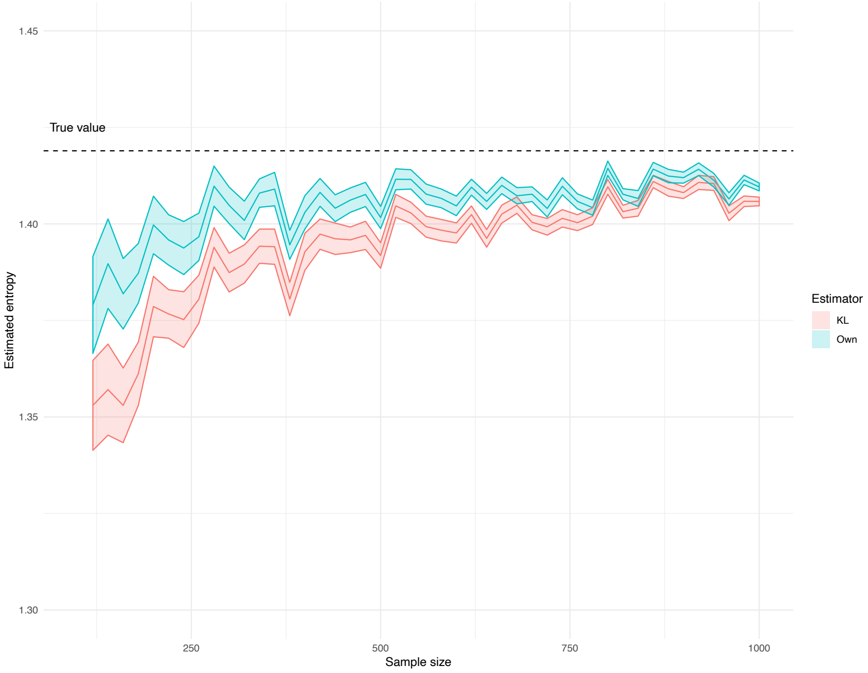

Figure 4.1: *Comparison of the KL-estimator of entropy (red) and the resubstitution estimator using log-spline density estimation (blue) as given in* (4.1). *Both estimators were used on the same sample of a* $\mathcal{N}(0,1)$*-distributed random variable* 50 *times and then averaged. This was done for samples of sizes* 100 *to* 1000 *in increments of twenty. The confidence bands are based on the bootstrapped standard deviation.*

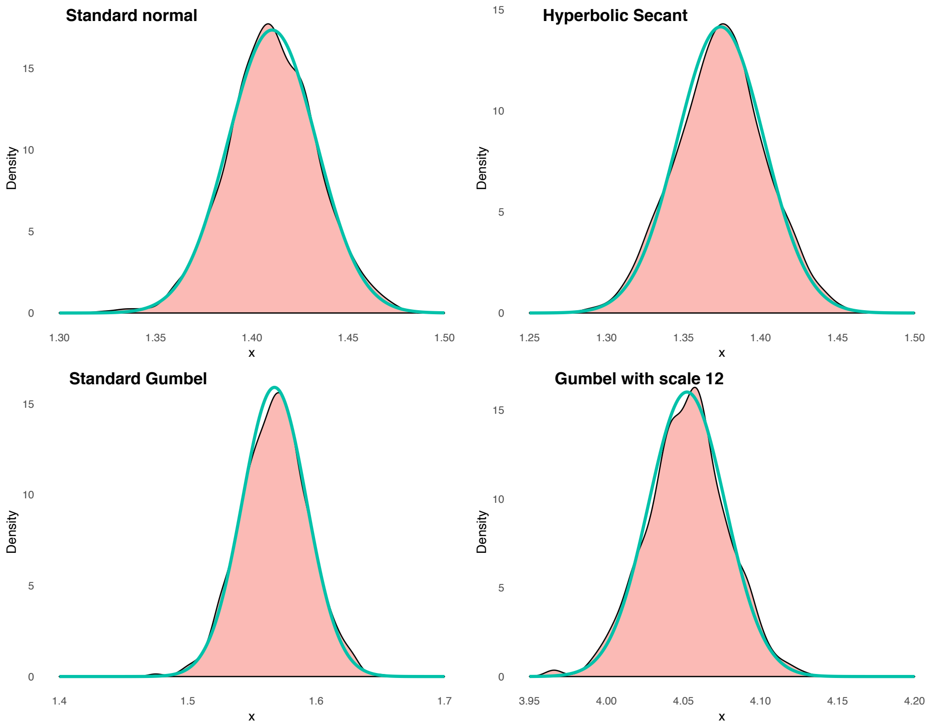

Figure 4.2: *Simulated estimates of* $\hat{\mathbb{H}}_n$ *for four different distributions, each with with sample size* 1000 *and* 1000 *repetitions. The superimposed lines are Guassian densities with standard deviation as estimated by* $\hat{\mathbb{V}}_n\hat{\mathbb{H}}_n$ *on a single sample. The filled areas mark the kernel density estimate of the* 1000 *realizations of* $\hat{\mathbb{H}}_n$.

we can implement a significance test of the marginal entropy score. For this, we employ a standard one-tailed z-test. In our first implementation, we stopped the Greedy entropy-search the first time it tried to add an edge that did not significantly increase the marginal entropy score. Pseudocode for this implementation can be found in Algorithm 2 on page 67.[7] Following this recipe led to very poor results when applied to larger[8] graphs. The Greedy entropy-search seemed to identify correct edges at first, but then exit prematurely. In many cases, the Gaussian score would identify an edge which was in fact not present in the true graph. The estimated marginal entropy score would then not be significant, and the search would terminate. To circumvent this issue, we allowed the search to continue a set number of times after reaching an edge that did not significantly increase the entropy score, which improved the algorithm's performance markedly. Pseudocode-code for the final implementation can be found in Algorithm 3 on page 68.

In summary, we are going to perform regressions using the `gam` function provided in the `mgcv` package [Wood, 2003]. For variance estimation, we use `gam`'s built-in estimator and for entropy estimation we use the package `logspline` [Kooperberg, 2019] to provide density estimates which are then plugged into the resubstitution estimator given in equation (4.1). We do not claim that these are the best possible estimators, but feel that they at least provide a good starting point.

4.3 SIMULATION STUDIES

All simulation studies in this section were carried out using R version 3.5.2, running on a Linux® device with 24 virtual CPU cores and 295 gigabytes of RAM.

We have written a package in R which implements the Greedy entropy-search algorithm as described in Algorithm 3. The algorithm can be run both with and without sample splitting. For convenience, all packages used either in the implementation or in the simulations in this section are collected in Table 4.1 along with their references. Throughout this section we conduct simulation experiments meant to give an empirical assessment of the performance of the Greedy entropy-search in finite sample-settings.

Baseline performance and sensitivity towards linearity

To provide a baseline for the performance of our algorithm, we begin by repeating a simulation experiment carried out first by Hoyer et al. [2009b] and later by Nowzohour and Bühlmann [2016]. In this experiment, we consider a bivariate ANM with graph $\mathcal{G} := X_\alpha \to X_\beta$ and

$$\begin{aligned} X_\alpha &:= \operatorname{sign}(N_\alpha) \cdot |N_\alpha|^q, && N_\alpha \sim \mathcal{N}(0,1), && q \geq 0.5 \\ X_\beta &:= X_\alpha + b \cdot X_\beta^3 + \operatorname{sign}(N_\beta) \cdot |N_\beta|^q, && N_\beta \sim \mathcal{N}(0,1), && \beta \in \mathbb{R}. \end{aligned}$$

[7] To get a version that employs sample splitting, simply insert an extra dataset and keep track of what is estimated where. We do not typeset such a version because it is considerably longer and is no more instructive.

[8] Say, more than 15 nodes

Algorithm 2: The Greedy Entropy-Search without sample-splitting and early exit.

```
Input: Graph $\mathcal{G}$ with $d$ nodes, i.i.d data, a regression $\hat{\phi}_n$, estimator $\hat{\mathbb{H}}_n$
Initialization
Make a $d \times d$ matrix, $M$, with $\operatorname{diag}(M) = 0$
$k := 0$
$\mathcal{G}^{s=k} := \mathcal{G}$
Step 1:
for $\nu, \gamma \in V(\mathcal{G})$, $\nu \neq \gamma$ do
    Obtain estimates $\hat{\phi}_{X_\nu | P_\nu^{\mathcal{G}^{s=k}}}$ and $\hat{\phi}_{X_\nu | P_\nu^{\mathcal{G}^{s=k}_{\nu\to\gamma}}}$
    Estimate residuals $\hat{\boldsymbol{N}}_\nu^{\mathcal{G}^{s=k}}$ and $\hat{\boldsymbol{N}}_\nu^{\mathcal{G}^{s=k}_{\nu\to\gamma}}$
    Set $M_{\nu,\gamma} := \widehat{\Delta\ell_n^{\mathrm{g}}}\left(\mathcal{G}^{s=k}, \nu \to \gamma\right)$
Step 2:
Set $(\alpha \to \beta) := \arg\max_{\nu \neq \gamma} M_{\nu,\gamma}$
if $M_{\nu,\gamma} \leq 0$ then
    return $\mathcal{G}^{s=k}$
Estimate $\widehat{\Delta\ell_n}\left(\mathcal{G}^{s=k}, \beta \to \alpha\right)$ and $\widehat{\Delta\ell_n}\left(\mathcal{G}^{s=k}, \alpha \to \beta\right)$
Set $(\tilde{\alpha} \to \tilde{\beta}) := \arg\max_{(\alpha\to\beta),(\beta\to\alpha)} \widehat{\Delta\ell_n}\left(\mathcal{G}, \cdot \to \cdot\right)$
if $\widehat{\Delta\ell_n}\left(\mathcal{G}^{s=k}, \tilde{\alpha} \to \tilde{\beta}\right)$ is not significant. then
    return $\mathcal{G}^{s=k}$
else
    $k := k + 1$
    $\mathcal{G}^{s=k} := \mathcal{G}^{s=k-1}_{\tilde{\alpha}\to\tilde{\beta}}$
Set $M_{\tilde{\alpha},\tilde{\beta}} := M_{\tilde{\beta},\tilde{\alpha}} := 0$
Re-estimate $\widehat{\Delta\ell_n^{\mathrm{g}}}\left(\mathcal{G}^{s=k}, \nu \to \tilde{\beta}\right)$ for every $\nu$
Set $M_{\nu,\tilde{\beta}} := \widehat{\Delta\ell_n^{\mathrm{g}}}\left(\mathcal{G}^{s=k}, \nu \to \tilde{\beta}\right)$ for every $\nu$
Step 3:
Repeat step 2 until every entry in $M$ is weakly below zero or until the marginal score does
  not increase significantly
```

Algorithm 3: The Greedy Entropy-Search with modified exit.

```
Input: Graph $\mathcal{G}$ with $d$ nodes, i.i.d data, a regression $\hat{\phi}_n$, estimator $\hat{\mathbb{H}}_n$, integer $m_{max}$
Initialization
Make a $d \times d$ matrix, $M$, with $\operatorname{diag}(M) = 0$
$k := 0$
$\mathcal{G}^{s=k} := \mathcal{G}$
$m := 0$
Step 1:
for $\nu, \gamma \in V(\mathcal{G})$, $\nu \neq \gamma$ do
    Obtain estimates $\hat{\phi}_{X_\nu | P_\nu^{\mathcal{G}^{s=k}}}$ and $\hat{\phi}_{X_\nu | P_\nu^{\mathcal{G}^{s=k}_{\nu\to\gamma}}}$
    Estimate residuals $\hat{\boldsymbol{N}}_\nu^{\mathcal{G}^{s=k}}$ and $\hat{\boldsymbol{N}}_\nu^{\mathcal{G}^{s=k}_{\nu\to\gamma}}$
    Set $M_{\nu,\gamma} := \widehat{\Delta\ell_n^{\mathrm{g}}}\left(\mathcal{G}^{s=k}, \nu \to \gamma\right)$
Step 2:
Set $(\alpha \to \beta) := \arg\max_{\nu\neq\gamma} M_{\nu,\gamma}$
if $M_{\nu,\gamma} \leq 0$ then
    return $\mathcal{G}^{s=k}$
Estimate $\widehat{\Delta\ell_n}\left(\mathcal{G}^{s=k}, \beta \to \alpha\right)$ and $\widehat{\Delta\ell_n}\left(\mathcal{G}^{s=k}, \alpha \to \beta\right)$
Set $(\tilde{\alpha} \to \tilde{\beta}) := \arg\max_{(\alpha\to\beta),(\beta\to\alpha)} \widehat{\Delta\ell_n}\left(\mathcal{G}, \cdot \to \cdot\right)$
if $\widehat{\Delta\ell_n}\left(\mathcal{G}^{s=k}, \tilde{\alpha} \to \tilde{\beta}\right)$ is not significant. then
    if $m < m_{max}$ then
        $M_{\tilde{\alpha},\tilde{\beta}} := M_{\tilde{\beta},\tilde{\alpha}} := 0$
        $m := m + 1$
        Jump to start of step 2
    else
        return $\mathcal{G}^{s=k}$
else
    $k := k + 1$
    $\mathcal{G}^{s=k} := \mathcal{G}^{s=k-1}_{\tilde{\alpha}\to\tilde{\beta}}$
Set $M_{\tilde{\alpha},\tilde{\beta}} := M_{\tilde{\beta},\tilde{\alpha}} := 0$
Re-estimate $\widehat{\Delta\ell_n^{\mathrm{g}}}\left(\mathcal{G}^{s=k}, \nu \to \tilde{\beta}\right)$ for every $\nu$
Set $M_{\nu,\tilde{\beta}} := \widehat{\Delta\ell_n^{\mathrm{g}}}\left(\mathcal{G}^{s=k}, \nu \to \tilde{\beta}\right)$ for every $\nu$
Step 3:
Repeat step 2 until every entry in $M$ is weakly below zero or until $m = m_{max}$
```

Package	Usage	Reference
`mgcv`	Non-parametric regression.	Wood 2003
`mboost`	Non-parametric regression.	Hofner et al. 2015
`logspline`	Density estimation	Kooperberg 2019
`SID`	Calculates Structural Intervention Distance between DAGs	Peters 2015
`gRbase`	For construction and manipulation of graph objects	Dethlefsen and Højsgaard 2005
`pcalg`	Simulation of random graphs, learning algorithms	Hauser and Bühlmann 2012
`CAM`	Implements a learning algorithm for ANMs	Peters and Ernest 2015
`CompareCausalNetworks`	Contains a unified method for calling causal discovery algorithms.	Heinze-Deml et al. 2018

Table 4.1: *Packages used to implement the Greedy entropy-search in R*

When $q = 1$ and $b = 0$, the model above reduces to $X_\beta = X_\alpha + N_\beta$ and $X_\alpha = N_\alpha$. For this model, Theorem 3.11 does not hold, and we will not be able to correctly direct the edge between α and β. This is because both sets of residuals are normally distributed and so the Gaussian score becomes proportional to the entropy score[9]. Furthermore, the two variables X_α and X_β entail conditional distributions that are equal up to a translation and so the residual variances are identical when $b = 0$ and $q = 1$. For any other combination of b and q, however, we can recover the graph of $\mathscr{C}$ – in theory, at least. Whenever q is close to one and b is close to zero, we find ourselves in a region of 'almost'-identifiability, so to speak. If we are given a finite sample of data from the above model, say with $b = 10^{-100}$ and $q = 1$, it seems almost impossible that we should be able to discern this from the unidentifiable case of $b = 0$ and $q = 1$. This simulation experiment, then, also serves a secondary purpose, namely of investigating what happens when the ANM is *almost* unidentifiable.

Mimicking Hoyer et al. [2009b] and Nowzohour and Bühlmann [2016], we simulated 300 observations of (X_α, X_β) and ran two Greedy entropy-searches to recover the graph; one with sample splitting and one without. We repeated this experiment 100 times and recorded the number of times in which we added the correct edge to estimate the probability of making the correct decision. We then did this for different values of $b \in [-1, 1]$ and $q \in [0.5, 2]$. The results of the simulation can be seen in Figure 4.3. The results of the simulation were largely comparable to the results of Nowzohour and Bühlmann [2016], when we did not split the sample. When we employed sample splitting, the results were considerably worse for varying values of b, but only slightly worse when varying q. At lower sample sizes, the difference between sample splitting and not sample splitting becomes more

[9] Recall Proposition 3.3.

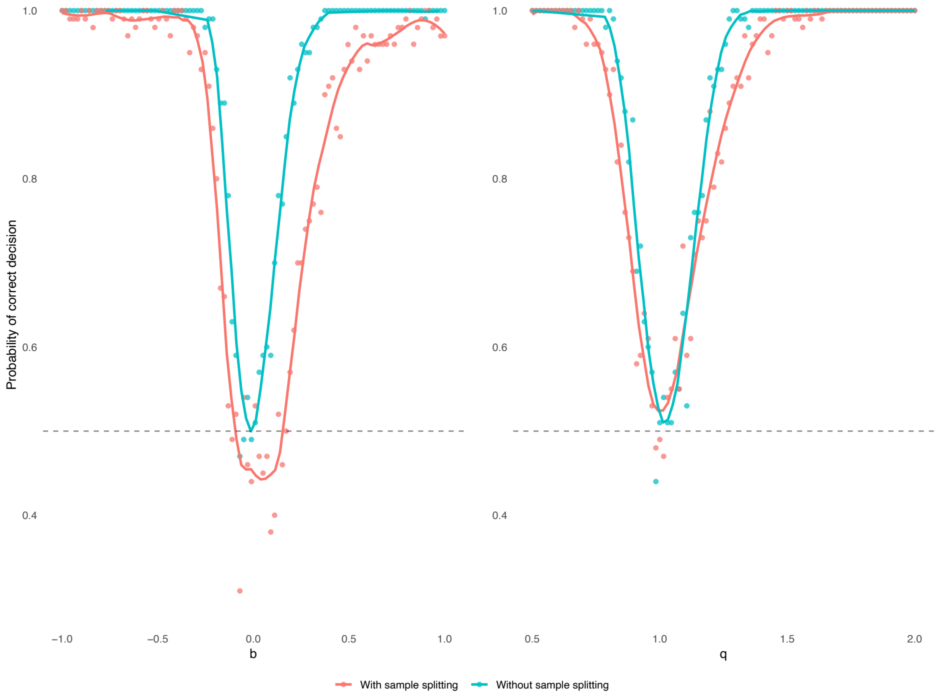

Figure 4.3: *Estimated probabilities of correctly identifying $\mathcal{G}$ for different values of b (left) and q (right), respectively. Light blue points are results where the sample was split randomly in half during estimation. Red points are results from using the full sample. Superimposed lines are Loess smoothers included to indicate the general trends.*

pronounced. See Appendix A.2 for a version of Figure 4.3 in which the sample size was lowered to 50. We conclude two things from this simulation experiment. Firstly, that the Greedy entropy-search appears to be quite sensitive to functions that are almost linear when the noise is normal. In contrast, when the function is linear, it seems that the algorithm is better capable of handling noise that is close to being normal. These findings are similar to those of Hoyer et al. [2009b] and Nowzohour and Bühlmann [2016]. Our second conclusion is that it does not seem beneficial to employ splitting – in fact, it seems to worsen the performance of the Greedy entropy-search. From this point on, we do not employ sample splitting in any of our estimations.

Having established that we can successfully recover a bivariate, identifiable graph we are now going to investigate how well the algorithm performs in larger graphs with randomly generated edge functions. Following Nowzohour and Bühlmann [2016] once again, we generate random functions by simulating a random walk on an interval which we then smooth with a cubic spline using the `smooth.spline` function available in Base R. We adjust the linearity of the random functions by adjusting the smoothing parameter of `smooth.spline`. See Figure 4.4 for an example of how these random functions look. In all simulations from hereon out, we used this method to generate random functions unless otherwise stated. Furthermore, we always use a significance level of $\alpha = 0.001$ unless otherwise specified. We start by considering a fixed graph with three nodes, random edge functions and standard Gaussian noise. That is, we do not change which edges are present in the graph, but we do change the corresponding structural equation. We consider the graph with edges $\alpha \to \beta \leftarrow \gamma$. As we did in the bivariate case, we repeat the experiment for different values of smoothing parameters in an effort to examine the model's sensitivity to functions that are close to linear. In light of Corollary 3.29, we also ran the same experiment but with the noise variables following a Hyperbolic Secant distribution instead, to compare the difference. For every value of the smoothing parameter, s, we sampled 300 observations from each ANM and ran a Greedy entropy-search. We repeated this process 100 times and recorded the number of times the correct graph was recovered, which we then scaled down to a probability. The results are summarized in Figure 4.5. Even though the model is identifiable when $s = 0$ and the noise distribution is Hyperbolic Secant, the true graph is recovered in less than half of the simulations. While it does recover the graph more often in the Hyperbolic Secant case than it does in the Gaussian, we had expected a higher success rate. However, since the normal distribution and the Hyperbolic Secant distribution are very similar in shape, this is likely a case of the noise not being sufficiently 'non-Gaussian', as we also observed in the bivariate case.

Estimation of larger graphs

We now apply the Greedy entropy-search algorithm to larger graphs. For a chosen number of nodes, say p, we draw graphs randomly from the class of unrelated graphs with an expected number of edges equal to p. That is, whenever we simulate a graph of size p, each edge has probability $2/(p-1)$ of appearing. For the random selection of graphs, we used the function `randomDAG` from the package `pcalg`. We slightly modified the `randomDAG`-function to reflect that we have only proven the Greedy entropy-search for unrelated graphs. However, we were not able to find a computationally fast method of checking whether a graph is unrelated. Instead, we generated a slightly larger class of graphs, in which every cycle is allowed to have only two colliders. This allows for graphs with a diamond-like structure, for instance[10]. Such graphs can equivalently be characterized by the ancestral graph of any node having exactly as many connected components[11] as it has parents. Since connected components of a graph can be found in linear time [Hopcroft and Tarjan, 1973], it requires little computing power to check if a graph has at most two colliders in every cycle. We do not have a similar method available for unrelated graphs. As a consequence, we are on occasion going to apply the Greedy entropy-search on a type of graph, for which we have no guarantee that it works.

As we begin to estimate larger and larger graphs, the probability of recovering exactly the true graph becomes lower and lower. Instead of measuring the performance of the Greedy entropy-search by the probability of correctly identifying the true graph, we use instead the Structural Hamming Distance (SHD), which counts the number of differing edges between two graphs. The SHD provides an intuitive measure of the differences between two graphs, but Peters and Bühlmann [2013a] argue

[10] Recall Figure 3.1 for an example.

[11] A connected component is a collection of nodes that are all reachable, starting from any node in the component.

that it is not well suited for measuring a models capacity for causal inference. As an alternative, Peters and Bühlmann [2013a] define the Structural Intervention Distance (SID). Instead of counting the number of incorrect edges in an estimated graph, it counts the number of incorrectly inferred interventional distributions in the estimated graph [Peters and Bühlmann, 2013a, page 5]. In this sense, the SID serves as a measure of capacity for causal inference in an estimated graph compared to a true graph. It is weakly positive, with equality only when the estimated graph equals the true graph. For both the SHD and SID, a lower score is better.

We first compared our Greedy entropy-search with the Causal Additive Models (CAM) algorithm introduced in Bühlmann et al. [2014]. This algorithm is implemented in the `R` package `CAM`. It is an ambitious starting point to compare our method to the CAM algorithm, as the CAM algorithm has been shown by simulation to perform considerably better than its competitors when applied to ANMs [Bühlmann et al., 2014]. However, due to time constraints we were not able to compare our method to the CAM algorithm in as large a scale as we would have liked. We simulated 50 samples of size 300 from ANMs with respectively 10, 20, 30 and 50 nodes. In all simulations, we added standard normal noise. We then ran the CAM algorithm and the Greedy entropy-search on each simulated dataset, each time recording the SID and the SHD. In the setup of the CAM algorithm, we used the example provided by the package authors, Peters and Ernest, in the package vignette. However, we also chose to use the Preliminary Neighborhood Step (PNS), as described in Bühlmann et al. [2014, Section 5.1], in order to keep computation times at a minimum. The results of the simulation are shown in Figure 4.6. In all cases, the trends were the same; the CAM algorithm performed better when measured by the SID, while the Greedy entropy-search achieved the lowest SHD. A possible explanation of this phenomenon could be, that the Greedy entropy-search is good at identifying edges, but occasionally has trouble directing them. A single misdirected edge can potentially result in a large SID, but will have a low SHD. In fact, when the SHD is equal to one, the SID can be as large as $2 \cdot (p-1)$ [Peters and Bühlmann, 2013a, Proposition 8], where p denotes the size of the graph. This explanation is also consistent with our claim that it is difficult to estimate the differential entropy, since we use the entropy to direct edges. In contrast, the CAM algorithm appears to make more mistakes when adding edges, but does so in manner that does not affect the SID. Based on this, albeit small, simulation exercise, it appears that the Greedy entropy-search is capable of recovering a graph that is close to the true graph in term of SHD, but that the CAM algorithm is the better choice for drawing causal inference.

An important feature of the Greedy entropy-search is that it is computationally efficient. In the same simulation as above, we also recorded the running time of each algorithm – that is, the time it took for the algorithm to run to completion. The results can be found in Figure 4.7. Each call of each algorithm was parallelized onto 24 virtual CPU cores[12]. The simulations indicate that the Greedy entropy-search scales considerably better with the size of the ANM than the CAM algorithm

[12] Although this procedure may, in fact, not provide any benefit in small graphs, when accounting for the time it takes to set up each task. However, already at $p = 10$, we perform 90 regressions in the first step.

does, although the speed-up comes at a cost; we estimate the graph quicker when using the Greedy entropy-search, but the output graph has lower capacity for causal inference. In defense of the CAM algorithm, we did not attempt to tune the algorithm parameters to achieve a lower running time. It is possible that another configuration exists which provides a lower running time with similar results.

Next we compared the Greedy entropy-search to other existing methods. Due to time constraints we chose only a few different algorithms based on their running times. We compared the Greedy entropy-search with the 'Linear Non-Gaussian Acyclic Models' (LiNGAM) algorithm [Shimizu et al., 2006] as well as the PC algorithm (Spirtes et al. 2000 as cited in Bühlmann et al. [2014]) in a simulation identical to the one described above – except that the number of repetitions is increased to 100. We used the package `CompareCausalNetworks` to run the algorithms. The results of the simulation are shown in Figure 4.8. We see in Figure 4.8 a considerable difference between the Greedy entropy-search and its competitors, especially in larger graphs. It appears then, that the Greedy entropy-search is superior to both the LiNGAM and PC algorithms for ANMs of this type, both in terms of SHD and SID. This finding is consistent with what we saw in Figure 4.6, where we found the Greedy entropy-search to be comparable to the CAM algorithm for which Bühlmann et al. [2014, Figure 5] reach a similar conclusion. Note that the PC algorithm can return bi-directed graphs. The `SID` package then computes upper and lower SID bounds by calculating respectively the highest and lowest possible SID that can be obtained within the Markov equivalence class of the graph.

We then tried to run the Greedy entropy-search on an even larger scale. This time, we simulated random ANMs with 100 nodes and 1000 observations. Unfortunately time did not permit us to include the CAM algorithm in this experiment; based on two practice runs, the CAM algorithm took just over 20 minutes to run to completion. Instead, we included only the PC algorithm and the LiNGAM algorithm for comparison. The results are shown in Figure 4.9. From Figure 4.9 it appears that the Greedy entropy-search performs well even in large graphs. Although the SID does become very high on some occasions, this should be held in contrast to the fact the SID has an upper bound that increases quadratically in the size of the graph. That is, when the graphs have 100 nodes, the SID can be as high as $100 \cdot (100 - 1)$ [Peters and Bühlmann, 2013a, Section 2.3]. The Greedy entropy-search had a median SID of six and in the worst observed case, the Greedy entropy-search achieved an SID of 89. In fact, only the single worst outcome of the Greedy entropy-search, as measured by the SID, reached a higher SID than the best possible outcome of the remaining algorithms combined. That is, the second highest SID reached by the Greedy entropy-search was equal to 50; the lowest SID reached by the PC algorithm had a lower bound of 58. However, in running time, the Greedy entropy-search took an average of 5.9 minutes to run. In contrast, the PC algorithm averaged a running time of just below two seconds and the LiNGAM algorithm nine seconds.

In summary, we find in our simulations that the Greedy entropy-search outperforms a small selection of competing algorithms when applied to a class of non-linear ANMs with Gaussian noise. Furthermore, we find that it performs at a level comparable to that of the CAM algorithm in terms of SHD, although worse when measured by the SID.

Causal discovery of linear, non-Gaussian Additive Noise Models

We argued in Chapter 3 that the Greedy entropy-search should in theory also work on a selection of linear, non-Gaussian ANMs. As a particular example, we should be able to recover the graph of an ANM with linear assignments and noise variables that follow a Hyperbolic Secant distribution. We also briefly touched upon the subject in the simulation shown in Figure 4.5. We now extend the simulation experiment to large, random graphs in which all assignments are linear and all noise variables follow a Hyperbolic Secant distribution. By Corollary 3.29, such a model is indeed recoverable by the Greedy entropy-search. To reflect the change of model class, we changed the regression function in our implementation to a linear regression. For comparison, we used the LiNGAM and PC algorithms, as well as the CAM algorithm set to use its built-in linear regression method. The change from non-parametric regression to linear regression speeds up the estimations procedures by a considerable amount, which allows for larger simulations. We simulated 1000 observations from linear ANMs with random graphs of sizes 10, 20, 30 and 50 respectively. Each simulation was repeated 100 times, and the results are shown in Figure 4.10. In this setting, the Greedy entropy-search actually out-performed the CAM algorithm in all cases, both in terms of SID and SHD. However, in this setting, the Greedy entropy-search was outperformed on SID by the LiNGAM method, although comparable in SHD. It seems, then, that the Greedy entropy-search is capable of estimating a graph that is close to the true graph in terms of SHD, but that it is not well suited for causal inference in a linear ANM with Hyperbolic Secant noise.

Causal discovery of general non-Gaussian Additive Noise Models

We have seen throughout this section, that the Greedy entropy-search works well in a non-linear, Gaussian setting and decently in a linear, non-Gaussian setting. However, in both scenarios it was outperformed by either the CAM algorithm or the LiNGAM algorithm. Interestingly, neither of these two algorithms worked particularly well in both settings. That is, the CAM algorithm excelled in a non-linear, Gaussian setting, but not in a linear, non-Gaussian setting and vice versa for the LiNGAM algorithm. As our final simulation study, we consider an ANM with noise variables following a Hyperbolic Secant distribution. The twist is, that we do not fix a smoothing parameter for the function class. Instead, we draw a smoothing parameter uniformly from $[0, 0.5]$ and then randomly sample an ANM. In this setting, the algorithms will encounter both non-linear and linear[13] functions. Due to time constraints, we only simulations with 10 nodes and 30 nodes. Each time, we drew a random sample of size 1000 from a random graph. We repeated each of these experiments

[13] Strictly speaking, they are not completely linear, but almost.

100 times. In order to reduce the running time, we disabled the so-called pruning step of the CAM algorithm. Bühlmann et al. [2014] show that the pruning step only has a noticeable influence on the SHD, but not the SID. Therefore, we only record the SID in this experiment. The results are summarized in Figure 4.11. This time, the Greedy entropy-search outperforms both the CAM algorithm and the LiNGAM algorithm. Although the Greedy entropy-search still appears to be somewhat volatile, in the sense that it on occasion achieves a very high SID, it generally scores well in this scenario.

4.4 APPLYING THE GREEDY ENTROPY-SEARCH TO REAL DATA

Mooij et al. 2016 have collected a database of cause-effect pairs (41 of which are taken from the UCI Machine Learning Repository [Dua and Karra Taniskidou, 2017]), which are real-life datasets that each comprise a *known* causal effect. As an example, pair number fifteen consists of observations of fuel consumption and observations of vehicle weight. In this instance, we know that it is the vehicles weight that influences its fuel consumption – not the other way around. A total of 108 datasets are available at the time of writing – all of which can be found at `http://webdav.tuebingen.mpg.de/cause-effect/`. Of these 108 datasets, we restricted attention to those that are bivariate. Furthermore, three datasets were removed from consideration because these did not contain enough unique values for `logspline`-procedure to run. Lastly, in one dataset we log-transformed one of the variables in order to obtain convergence of our density estimate. In total, this left 96 cause-effect pairs to be analyzed. As in the previous sections, we used the `mgcv` package in R for regression and a resubstitution estimator of entropy based on density estimates from the `logspline` package. We did not utilize any variance estimator, as there was only one candidate edge in each dataset. We correctly identified the causal effect in 58 (60.4%) of 96 cases. Using the weights supplied by Mooij et al. 2016 to calculate a weighted mean, this corresponded to a weighted accuracy of 65.2%. If we instead used the `gamboost` function from the `mboost` package, we correctly identified 64 (66.7%) of 96 cases, amounting to a weighted accuracy of 68.9%. In the implementation, we forced the Greedy entropy-search to add an edge, even if it could not find one that increased the score significantly. It is only possible to this because we know *a priori* that a causal relationsship exists. Both when using `gam` and when using `gamboost`, we forced a decision in two cases that were not significant at a 0.1% percent significance level. If we consider these two cases to be wrong decisions, the weighted accuracy drops to 64.6% and 68.1% for `gam` and `gamboost` respectively. In comparison, Nowzohour and Bühlmann [2016] report accuracies ranging from 58%to 75% for different algorithms, although it is not clear whether these are weighted or not. As a disclaimer, it should be noted that these results are not directly comparable, since more datasets have been added to the database in the time since Nowzohour and Bühlmann [2016] was published.

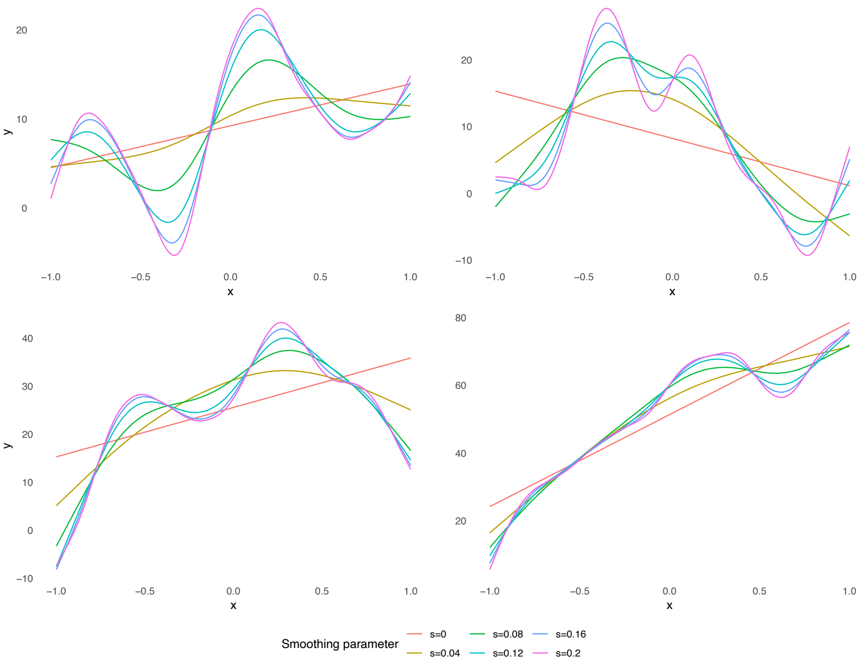

Figure 4.4: *Examples of the generated random functions for differing values of the smoothing parameter. Each plot is based on a single realization of a random walk on* $(-1, 1)$.

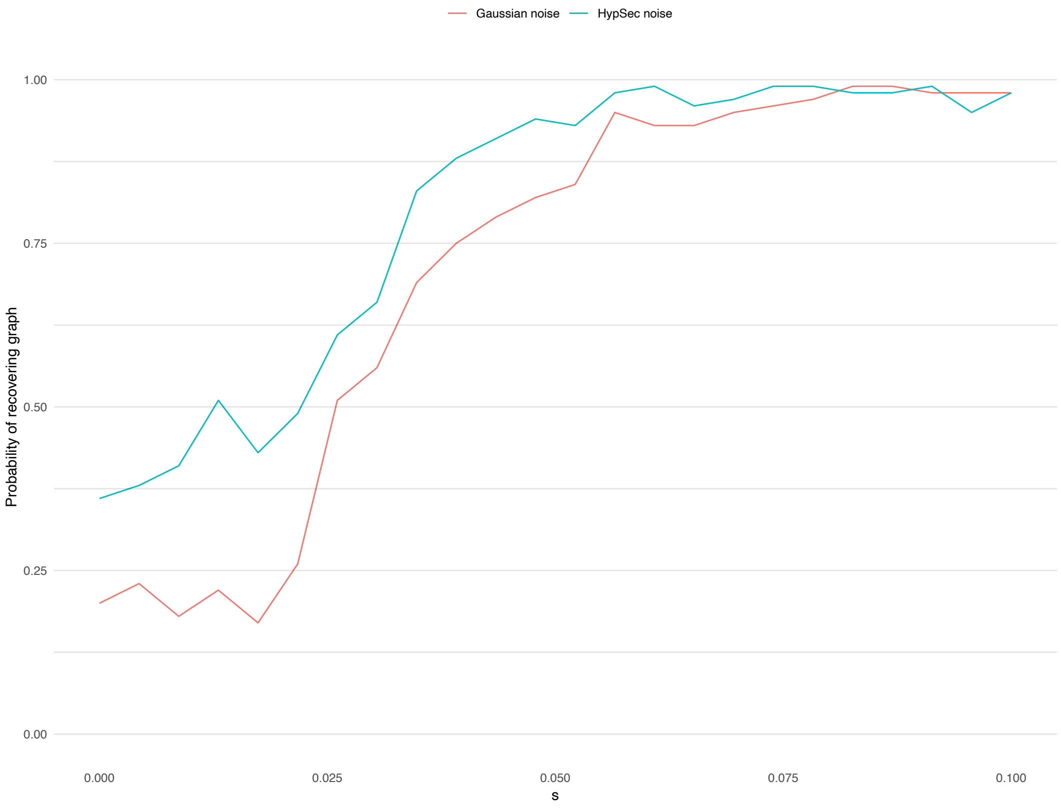

Figure 4.5: *Estimated probabilities of recovering the true graph for different values of a smoothing parameter.* $s = 0$ *corresponds to linear assignments. Red line: Gaussian noise variables. Blue line: Hyperbolic Secant distributed noise variables. Each point is an average of* 100 *repetitions. In each repetition,* 300 *observations were sampled from the ANM with graph* $X_\alpha \to X_\beta \leftarrow X_\gamma$

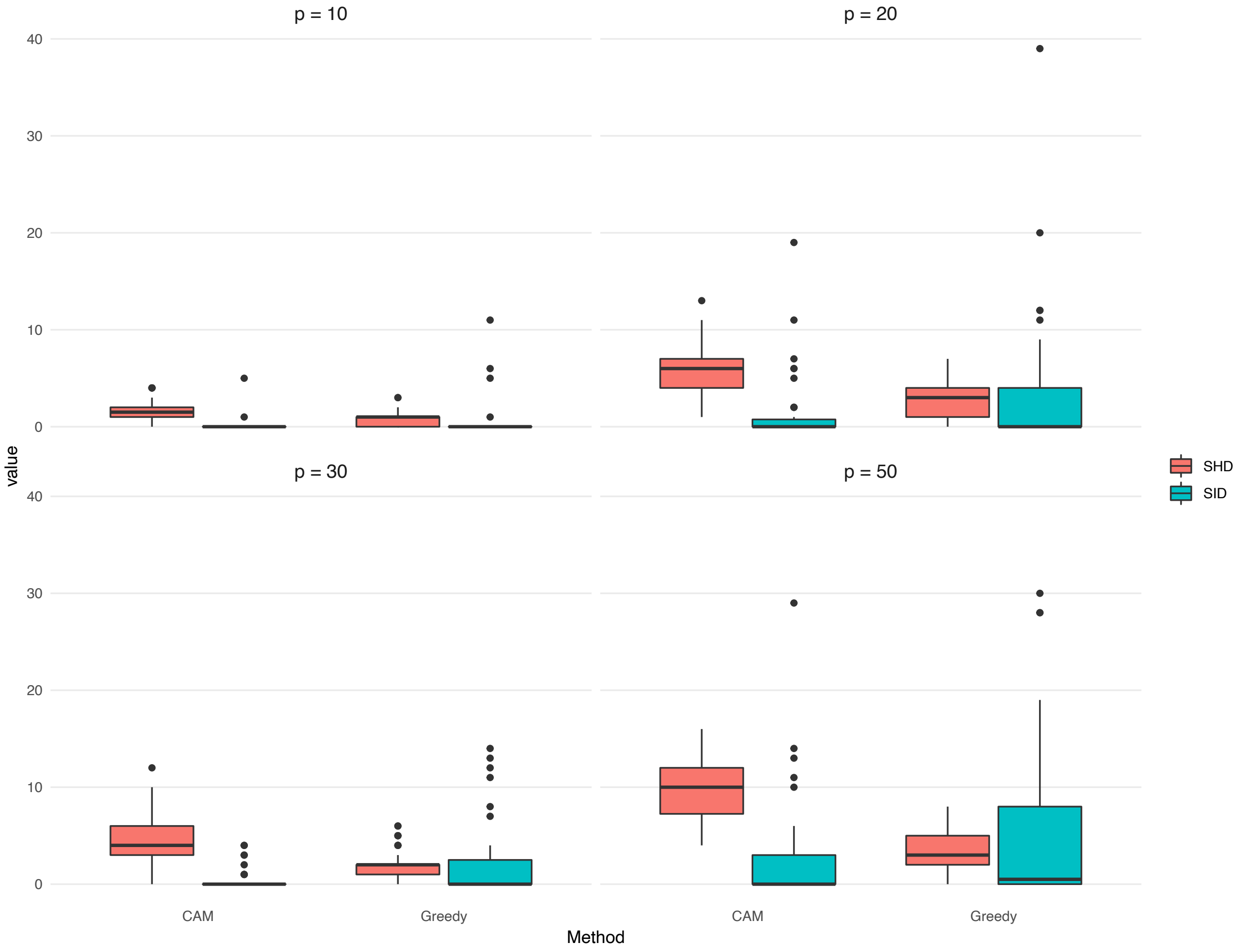

Figure 4.6: *A comparison of the CAM algorithm and the Greedy entropy-search based on* 50 *simulations, each with sample size* 300*. The simulations were run on ANMs with respectively* 10*,* 20*,* 30 *and* 50 *nodes. The number of nodes are denoted by* p*.*

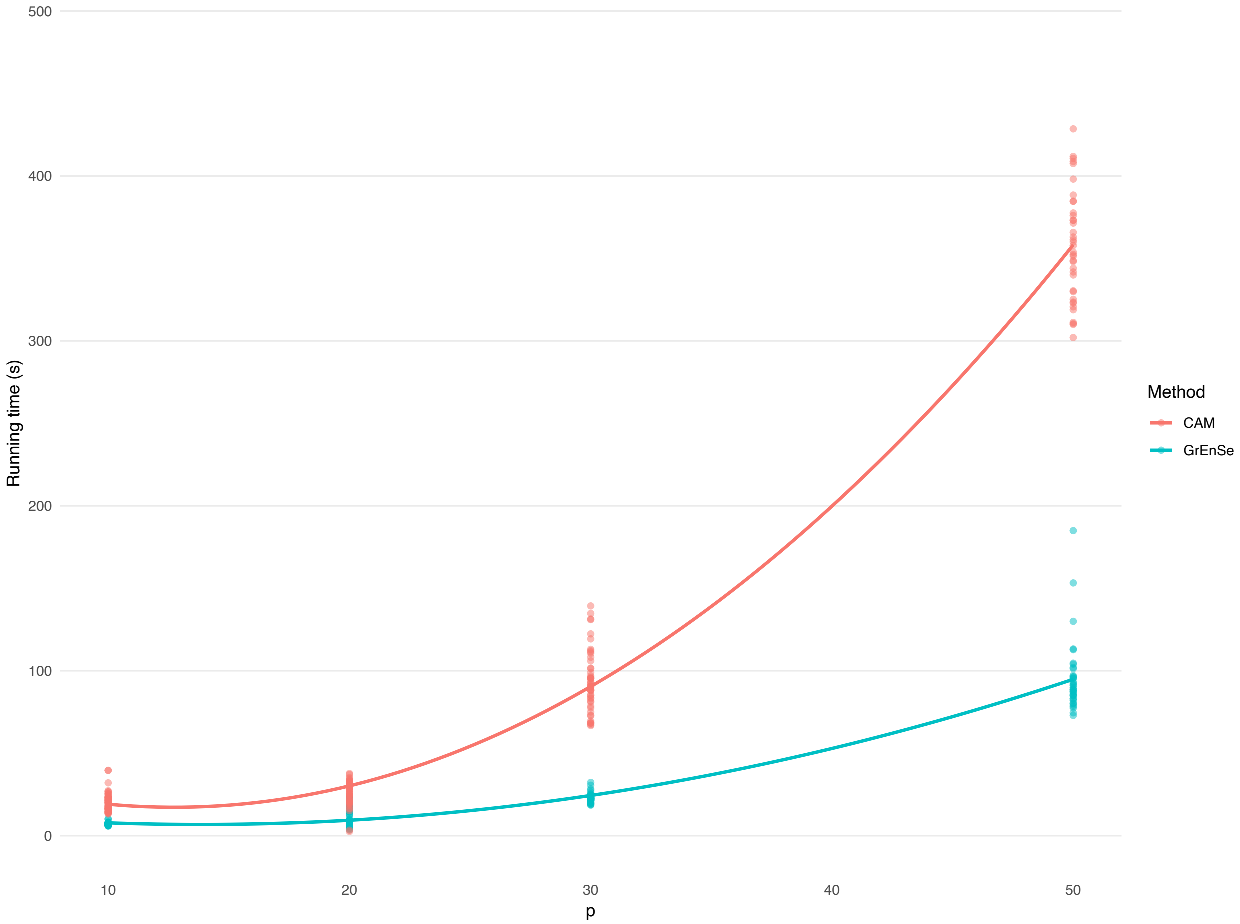

Figure 4.7: *A comparison of the running times of the CAM algorithm and the Greedy entropy-search. The results are based on the simulation described in Figure 4.6. Running time is displayed in seconds. The superimposed lines are Loess smoothers, included to display the general trends. The method 'GrEnSe' is short for Greedy entropy-search. Two extreme points (*≥ 3000 *seconds) have been removed from the graph in order to better compare the two methods. Both extreme points were from the CAM algorithm.*

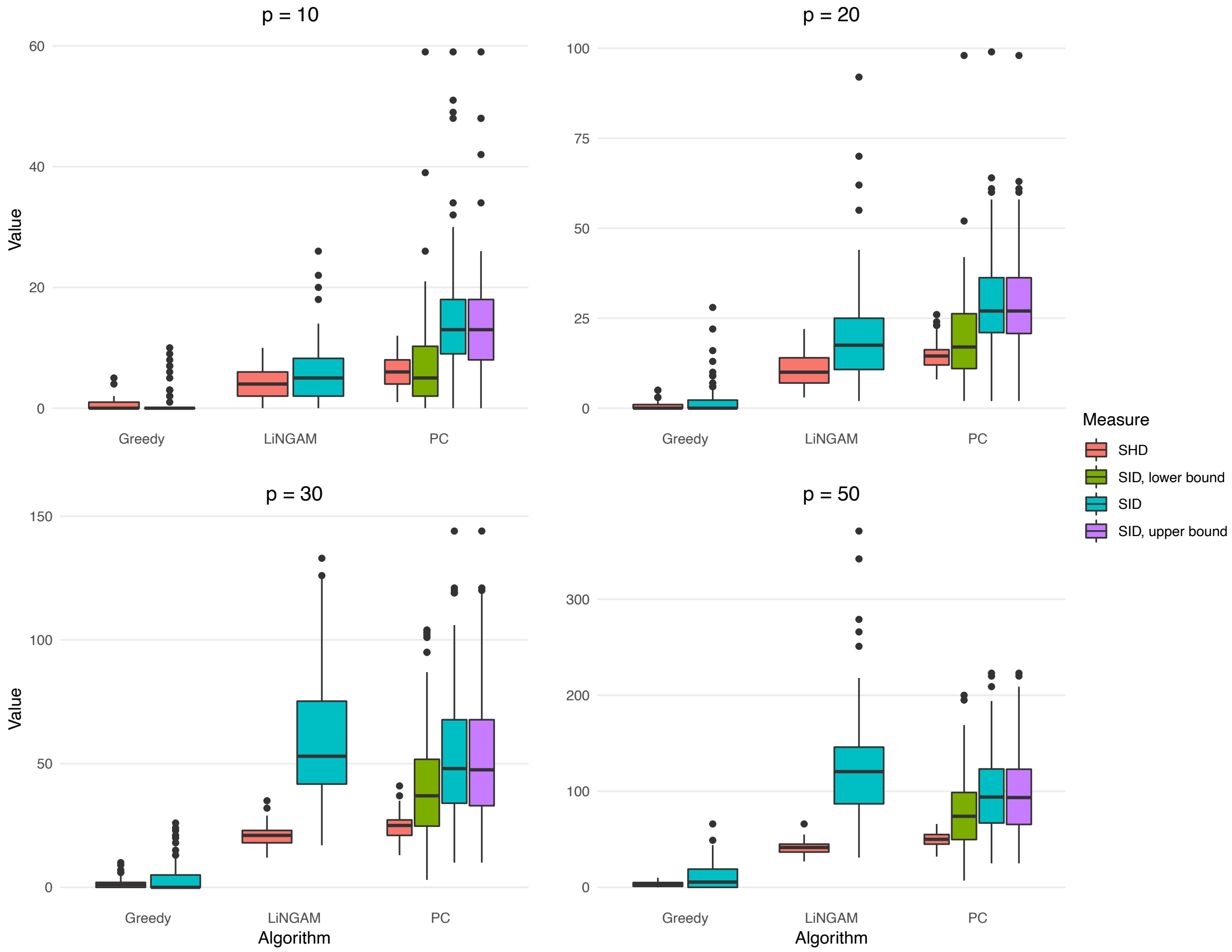

Figure 4.8: *A comparison of the Greedy entropy-search and existing methods based on* 100 *simulations, each with sample size* 300*. The simulations were run on ANMs with respectively* 10*,* 20*,* 30 *and* 50 *nodes. The number of nodes are denoted by p. Note that the scales are not fixed, but vary by subfigure.*

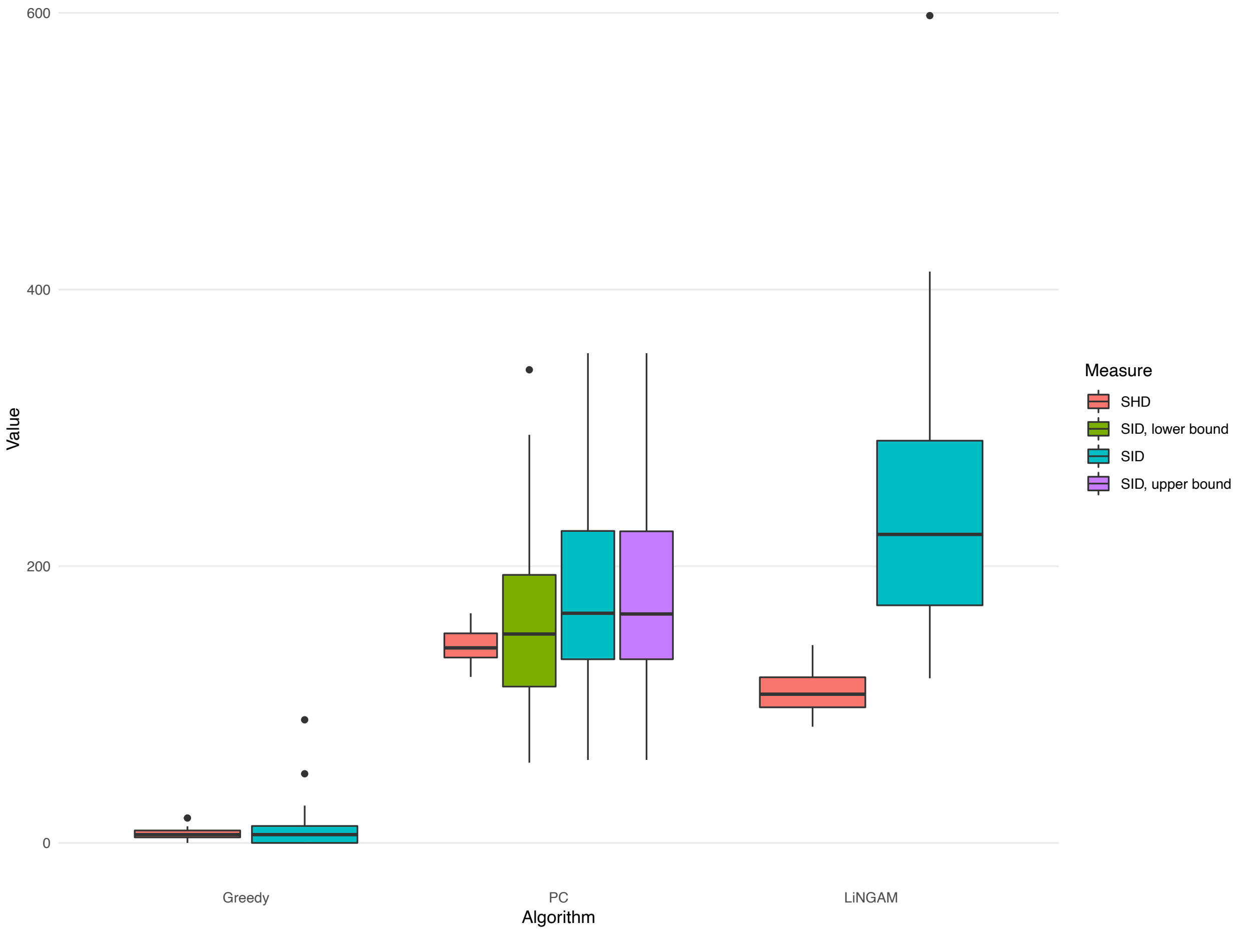

Figure 4.9: *Performance of the Greedy entropy-search based on* 40 *simulations compared to the PC and LiNGAM algorithms, each with sample size* 1000*. The simulations were run on ANMs with* 100 *nodes.*

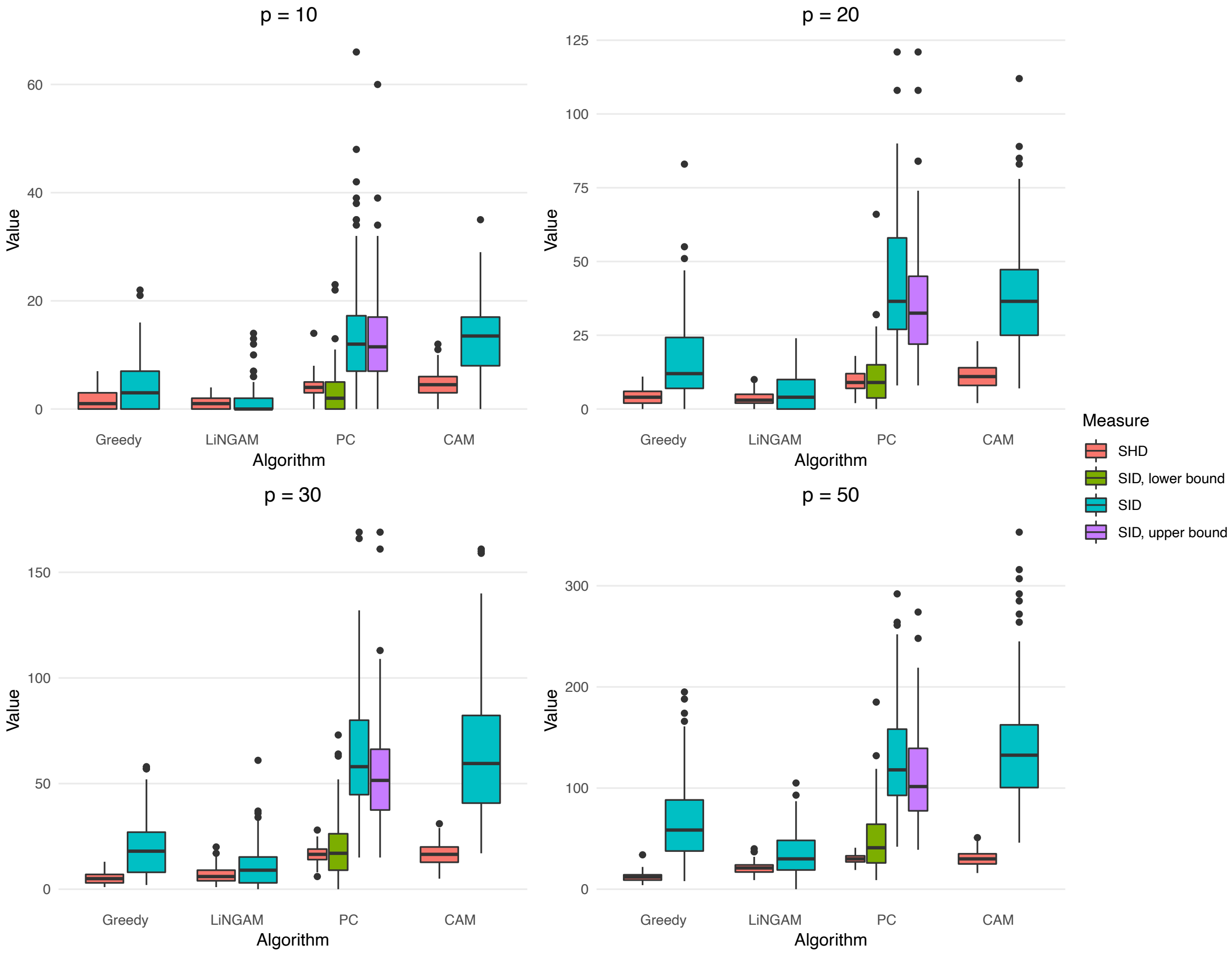

Figure 4.10: *Comparison of the Greedy entropy-search, the CAM, the LiNGAM and the PC algorithms based on* 100 *simulations, each with sample size* 1000*. The simulations were run on ANMs with* 10, 20, 30 *and* 50 *nodes respectively. Note that the y-scales are not fixed but vary by subfigure.*

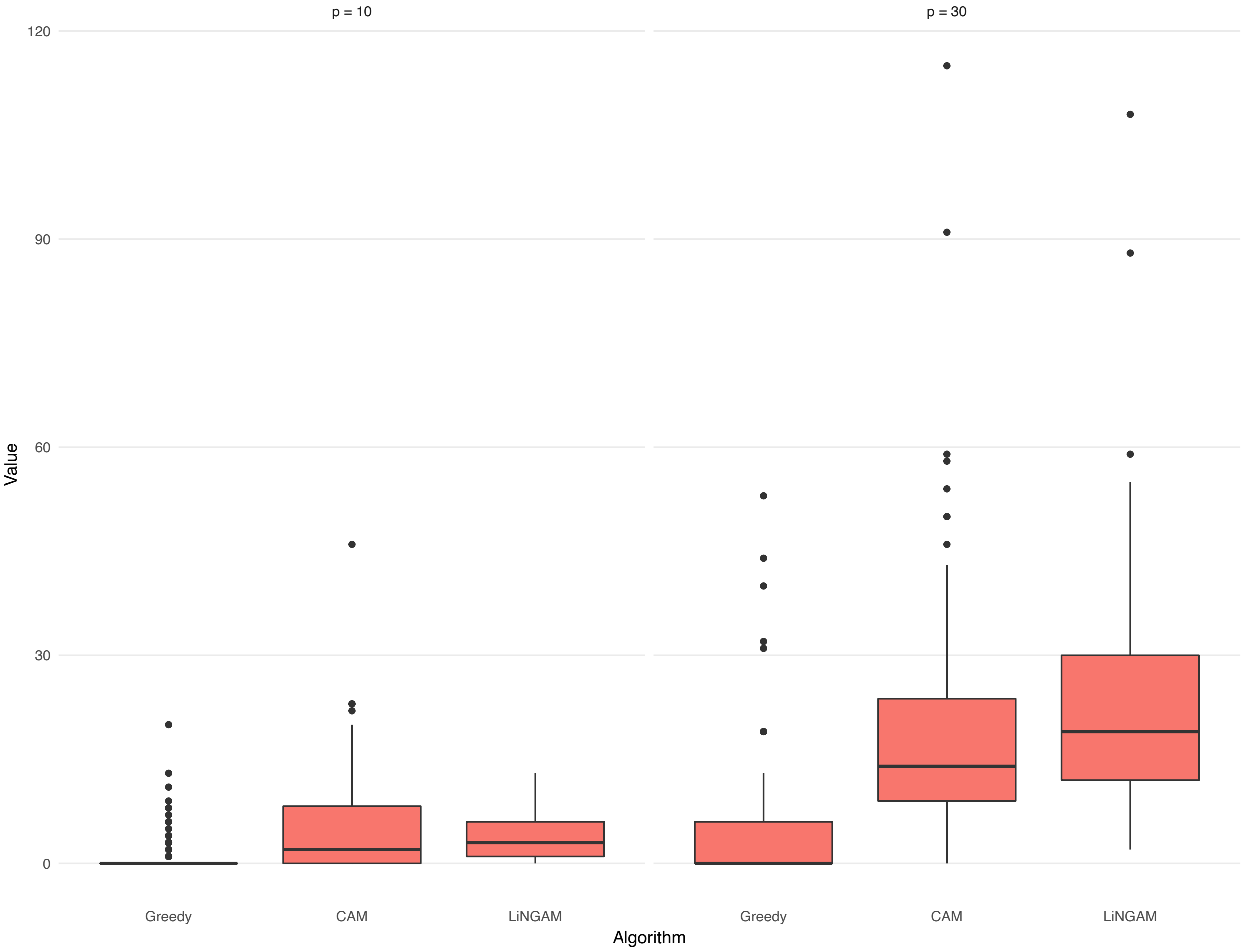

Figure 4.11: *Comparison of the Greedy entropy-search, the CAM and the LiNGAM algorithms based on* 100 *simulations, each with sample size* 1000*. The simulations were run on ANMs with* 10 *(left) and* 30 *nodes (right) respectively.*

CHAPTER 5

CONCLUSIONS AND FUTURE PERSPECTIVES

In this thesis we have proven optimality of a Greedy-like search algorithm for causal discovery in non-linear Additive Noise Models, proposed by Jonas Peters and Martin Wainwright in an unpublished manuscript. In order to prove optimality, we first proved that any arbitrary sub-model of an identifiable Additive Noise Model is itself an identifiable Additive Noise Model, under regularity conditions. A large part of proving this result was made possible by proving two different results. In one, we showed that any real analytic function either solves the differential equation $(\log f)'' = 0$ on the entire real or nowhere on the real line. In the second, we proved that the class of essentially bounded and real analytic functions is closed under convolution with integrable functions. By restricting attention to Additive Noise Models with a particular independence structure, it followed as a corollary, that any arbitrary sub-model of an identifiable Additive Noise Model that has non-linear assignments and noise variables with real analytic and strictly positive density functions, must itself be an identifiable Additive Noise Model. The version of the Greedy entropy-search proposed by Peters and Wainwright was restricted to Additive Noise Models whos graphs have no undirected cycles. We extended the Greedy entropy-search method work on a larger class of models, in which the graphs satisfy that every cycle contains at least three colliders. By imposing an additional constraint on the family of noise distributions, we proved the Greedy entropy-search to also be optimal in a subclass of linear, non-Gaussian Additive Noise Models.

By means of simulation, we found the Greedy entropy-search to perform at a comparable level to the Causal Additive Models (CAM) algorithm in term of Structural Hamming Distances (SHD), although slightly worse in terms of Structural Intervention Distances (SID). We applied the Greedy entropy-search on simulated datasets with up to 100 variables and found that it was able to quickly estimate graphs with high precision, as measured by SID. In a simulation example, we achieved a median SID of six and an average running time of just below six minutes when applied to a 1000 by 100 dataset. However, this simulation was performed on a small scale and the Greedy entropy-search suffered from being more volatile in its estimation procedure than the CAM algorithm. In linear Additive Noise Models, we found the Greedy entropy-search to be outperformed by the LiNGAM algorithm in terms of SID, but to be comparable in terms of SHD. However, when we considered non-Gaussian ANMs in which the functions were not restricted to being either linear or non-linear, the Greedy entropy-search out-performed both the CAM algorithm and the LiNGAM algorithm. When applying the Greedy entropy-search on 96 real datasets of known cause-effect pairs, we correctly identified the correct causal effect in roughly 65% of the cases, depending on the choice of regression method.

In conclusion, we believe that the Greedy entropy-search provides a promising method of causal discovery, as it is both computationally efficient and on a level comparable to existing methods in both a non-linear, Gaussian setting and a more general non-Gaussian setting.

5.1 FUTURE PERSPECTIVES AND EXTENSIONS

We believe that the proof of optimality that we gave in Chapter 3 can be made stronger than it currently is. In particular, it would be interesting to try and find a different way of proving the cases in which we made use of Lemma 3.17, since using this imposes a strong assumption on the class of functions we allow for. In addition, it remains to find consistent estimators of the marginal scores.

APPENDIX A

ADDITIONAL FIGURES

A.1 SUPPLEMENT TO FIGURE 4.2

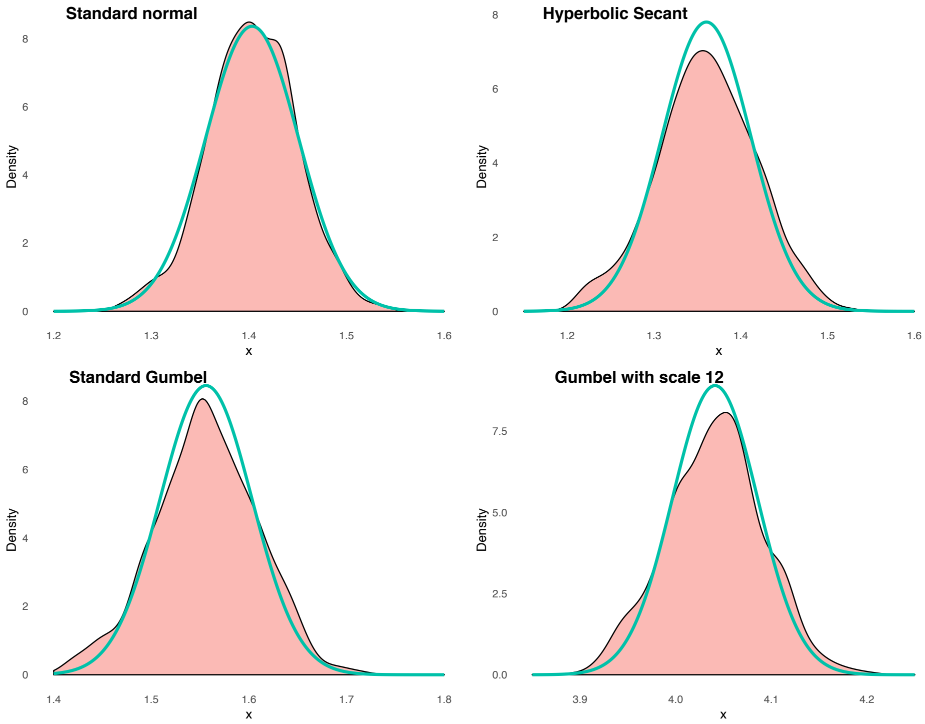

Figure A.1: *Simulated estimates of* $\hat{\mathbb{H}}_n$ *for four different distributions, each with with sample size* 250 *and* 1000 *repetitions. The superimposed lines are Guassian densities with standard deviation as estimated by* $\hat{\mathbb{V}}_n\hat{\mathbb{H}}_n$ *on a single sample. The filled areas mark the kernel density estimate of the* 1000 *realizations of* $\hat{\mathbb{H}}_n$.

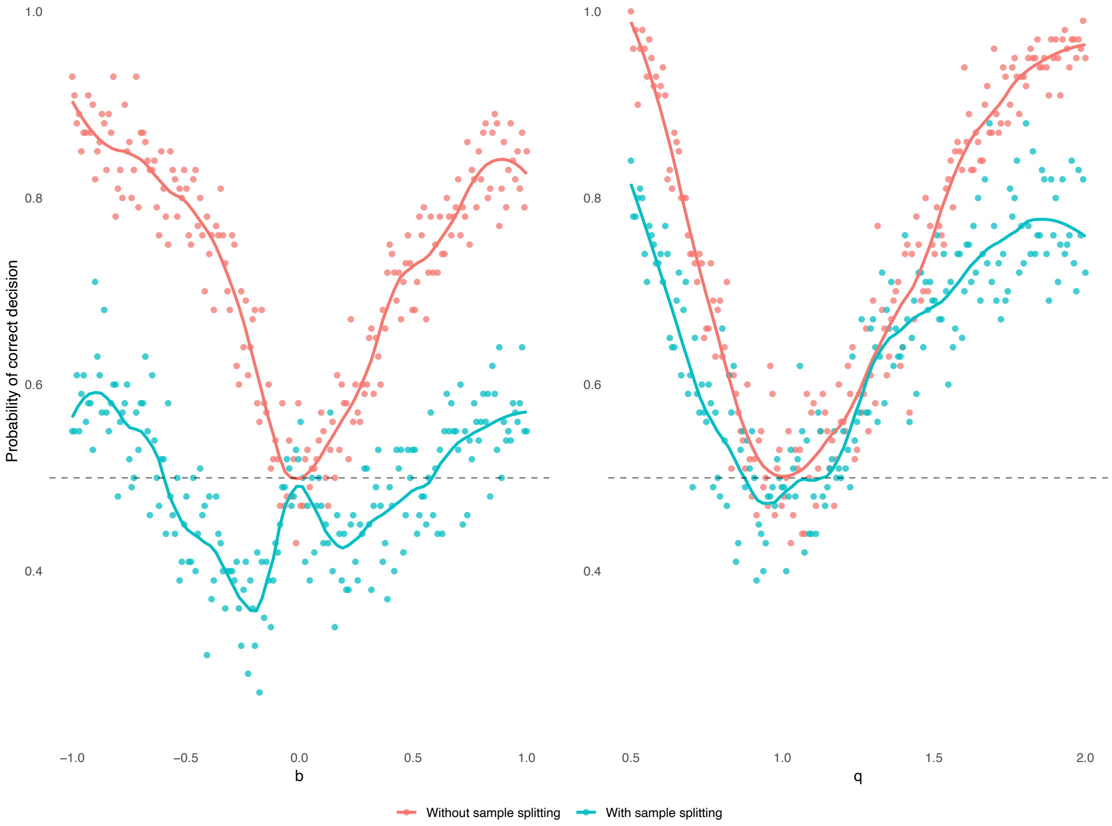

Figure A.2: *Estimated probabilities of correctly identifying $\mathcal{G}$ for different values of b (left) and q (right), respectively. Light blue points are results where the sample was split randomly in half during estimation. Red points are results from using the full sample. Superimposed lines are Loess smoothers indicate the general trends.*

A.2 SUPPLEMENT TO FIGURE 4.3

APPENDIX B

ADDITIONAL MATERIAL

B.1 DEFINITIONS OF THE DISTRIBUTIONS IN TABLE 2.1

A random variable, X, is log-mix-lin-exp distributed, if it has density on the form

$$\partial\mathbb{P}_X(x) = \exp\left(c_1 \exp(c_2 x) + c_3 x + c_4\right),$$

for choices of real constants, c_1 through c_4 such that the above is a density. X is one-sided asymptotically exponential if $\log \partial\mathbb{P}_X$ has a non-zero limit as either $x \to \infty$ or $x \to -\infty$. X is two-sided asymptotically exponential if $\log \partial\mathbb{P}_X$ does not vanish at *neither* ∞ *nor* at $-\infty$. X is a mixture of two exponentials if it has density on the form

$$\partial\mathbb{P}_X(x) = \exp\Big(c_1 x + c_2 \log\left(c_3 + c_4 \exp(c_5 x)\right) + c_6\Big),$$

for choices of real constants c_1 through c_6 such that the above is a density.

B.2 A DIFFERENT PROOF OF THEOREM 3.11, CASE 2.1

Case 2.1 - ϵ traverses $\mathbf{PA}_{\mathcal{G}^0}(\beta)$ and $\pi \in \Omega_\alpha$:
Suppose the d-connecting path ϵ traverses a parent of β and that π belongs to Ω_α. Denote by ρ the parent of β on ϵ. As π is a $\mathcal{G}^s$-parent of α, then

$$X_\alpha \perp\!\!\!\perp X_\beta \mid X_\pi,$$

by the local Markov property. Proposition 3.13 then implies that

$$\Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \beta \to \alpha\right) = 0.$$

By the same argument, if $\rho \in \Omega_\beta$ then $\Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \alpha \to \beta\right) = 0$, which would be a contradiction. Suppose then that ρ is *not* a $\mathcal{G}^s$ parent of β. We will show that $\Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \rho \to \beta\right) > \Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \alpha \to \beta\right)$. As ρ and Ω_β both belong to $\mathbf{PA}_{\mathcal{G}^0}(\beta)$, we can apply Remark 3.14 to find

$$\mathbb{V}\left(X_\beta - \sum_{\omega\in\Omega_\beta} \hat{f}_\omega(X_\omega) \;\; - \hat{f}_\rho(X_\rho)\right)$$
$$\overset{(a)}{=} \mathbb{V}\Big(X_\beta - \mathbb{E}[X_\beta \mid (X)_{\Omega_\beta}, X_\rho]\Big)$$

$$
\begin{aligned}
&= \mathbb{V}\left(N_\beta + \sum_{\omega\in\Omega_\beta} f^0_{\beta,\omega}(X_\omega) + \sum_{\theta\in\Theta_\beta\setminus\{\rho\}} f^0_{\beta,\theta}(X_\theta) + f^0_{\beta,\rho}(X_\rho) \right.\\
&\qquad \left. - \ \mathbb{E}\left[N_\beta + \sum_{\omega\in\Omega_\beta} f^0_{\beta,\omega}(X_\omega) + \sum_{\theta\in\Theta_\beta\setminus\{\rho\}} f^0_{\beta,\theta}(X_\theta) + f^0_{\beta,\rho}(X_\rho) \,\middle|\, (X)_{\Omega_\beta}, X_\rho \right]\right)\\
&\overset{(b)}{=} \mathbb{V}\left(N_\beta - \mathbb{E}N_\beta + \sum_{\theta\in\Theta_\beta\setminus\{\rho\}} f^0_{\beta,\theta}(X_\theta) - \mathbb{E}\left(\sum_{\theta\in\Theta_\beta\setminus\{\rho\}} f^0_{\beta,\theta}(X_\theta)\right)\right)\\
&= \mathbb{V}\left(N_\beta + \sum_{\theta\in\Theta_\beta\setminus\{\rho\}} f^0_{\beta,\theta}(X_\theta)\right). \qquad \text{(B.1)}
\end{aligned}
$$

In the above, equality (a) is a direct application of Remark 3.14. To get equality (b) we use that X_ρ and $(X)_{\Omega_\beta}$ are $\sigma((X)_\Omega, X_\rho)$-measurable to pull these terms out of the conditional expectations, after which they cancel out, and that

$$
(N_\beta, (X)_{\Theta_\beta\setminus\{\rho\}}) \perp\!\!\!\perp (X_\rho, (X)_{\Omega_\beta}),
$$

as $\mathcal{G}^0$ is unrelated by assumption, to conclude (b), to change the conditioning algebra on the remaining terms. Next, we consider the residual variance when adding ($\alpha \to \beta$):

$$
\begin{aligned}
&\mathbb{V}\left(X_\beta - \sum_{\omega\in\Omega_\beta} \hat{f}_\omega(X_\omega) \ - \hat{f}_\alpha(X_\alpha)\right)\\
&\qquad = \mathbb{V}\left(N_\beta + \sum_{\omega\in\Omega_\beta} f^0_{\beta,\omega}(X_\omega) + \sum_{\theta\in\Theta_\beta\setminus\{\rho\}} f^0_{\beta,\theta}(X_\theta) + f^0_{\beta,\rho}(X_\rho) - \sum_{\omega\in\Omega_\beta} \hat{f}_\omega(X_\omega) - \hat{f}_\alpha(X_\alpha)\right)\\
&\qquad \overset{(c)}{=} \mathbb{V}\left(N_\beta + \sum_{\theta\in\Theta_\beta\setminus\{\rho\}} f^0_{\beta,\theta}(X_\theta)\right) + \mathbb{V}\left(\sum_{\omega\in\Omega_\beta} f^0_{\beta,\omega}(X_\omega) + f^0_{\beta,\rho}(X_\rho) - \sum_{\omega\in\Omega_\beta} \hat{f}_\omega(X_\omega) - \hat{f}_\alpha(X_\alpha)\right)\\
&\qquad > \mathbb{V}\left(N_\beta + \sum_{\theta\in\Theta_\beta\setminus\{\rho\}} f^0_{\beta,\theta}(X_\theta)\right)\\
&\qquad \overset{(d)}{=} \mathbb{V}\left(X_\beta - \sum_{\omega\in\Omega_\beta} \hat{f}_\omega(X_\omega) - \hat{f}_\rho(X_\rho)\right), \qquad \text{(B.2)}
\end{aligned}
$$

where equality (c) follows from

$$
(N_\beta, (X)_{\Theta_\beta\setminus\{\rho\}}) \perp\!\!\!\perp (X_\rho, X_\alpha, (X)_{\Omega_\beta}).
$$

As the functions $(\hat{f})_{\Omega_\beta}$ and $\hat{f}_\alpha$ are all elements in $\mathcal{F}$, they are continuous and therefore measurable. As independence is retained over measurable mappings (see Sokol and Rønn-Nielsen [2016, Lemma 1.3.7]), equality (c) follows. To get equality (d), we simply use the result from (B.1). But this implies that

$$\Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \rho \to \beta\right) > \Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \alpha \to \beta\right)$$

which is a contradiction.

BIBLIOGRAPHY

Jan Beirlant, Edward J Dudewicz, László Györfi, and Edward C Van der Meulen. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39, 1997.

Thomas Berrett, Richard Samworth, and Ming Yuan. Efficient multivariate entropy estimation via k-nearest neighbour distances. *Annals of Statistics*, 47(1), 2019. ISSN 00905364. URL `http://search.proquest.com/docview/2155913331/`.

Thomas B. Berrett, Daniel J. Grose, and Richard J. Samworth. *IndepTest: Nonparametric Independence Tests Based on Entropy Estimation*, 2018. URL `https://CRAN.R-project.org/package=IndepTest`. R package version 0.2.0.

Alina Beygelzimer, Sham Kakadet, John Langford, Sunil Arya, David Mount, and Shengqiao Li. *FNN: Fast Nearest Neighbor Search Algorithms and Applications*, 2019. URL `https://CRAN.R-project.org/package=FNN`. R package version 1.1.3.

Peter Bühlmann, Jonas Peters, Jan Ernest, et al. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

T. M Cover. *Elements of information theory.* A Wiley-Interscience publication. Wiley-Interscience, Hoboken, N.J, second edition edition, 2006. ISBN 0471748811.

Claus Dethlefsen and Søren Højsgaard. A common platform for graphical models in R: The gRbase package. *Journal of Statistical Software*, 14(17):1–12, 2005. URL `http://www.jstatsoft.org/v14/i17/`.

Peng Ding. Three occurrences of the hyperbolic-secant distribution. *The American Statistician*, 68 (1):32–35, 2014.

Peng Ding and Joseph K. Blitzstein. On the gaussian mixture representation of the laplace distribution. *The American Statistician*, 72(2):172–174, 2018. ISSN 0003-1305.

Dheeru Dua and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Encyclopedia of Mathematics. Hilbert space. Encyclopedia of Mathematics. . Website, a. URL: `http://www.encyclopediaofmath.org/index.php?title=Hilbert_space&oldid=24082`. Accessed on 19/02/2019.

Encyclopedia of Mathematics. Real analytic function. Website, b. URL: http://www.encyclopediaofmath.org/index.php?title=Real_analytic_function&oldid=31091. Accessed on 24/02/2019.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of non-parametric regression*. Springer series in statistics. Springer, New York, 2002. ISBN 0387954414.

Ernst Hansen. *Measure theory*. University of Copenhagen. Department of mathematical sciences, Copenhagen, 4th ed. edition, 2009. ISBN 9788791927447.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464, 2012. URL http://jmlr.org/papers/v13/hauser12a.html.

Christina Heinze-Deml, Marloes H. Maathius, and Nicolai Meinshausen. Causal structure learning. *Annual Review of Statistics and Its Application*, 8, 2018.

Benjamin Hofner, Luigi Boccuto, and Markus Goeker. Controlling false discoveries in high-dimensional situations: Boosting with stability selection. *BMC Bioinformatics*, 16(144), 2015.

John Hopcroft and Robert Tarjan. Algorithm 447: Efficient algorithms for graph manipulation. *Commun. ACM*, 16(6):372–378, June 1973. ISSN 0001-0782. doi: 10.1145/362248.362272. URL http://doi.acm.org/10.1145/362248.362272.

Patrik O. Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 689–696. Curran Associates, Inc., 2009a. URL http://papers.nips.cc/paper/3548-nonlinear-causal-discovery-with-additive-noise-models.pdf.

Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009b.

Menno Hulswit. A short history of causation. *SEED Journal (Semiotics, Evolution, Energy, and Development)*, 4(3):16–42, 2004.

Olav Kallenberg. *Foundations of Modern Probability*. Probability and its applications. Springer, New York, 2. ed. edition, 2002. ISBN 0387953132.

Charles Kooperberg. *logspline: Routines for Logspline Density Estimation*, 2019. URL https://CRAN.R-project.org/package=logspline. R package version 2.1.12.

LF Kozachenko and Nikolai N Leonenko. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16, 1987.

Samory Kpotufe, Eleni Sgouritsa, Dominik Janzing, and Bernhard Schölkopf. Consistency of causal inference under the additive noise model. 2013.

Steven George Krantz and Harold R Parks. *A primer of real analytic functions*. Birkhäuser advanced texts Basler Lehrbücher. Birkhäuser, Boston, 2. ed. edition, 2002. ISBN 3764342641.

Steffen L. Lauritzen. *Lectures on Graphical Models*. Department of Mathematical Sciences, Faculty of Science, University of Copenhagen, 2017. ISBN 978-87-70787-53-6.

Brendan D. McKay, Frederique E. Oggier, Gordon F. Royle, N. J. A. Sloane, Ian M. Wanless, and Herbert S. Wilf. Acyclic digraphs and eigenvalues of (0,1)-matrices. 2003.

Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016. URL `http://jmlr.org/papers/v17/14-518.html`.

Christopher Nowzohour and Peter Bühlmann. Score-based causal learning in additive noise models. *Statistics*, 50(3):471–485, 2016.

Judea Pearl. *Causality : Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2009. ISBN 9780511803161.

J Peters, Jm Mooij, D Janzing, and B Scholkopf. Causal discovery with continuous additive noise models. *Journal Of Machine Learning Research*, 15:2009–2053, 2014. ISSN 1532-4435.

Jonas Peters. *SID: Structural Intervention Distance*, 2015. URL `https://CRAN.R-project.org/package=SID`. R package version 1.0.

Jonas Peters and Peter Bühlmann. Structural intervention distance (sid) for evaluating causal graphs. *arXiv preprint arXiv:1306.1043*, 2013a.

Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2013b.

Jonas Peters and Jan Ernest. *CAM: Causal Additive Model (CAM)*, 2015. URL `https://CRAN.R-project.org/package=CAM`. R package version 1.0.

Jonas Peters and Martin Wainwright. Estimating the structure of additive noise models with greedy search algorithms. unpublished, Unpublished.

Jonas Martin Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference, foundations and learning algorithms*. Adaptive computation and machine learning series. MIT Press, Cambridge, MA, 2017. ISBN 9780262037310.

Walter Rudin. *Principles of mathematical analysis.* International series in pure and appled mathematics. McGraw-Hill, Tokyo, 3. ed. edition, 1976. ISBN 007054235X.

Walter Rudin. *Real and complex analysis.* McGraw-Hill International Editions, Mathematics Series. McGraw-Hill, New York, USA, 3. edition, 1987. ISBN 0071002766.

René L. Schilling. *Measures, Integrals and Martingales.* Cambridge University Press, 2005. ISBN 9780521850155.

C Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001. ISSN 1931-1222.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.

Alexander Sokol and Anders Rønn-Nielsen. *Advanced probability.* Department of Mathematical Sciences, University of Copenhagen, Copenhagen, 4. ed. edition, 2016. ISBN 9788770789479.

Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search.* MIT press, 2000.

Eric W. Weisstein. "binomial theorem." from mathworld–a wolfram web resource, a. URL `http://mathworld.wolfram.com/BinomialTheorem.html`. Accessed on 23/02/2019.

Eric W. Weisstein. "gumbel distribution." from mathworld–a wolfram web resource, b. URL `http://mathworld.wolfram.com/GumbelDistribution.html`. Accessed on 24/02/2019.

Eric W. Weisstein. "logistic distribution." from mathworld–a wolfram web resource., c. URL `http://mathworld.wolfram.com/LogisticDistribution.html`. Accessed on 24/02/2019.

S. N. Wood. Thin-plate regression splines. *Journal of the Royal Statistical Society (B)*, 65(1):95–114, 2003.

Kun Zhang and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. 2012.