Most Frugal Explanations: Occam's Razor Applied to Bayesian Abduction

Johan Kwisthout

Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen
PO Box 9104, 6500HE Nijmegen, The Netherlands, j.kwisthout@donders.ru.nl

Abstract

What constitutes 'Best' in 'Inference to the Best Explanation' has been hotly debated. In Bayesian models the traditional interpretation is 'Best = Most Probable'. We propose an alternative notion, denoted as Most Frugal Explanation (MFE), that utilizes the fact that only few variables actually are relevant for deciding upon the best explanation. We show that MFE is intractable in general, but can be tractably approximated under plausible situational constraints.

1 Introduction

Abduction or inference to the best explanation refers to the process of finding a suitable explanation of observed data or phenomena. In the last decades, Bayesian notions of abduction have emerged due to the widespread popularity of Bayesian techniques for representing and reasoning with knowledge [16, 24]. They are used in decision support systems in a wide range of problem domains such as [3, 7, 13] and as computational models of economic, social, or cognitive processes [6, 25, 26]. A natural question is of course *what is seen as best*. Apart from the obvious interpretation—the best explanation is the one with maximum posterior probability—other relationships have been proposed to describe *why* we judge one explanation to be preferred over another [22], like various measures based on a Bayesian account of *coherence theory* [10, 14]. Such alternative formalisms put an emphasis on different properties of 'good' explanations, e.g., that they are coherent with the available evidence. While the posterior probability of such explanations is not the deciding criterion to prefer one explanation over another, it is typically so that explanations we consider to be good for other reasons also have a high probability compared to alternative explanations [15].

However, computing explanations is computationally costly, especially when there are many intermediate (neither observed nor to be explained) variables that may influence the explanation. One way of dealing with this intractability might be by assuming modularity of knowledge representations. However, this is problematic as we cannot know beforehand which elements of background knowledge or observations may be relevant for determining the best explanation [11]. Fortunately, even when a full Bayesian analysis may not be feasible, we need not constrain inferences only to small or encapsulated knowledge structures. It is known that in general only few of the variables in a network are relevant to a particular inference query [9]. We propose to utilize this property of Bayesian networks in order to make tractable approximate inferences to the best explanation over large unencapsulated knowledge structures. This novel explanation formalism, denoted as Inference to the Most Frugal Explanation (MFE), is explicitly designed to reflect that only few variables are typically relevant in real-world situations. Our aim here is to leave out those variables not needed to deciding upon an explanation, in a loose sense thus applying *Occam's razor* to Bayesian abduction.

MOST FRUGAL EXPLANATION (MFE)
Instance: A Bayesian network, partitioned into a set of observed evidence variables, a set of explanation variables, a set of 'relevant' intermediate variables that are marginalized over, and a set of 'irrelevant' intermediate variables that are not marginalized over.
Output: The joint value assignment to the nodes in the explanation set that is most probable for the maximal number of joint value assignments to the irrelevant intermediate variables.

In the remainder of this paper, we will discuss some needed preliminaries in Section 2. In Section 3 we discuss MFE more thoroughly. We give a more formal definition, including a formal definition of (normative) relevance in the context of Bayesian networks. We show that, despite intractability of the problem in general, MFE can be tractably approximated under plausible assumptions. We conclude in Section 4.

2 Preliminaries

In this section we will introduce some preliminaries from Bayesian networks, in particular the MAP problem as standard formalization of Bayesian abduction. We will discuss the ALARM network which we will use as a running example throughout this paper. Lastly, we introduce some needed concepts from parameterized complexity theory.

2.1 Bayesian networks

A Bayesian or probabilistic network $\mathcal{B}$ is a graphical structure that models a set of stochastic variables, the conditional independences among these variables, and a joint probability distribution over these variables. $\mathcal{B}$ includes a directed acyclic graph $\mathbf{G}_\mathcal{B} = (\mathbf{V}, \mathbf{A})$, modeling the variables and conditional independences in the network, and a set of parameter probabilities $\Pr$ in the form of conditional probability tables (CPTs), capturing the strengths of the relationships between the variables. The network models a joint probability distribution $\Pr(\mathbf{V}) = \prod_{i=1}^{n} \Pr(V_i \mid \pi(V_i))$ over its variables, where $\pi(V_i)$ denotes the parents of V_i in $\mathbf{G}_\mathcal{B}$. We will use upper case letters to denote individual nodes in the network, upper case bold letters to denote sets of nodes, lower case letters to denote value assignments to nodes, and lower case bold letters to denote joint value assignments to sets of nodes. By convention, we will use $\mathbf{E}$, $\mathbf{H}$, and $\mathbf{I}$, to denote the set of evidence variables, the set of explanation variables, and the set of intermediate variables, respectively. The problem of determining the most probable joint value assignment to the explanation set given evidence is defined as MAP[1]. As a decision problem, MAP is formally defined as follows.

MAXIMUM A POSTERIORI PROBABILITY (MAP)
Instance: A Bayesian network $\mathcal{B} = (\mathbf{G}_\mathcal{B}, \Pr)$, where $\mathbf{V}$ is partitioned into evidence variables $\mathbf{E}$ with joint value assignment $\mathbf{e}$, explanation variables $\mathbf{H}$, and intermediate variables $\mathbf{I}$; a rational number $0 \leq q < 1$.
Question: Is there a joint value assignment $\mathbf{h}$ to the nodes in $\mathbf{H}$ such that $\Pr(\mathbf{h}, \mathbf{e}) > q$?

2.2 The ALARM network

The ALARM network [1] will be used throughout this paper as a running example. It consists of thirty-seven discrete variables. Eight of these variables are diagnostic variables, indicating problems like pulmonary embolism or a kinked tube; another sixteen variables indicate measurable or observable findings. The remaining thirteen variables are intermediate variables, i.e., they are neither diagnostic variables, nor can be observed (in principle or in practice). As an example, consider that a high breathing pressure was detected (PRSS = high) and that minute ventilation was low (MINV = low); all other observable variables take their default (i.e., non-alarming) value. From these findings a probability of 0.92 for the diagnose 'kinked tube' (KINK = true) can be computed. Likewise, we can compute that the most probable joint explanation for the diagnostic variables, given that PCWP (pulmonary capillary wedge pressure) and BP (blood pressure) are high, is that HYP = true (hypovolemia, viz., loss of blood volume) and all other diagnostic variables are negative. This joint value assignment has probability 0.58. The second-best explanation (all diagnostic variables are negative, despite the two alarming conditions) has probability 0.11.

2.3 Parameterized complexity theory

In the remainder, we assume that the reader is familiar with basic concepts of computational complexity theory inasmuch they are related to Bayesian computations. In particular we assume familiarity with Turing Machines, the complexity classes NP and PP, oracles, and intractability proofs. For more background we refer to textbooks like [12] and [4]. In addition to these basic concepts we will shortly, and somewhat informally, introduce parameterized complexity theory. A more thorough introduction can be found in [8].

[1] Also PARTIAL MAP or MARGINAL MAP to emphasize that the probability of any such joint value assignment is computed by marginalization over the intermediate variables.

Sometimes problems are intractable (i.e., NP-hard) in general, but become tractable if some *parameters* of the problem can be assumed to be small. Informally, a problem is called fixed-parameter tractable for a set of parameters $k = \{k_1, \ldots, k_m\}$ if it can be solved in time, exponential *only* in k and polynomial in the input size $|x|$. In practice, this means that problem instances can be solved efficiently, even when the problem is NP-hard in general, if $\{k_1, \ldots, k_m\}$ are known to be small. The notion of fixed-parameter tractability can be extended to deal with *rational*, rather than integer, parameters [18]. Informally, if a problem is fixed-rational tractable for a (rational) parameter k_i, then the problem can be solved tractably if k_i is close to 0. For readability, we will liberally mix integer and rational parameters in the remainder.

3 Most Frugal Explanations

In real-world applications there are many intermediate variables that are neither observed nor to be explained, yet may influence the explanation. Some of these variables can considerably affect the outcome of the abduction process. Most of these variables, however, are irrelevant as they are not expected to influence the outcome of the abduction process in all but maybe the very rarest of cases [9]. To compute the most probable, most likely, or most coherent explanation of the evidence, however, one needs to marginalize over all these variables, that is, take their prior or conditional probability distribution into account. This seems a waste of computing resources when we might as well have assigned an arbitrary value to these variables and still arrive at the same explanation. One way of ensuring tractability of inference may be by 'weeding out' the irrelevant aspects in the knowledge structure prior to inference. Yet, it is quite unpractical to construct and represent a subset of the entire knowledge structure for every new query of the belief system: this may buy tractability for the abductive inference itself, but requires extensive computations to construct a subset of the 'relevant' variables and the probabilistic relationships between them. Therefore we assume that inferences are made on the (entire) knowledge structure, rather than re-representing priors and conditionals in order to do inference on subsets of the knowledge structures. We propose that marginalization is done only on a subset of the intermediate variables (the variables that are considered to be relevant), and that a sampling strategy is used for the remaining intermediate variables that are not considered to be relevant. Such a sampling strategy may be very simple ('decide using a singleton sample') or more complex ('compute the best explanation on N samples and take a majority vote').

Example 1. In the ALARM network, let us assume that, given the observations that PCWP and BP are high, we consider VTUB, SHT, VLNG, VALV and LVV to be relevant intermediate variables, and VMCH, PVS, ACO2, CCHL, ERLO, STKV, HR, and ERCA to be irrelevant variables. The most *frugal* joint explanation for the diagnostic variables is still that HYP = true while all other diagnostic variables are negative: in 31% of the joint value assignments to these irrelevant intermediate variables, this is the most probable explanation. In 16% of the assignments 'all negative' is the most probable explanation, and in 24% of the assignments HYP = true and INT = onesided (onesided intubation, rather than normal) is the most probable explanation of the observations. If, in addition, we also consider VMCH, PVS, and STKV to be relevant, then every joint value assignment to ACO2, CCHL, ERLO, HR, and ERCA will have HYP = true as the most probable explanation for the observations. In other words, rather than marginalizing over these variables, we might have assigned just an arbitrary joint value assignment over these variables, decreasing the computational burden. If we had considered less intermediate variables to be relevant, this strategy may still often work, but has a chance of error, if we pick a sample for which a different explanation is the most probable one. We can decrease this error by taking more samples and take a majority vote.

Note that MFE is not *guaranteed* to give the MAP explanation, unless we marginalize over all intermediate variables. Even with a voting strategy based on *all* joint value assignments to the irrelevant intermediate variables, we may still end up with a different explanation as explanations are computed differently.

3.1 Relevance

Until now, we have quite liberally used the notion 'relevance'. In this paper, we make a distinction between the *intrinsic* or normative and *expected* or subjective relevance of the intermediate variables. The intrinsic relevance is a statistical property of an intermediate variable that is based on Druzdzel and Suermondt's [9] definition of relevance of variables in a Bayesian model. According to Druzdzel and Suermondt a variable in a Bayesian model is relevant for a set $\mathbf{T}$ of variables, given an observation $\mathbf{E}$, if it is "needed to reason about the impact of observing $\mathbf{E}$ on $\mathbf{T}$" [9, p.60]. Our operationalization of "needed to reason" is inspired

by Wilson and Sperber's [28] relevance theory, who state that "an input is relevant to an individual when its processing in a context of available assumptions yields (...) a worthwhile difference to the individual's representation of the world" [28, p.608]. The term 'worthwhile difference' in this quote refers to the balance between the actual effects of processing that particular input and the effort required to do so. We define intrinsic relevance of a variable as a *measure*, indicating how sensitive explanations are to its actual value, so that this measure can be used to assess the 'worthwhileness' of considering this variable. Informally, an intermediate variable I has a low intrinsic relevance when there are only few 'possible worlds' in which the most probable explanation changes when the value of I changes.

Definition 2. Let $\mathcal{B} = (\mathbf{G}_{\mathcal{B}}, \Pr)$ be a Bayesian network partitioned into evidence nodes $\mathbf{E}$ with joint value assignment $\mathbf{e}$, intermediate nodes $\mathbf{I}$, and an explanation set $\mathbf{H}$. Let $I \in \mathbf{I}$, and let $\Omega(\mathbf{I} \setminus \{I\})$ denote the set of joint value assignments to the intermediate variables other than I. The *intrinsic relevance* of I is the fraction of joint value assignments $\mathbf{i}$ in $\Omega(\mathbf{I} \setminus \{I\})$ for which $\text{argmax}_{\mathbf{h}} \Pr(\mathbf{h}, \mathbf{e}, \mathbf{i}, i)$ is not identical for all $i \in \Omega(I)$.

The *expected* relevance of I is a subjective assessment of the intrinsic relevance of I which may or may not correspond to the actual value. Such a subjective assessment might be based on heuristics, previous knowledge, or by approximating the intrinsic relevance, e.g., by sampling a few instances of $\Omega(\mathbf{I} \setminus \{I\})$. Note that both intrinsic and expected relevance of a variable are relative to a particular set of candidate explanations $\mathbf{H}$, and conditional on a particular observation $\mathbf{e}$.

Example 3. Let, in the ALARM network, again PCWP and BP be high, and let all other observable variables take their non-alarming default values. The intrinsic relevance of the intermediate variables for the diagnosis is given in Table 1. Note that the left ventricular end-diastolic blood volume (LVV) is highly relevant for the diagnosis, while the amount of catecholamines in the blood (CCHL) is irrelevant given these observations.

Variable	VMCH	VTUB	SHNT	VLNG	VALV	PVS	ACO2	CCHL	LVV	ERLO	STKV	HR	ERCA
Relevance	0.53	0.80	0.88	0.76	0.64	0.24	0.00	0.00	0.94	0.00	0.57	0.00	0.00

Table 1: Intrinsic relevancy of intermediate variables in the ALARM network

When solving an MFE problem, we marginalize over the 'relevant intermediate variables'. We assume that the partition between relevant and irrelevant is made, based on some threshold on the (subjective) expected relevance of the intermediate variables. For example, if the threshold would be 0.85 then only SHNT and LVV would be relevant intermediate variables in the ALARM network, but if the threshold would be 0.40 then also VMCH, VTUB, VLNG, VALV, and STKV would be relevant variables. That influences the results, as the distribution of MFE explanations tends to be flatter when less variables are marginalized over. With a threshold of 0.85 there are 24 explanations that are sometimes the MFE, with the actual MAP explanation occurring most often (26%). With a threshold of 0.40 there are just three such explanations, with the MAP explanation occurring in 75% of the cases. Thus, the distribution of MFE explanations is more 'skewed' towards one explanation when more variables are considered to be relevant.

3.2 Complexity Analysis

To assess the computational complexity of MFE, we first define a decision variant.

MOST FRUGAL EXPLANATION (MFE)
Instance: A Bayesian network $\mathcal{B} = (\mathbf{G}_{\mathcal{B}}, \Pr)$, where $\mathbf{V}$ is partitioned into a set of evidence nodes $\mathbf{E}$ with a joint value assignment $\mathbf{e}$, an explanation set $\mathbf{H}$, a set of *relevant* intermediate variables $\mathbf{I}^+$, and a set of *irrelevant* intermediate variables $\mathbf{I}^-$; a rational number $0 \leq q < 1$ and an integer $0 \leq k \leq |\Omega(\mathbf{I}^-)|$.
Question: Is there a joint value assignment $\mathbf{h}$ to the nodes in $\mathbf{H}$ such that for more than k joint value assignments $\mathbf{i}$ to $\mathbf{I}^-$, $\Pr(\mathbf{h}, \mathbf{i}, \mathbf{e}) > q$?

It will be immediately clear that MFE is intractable, as it has the $\mathsf{NP}^{\mathsf{PP}}$-complete MAP and MSE [17] problems as degenerate cases for $\mathbf{I}^- = \varnothing$, respectively $\mathbf{I}^+ = \varnothing$. In this section we shown that MFE is $\mathsf{NP}^{\mathsf{PP}^{\mathsf{PP}}}$-complete, making it one of few real world-problems that are complete for that class[2]. The canonical SATISFIABILITY-variant that is complete for this class is E-MAJMAJSAT, defined as follows [27].

[2]Informally, one could imagine that for solving MFE one needs to counter *three* sources of complexity: selecting a joint value assignment out of potentially exponentially many candidate assignments to the explanation set; solving an inference problem over the

EMAJMAJSAT
Instance: A Boolean formula ϕ whose n variables $x_1 \dots x_n$ are partitioned into three sets $\mathbf{E} = x_1 \dots x_k$, $\mathbf{M_1} = x_{k+1} \dots x_l$, and $\mathbf{M_2} = x_{l+1} \dots x_n$ for some numbers k, l with $1 \leq k \leq l \leq n$.
Question: Is there a truth assignment to the variables in $\mathbf{E}$ such that for the majority of truth assignments to the variables in $\mathbf{M_1}$ it holds, that the majority of truth assignments to the variables in $\mathbf{M_2}$ yield a satisfying truth instantiation to $\mathbf{E} \cup \mathbf{M_1} \cup \mathbf{M_2}$?

As an example, consider the formula $\phi_{\text{ex}} = x_1 \wedge (x_2 \vee x_3) \wedge (x_4 \vee x_5)$ with $\mathbf{E} = \{x_1\}$, $\mathbf{M_1} = \{x_2, x_3\}$ and $\mathbf{M_2} = \{x_4, x_5\}$. This is a *yes* example of E-MAJMAJSAT: for $x_1 = \text{TRUE}$, three out of four truth assignments to $\{x_2, x_3\}$ (all but $x_2 = x_3 = \text{FALSE}$) are such that the majority of truth assignments to $\{x_4, x_5\}$ satisfy ϕ_{ex}.

To prove $\mathsf{NP}^{\mathsf{PP}^{\mathsf{PP}}}$-completeness of the MFE problem, we construct a Bayesian network $\mathcal{B}_\phi$ from an E-MAJMAJSAT instance $(\phi, \mathbf{E}, \mathbf{M_1}, \mathbf{M_2})$. For each propositional variable x_i in ϕ, a binary stochastic variable X_i is added to $\mathcal{B}_\phi$, with uniformly distributed values TRUE and FALSE. These stochastic variables in $\mathcal{B}_\phi$ are three-partitioned into sets $\mathbf{X_E}$, $\mathbf{X_{M_1}}$, and $\mathbf{X_{M_2}}$ according to the partition of ϕ. For each logical operator in ϕ an additional binary variable in $\mathcal{B}_\phi$ is introduced, whose parents are the variables that correspond to the input of the operator, and whose conditional probability table is equal to the truth table of that operator. The variable associated with the top-level operator in ϕ is denoted as V_ϕ, the set of remaining operators is denoted as $\mathbf{Op}_\phi$. Figure 1 shows the graphical structure of the Bayesian network constructed for the example E-MAJMAJSAT instance given above.

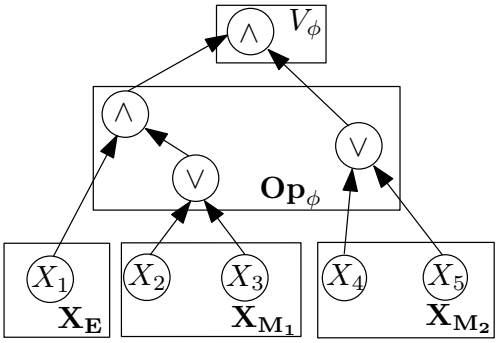

Figure 1: Example of the construction of $\mathcal{B}_{\phi_{\text{ex}}}$ for the Boolean formula $\phi_{\text{ex}} = x_1 \wedge (x_2 \vee x_3) \wedge (x_4 \vee x_5)$

Theorem 4. MFE *is* $\mathsf{NP}^{\mathsf{PP}^{\mathsf{PP}}}$*-complete.*

Proof. Membership in $\mathsf{NP}^{\mathsf{PP}^{\mathsf{PP}}}$ follows from the following algorithm: non-deterministically guess a value assignment $\mathbf{h}$ for which there are at least k joint value assignments $\mathbf{i}^-$ to $\mathbf{I}^-$ such that $\Pr(\mathbf{h}, \mathbf{i}^-, \mathbf{e}) > q$. The latter can be decided using an oracle for INFERENCE (marginalizing over the variables in $\mathbf{I}^+$) and we can decide whether there are at least k such joint value assignments $\mathbf{i}^-$ using an additional oracle for the threshold counting; note that we cannot 'merge' both oracles as the 'threshold' oracle machine must accept inputs for which the INFERENCE oracle answers 'no' as well as inputs for which the oracle answers 'yes'.

To prove $\mathsf{NP}^{\mathsf{PP}^{\mathsf{PP}}}$-hardness, we reduce MFE from E-MAJMAJSAT. We fix $q = \frac{1}{2}$ and $k = \frac{|\Omega(\mathbf{I}^-)|}{2}$. Let $(\phi, \mathbf{E}, \mathbf{M_1}, \mathbf{M_2})$ be an instance of E-MAJMAJSAT and let $\mathcal{B}_\phi$ be the network constructed from that instance as shown above. We claim the following: If and only if there exists a satisfying solution to $(\phi, \mathbf{E}, \mathbf{M_1}, \mathbf{M_2})$, there is a joint value assignment to $\mathbf{x_E}$ such that $\Pr(V_\phi = \text{TRUE}, \mathbf{x_E}, \mathbf{x_{M_2}}) > \frac{1}{2}$ for the majority of joint value assignments $\mathbf{x_{M_2}}$ to $\mathbf{X_{M_2}}$.

$\Rightarrow$ Let $(\phi, \mathbf{E}, \mathbf{M_1}, \mathbf{M_2})$ denote the satisfiable E-MAJMAJSAT instance. Note that in $\mathcal{B}_\phi$ any particular joint value assignment $\mathbf{x_E} \cup \mathbf{x_{M_1}} \cup \mathbf{x_{M_2}}$ to $\mathbf{X_E} \cup \mathbf{X_{M_1}} \cup \mathbf{X_{M_2}}$ yields $\Pr(V_\phi = \text{TRUE}, \mathbf{x_E}, \mathbf{x_{M_1}}, \mathbf{x_{M_2}}) = 1$, if and only if the corresponding truth assignment to $\mathbf{E} \cup \mathbf{M_1} \cup \mathbf{M_2}$ satisfies ϕ, and 0 otherwise. When marginalizing over $\mathbf{x_{M_1}}$ (and $\mathbf{Op}_\phi$) we thus have that a joint value assignment $\mathbf{x_E} \cup \mathbf{x_{M_2}}$ to $\mathbf{X_E} \cup \mathbf{X_{M_2}}$ yields $\Pr(V_\phi = \text{TRUE}, \mathbf{x_E}, \mathbf{x_{M_2}}) > \frac{1}{2}$ if and only if the majority of truth assignments

variables in the set $\mathbf{I}^+$, and deciding upon a threshold of the joint value assignments to the set $\mathbf{I}^-$. While the 'selecting' aspect is typically associated with problems in NP, 'inference' and 'threshold testing' are typically associated with problems in PP. Hence, as these three sub-problems work on top of each other, the complexity class that corresponds to this problem is $\mathsf{NP}^{\mathsf{PP}^{\mathsf{PP}}}$.

to $\mathbf{M_1}$, together with the given truth assignment to $\mathbf{E} \cup \mathbf{M_2}$, satisfy ϕ. Thus, given that this is the case for the majority of truth assignments to $\mathbf{M_2}$, we have that $\Pr(V_\phi = \text{TRUE}, \mathbf{x_E}, \mathbf{x_{M_2}}) > \frac{1}{2}$ for the majority of joint value assignments $\mathbf{x_{M_2}}$ to $\mathbf{X_{M_2}}$. We conclude that the corresponding instance $(\mathcal{B}_\phi, V_\phi = \text{TRUE}, \mathbf{X_E}, \mathbf{X_{M_1}} \cup \mathbf{Op}_\phi, \mathbf{X_{M_2}}, \frac{1}{2}, \frac{|\Omega(\mathbf{X_{M_2}})|}{2})$ of MFE is satisfiable.

$\Leftarrow$ Let $(\mathcal{B}_\phi, V_\phi = \text{TRUE}, \mathbf{X_E}, \mathbf{X_{M_1}} \cup \mathbf{Op}_\phi, \mathbf{X_{M_2}}, \frac{1}{2}, \frac{|\Omega(\mathbf{X_{M_2}})|}{2})$ be a satisfiable instance of MFE, i.e., there exists a joint value assignment $\mathbf{x_E}$ to $\mathbf{X_E}$ such that for the majority of joint value assignments $\mathbf{x_{M_2}}$ to $\mathbf{X_{M_2}}$, $\Pr(V_\phi = \text{TRUE}, \mathbf{x_E}, \mathbf{x_{M_2}}) > \frac{1}{2}$. For each of these assignments $\mathbf{x_{M_2}}$ to $\mathbf{X_{M_2}}$ $\Pr(V_\phi = \text{TRUE}, \mathbf{x_E}, \mathbf{x_{M_2}}) > \frac{1}{2}$ if and only if the majority of joint value assignments $\mathbf{x_{M_1}}$ to $\mathbf{X_{M_1}}$ satisfy ϕ.

Since the reduction can be done in polynomial time, this proves that MFE is $\mathsf{NP}^{\mathsf{PP}^{\mathsf{PP}}}$-complete. □

Given this intractability result, it may not be clear how MFE as mechanism for inference to the best explanation can scale up to task situations of real-world complexity. One approach may be to seek to approximate MFE, rather than to compute it exactly. Unfortunately, *approximating* MFE is NP-hard as well as computing it exactly. Given that MFE has MAP as a special case, it is intractable to infer an explanation that has a probability that is close to optimal [23] or that is similar to the most probable explanation [19]. By and of itself, for unconstrained domains, approximation of MFE does not buy tractability.

3.3 Parameterized Complexity

An alternative approach to ensure computational tractability is to study how the complexity of MFE depends on situational constraints, as described in Section 2. Building on known fixed parameter tractability results for MAP [18] and MSE [17], we will consider the parameters in Table 2:

Parameter	Description
Treewidth (t)	A measure on the network topology.
Cardinality (c)	The maximum number of values any variable can take.
#Relevants ($\lvert\mathbf{I}^+\rvert$)	The number of relevant intermediate variables that we marginalize over.
Skewedness (s)	A measure on the probability distribution [21], denoting the probability that for a given evidence set $\mathbf{E}$ with evidence $\mathbf{e}$ and explanation set $\mathbf{H}$, two random joint value assignments $\mathbf{i_1}$ and $\mathbf{i_2}$ to the irrelevant variables $\mathbf{I}^-$ would yield the same MFEs.

Table 2: Overview of parameters for MFE.

For $\mathbf{I}^+ = \varnothing$, MAP can be solved in $O(c^t \cdot n)$ for a network with n variables, and since $\Pr(X = x) = \sum_{y \in Y} \Pr(X = x, Y = y)$, we have that MAP can be solved in $O(c^t \cdot c^{|\mathbf{I}^+|} \cdot n)$. Note that even when we can tractably decide upon the most probable explanation for a given joint value assignment $\mathbf{i}$ to $\mathbf{I}^-$ (i.e., when c, t, and $|\mathbf{I}^+|$ are bounded) we still need to test at least $\lfloor \frac{c^{|\mathbf{I}^-|}}{2} \rfloor + 1$ joint value assignments to $|\mathbf{I}^-|$ to decide MFE exactly, even when $s = 1$. However, in that case we can tractably find an explanation that is *very likely* to be the MFE if s is close to 1. Consider the following algorithm for MFE (adapted from [17]):

Algorithm 1 Compute the Most Frugal Explanation

Sampled-MFE($\mathcal{B}, \mathbf{H}, \mathbf{I}^+, \mathbf{I}^-, \mathbf{e}, N$)

1: **for** $n = 1$ to N **do**
2: Choose $\mathbf{i} \in \mathbf{I}^-$ at random
3: Determine $\mathbf{h} = \text{argmax}_{\mathbf{h}} \Pr(\mathbf{H} = \mathbf{h}, \mathbf{i}, \mathbf{e})$
4: Collate the joint value assignments $\mathbf{h}$
5: **end for**
6: Decide upon the joint value assignment $\mathbf{h}_{\text{maj}}$ that was picked most often
7: **return** $\mathbf{h}_{\text{maj}}$

This randomized algorithm repeatedly picks a joint value assignment $\mathbf{i} \in \mathbf{I}^-$ at random, determines the most probable explanation, and at the end decides upon the explanation that was picked most often. Due to its stochastic nature, this algorithm is not guaranteed to give correct answers all the time. However, the error margin ϵ can be made sufficiently low by choosing N large enough. How large N needs to be for a particular ϵ depends on the probability of selecting a joint value assignment $\mathbf{i}$ for which $\mathbf{h}_{\text{maj}}$ is the most

probable explanation. This probability corresponds to the *skewedness* parameter s that was introduced in Table 2. If skewedness is high (e.g., $s = 0.85$), then N can be fairly low ($N \geq 10$) to ensure an error margin of less than $\epsilon = 0.1$. When determining the most probable explanation is easy—in particular, when the treewidth and cardinality of $\mathcal{B}$ are low and there are few relevant variables in the set $\mathbf{I}^+$—the algorithm thus runs in polynomial time. Since these parameters are independent of $\mathbf{i}$, MFE can in that case be decided in polynomial time, with a small possibility of error, when the skewedness is sufficiently large.

3.4 Discussion

We showed that MFE is intractable in general, yet can be tractably approximated (with a so-called expectation-approximation [21]) when the treewidth of the network is low, the cardinality of the variables is small, the number of relevant intermediate variables is low, *and* the probability distribution for a given explanation set $\mathbf{H}$, evidence $\mathbf{e}$ and relevant intermediate variables set $\mathbf{I}^+$ is skewed towards a single MFE explanation. We also know that MAP can be tractably computed exactly when the treewidth of the network is low, the cardinality of the variables is small, and either the MAP explanation has a high probability, or the total number of intermediate variables is low [18]. How do these constraints compare to each other?

For MAP, the constraint on the total number of intermediate variables seems implausible. In real-world knowledge structures there are many intermediate variables, and while only some of them may contribute to the MAP explanation, we still need to marginalize over all of them to compute MAP. Likewise, when there are many candidate hypotheses, it is not obvious that the most probable one has a high (i.e., close to 1) probability. Note that the actual fixed-parameter tractable algorithm [2, 18] has a running time with $\frac{\log p}{\log 1-p}$ in the exponent, where p denotes the probability of the MAP explanation. This exponent quickly grows with decreasing p, e.g., for $p = 0.1$ the exponent would be $\frac{\log 0.1}{\log 0.9} \approx 22$. Furthermore, treewidth and cardinality actually refer to the treewidth of the *reduced* junction tree, where observed variables are absorbed in the cliques. Given that we sample over the set $\mathbf{I}^-$ in MFE, but not in MAP, both parameters (treewidth and cardinality) will typically have much lower values in MFE as compared to MAP. That is, it is more plausible that these constraints are met in MFE than that they are met in MAP. Given the considerations in [9] it seems plausible that the *skewedness* constraint is met in many practical situations. Finally, the ALARM example suggests that the MFE results are fairly robust with respect to which variables are considered to be relevant.

4 Conclusion

In this paper we proposed Most Frugal Explanation (MFE) as an alternative to MAP. While this problem is intractable in general—it is $\mathsf{NP}^{\mathsf{PP}^{\mathsf{PP}}}$-complete, and thus even harder than MAP ($\mathsf{NP}^{\mathsf{PP}}$-complete [23]), Same-Decision Probability ($\mathsf{PP}^{\mathsf{PP}}$-complete [5]), and k-th MAP ($\mathsf{P}^{\mathsf{PP}^{\mathsf{PP}}}$-complete [20])—it can be tractably approximated under situational constraints that are arguably more realistic in large real-world applications than the constraints that are needed to render MAP (fixed-parameter) tractable. In future work we hope to explore the properties of MFE using simulations on (random) networks to investigate how MFE behaves under varying circumstances, like having a mismatch between intrinsic and expected relevant variables, having many competing explanations, and having varying degrees of 'skewedness' of the probability distribution.

References

[1] I. Beinlich, G. Suermondt, R. Chavez, and G. Cooper. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *2nd European Conference on AI and Medicine*, pages 247–256, 1989.

[2] H. L. Bodlaender, F. van den Eijkhof, and L. C. van der Gaag. On the complexity of the MPA problem in probabilistic networks. In *15th European Conference on Artificial Intelligence*, pages 675–679, 2002.

[3] A. S. Cofiño, R. Cano, C. Sordo, and J. M. Gutiérrez. Bayesian networks for probabilistic weather prediction. In *15th European Conference on Artificial Intelligence*, pages 695–699, 2002.

[4] A. Darwiche. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, Cambridge, UK, 2009.

[5] A. Darwiche and A. Choi. Same-decision probability: a confidence measure for threshold-based decisions under noisy sensors. In *5th European Workshop on Probabilistic Graphical Models*, 2010.

[6] R. Demirer, R.R. Mau, and C. Shenoy. Bayesian Networks: A decision tool to improve portfolio risk analysis. *Journal of Applied Finance*, 16(2):106–119, 2006.

[7] S. Dey and J. A. Stori. A Bayesian network approach to root cause diagnosis of process variations. *International Journal of Machine Tools and Manufacture*, 45(1):75–91, 2005.

[8] R. G. Downey and M. R. Fellows. *Parameterized Complexity*. Springer, Berlin, 1999.

[9] M.J. Druzdzel and H.J. Suermondt. Relevance in probabilistic models: "backyards" in a "small world". In *AAAI–1994 Fall Symposium Series: Relevance*, pages 60–63, 1994.

[10] B. Fitelson. A probabilistic theory of coherence. *Analysis*, 63:194–199, 2003.

[11] J. A. Fodor. *The Modularity of Mind*. MIT Press, Cambridge, MA, 1983.

[12] M. R. Garey and D. S. Johnson. *Computers and Intractability. A Guide to the Theory of NP-Completeness*. W.H. Freeman and Co., San Francisco, CA, 1979.

[13] P. L. Geenen, A. R. W. Elbers, L. C. van der Gaag, and W. L. A. van der Loeffen. Development of a probabilistic network for clinical detection of classical swine fever. In *11th Symposium of the International Society for Veterinary Epidemiology and Economics*, pages 667–669, 2006.

[14] D. H. Glass. Coherence measures and inference to the best explanation. *Synthese*, 157:275–296, 2007.

[15] D. H. Glass. Inference to the best explanation: does it track truth? *Synthese*, 185(3):411–427, 2012.

[16] E.T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[17] J. Kwisthout. Two new notions of abduction in Bayesian networks. In *22nd Benelux Conference on Artificial Intelligence*, pages 82–89, 2010.

[18] J. Kwisthout. Most probable explanations in Bayesian networks: Complexity and tractability. *International Journal of Approximate Reasoning*, 52(9):1452 – 1469, 2011.

[19] J. Kwisthout. Structure approximation of most probable explanations in Bayesian networks. In *12th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 340–351, 2013.

[20] J. Kwisthout, H. L. Bodlaender, and L. C. van der Gaag. The complexity of finding kth most probable explanations in probabilistic networks. In *37th International Conference on Current Trends in Theory and Practice of Computer Science*, pages 356–367, 2011.

[21] J. Kwisthout and I. van Rooij. Bridging the gap between theory and practice of approximate Bayesian inference. *Cognitive Systems Research*, 24:2–8, 2013.

[22] P. Lipton. *Inference to the Best Explanation*. Routledge, London, UK, 2004.

[23] J. D. Park and A. Darwiche. Complexity results and approximation settings for MAP explanations. *Journal of Artificial Intelligence Research*, 21:101–133, 2004.

[24] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, Palo Alto, CA, 1988.

[25] P. J. Sticha, D. M. Buede, and R. L. Rees. Bayesian model of the effect of personality in predicting decisionmaker behavior. In *4th Bayesian Modelling Applications Workshop*, 2006.

[26] J. B. Tenenbaum. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331:1279–1285, 2011.

[27] J. Torán. Complexity classes defined by counting quantifiers. *Journal of the ACM*, 38(3):752–773, 1991.

[28] D. Wilson and D. Sperber. Relevance theory. In *Handbook of Pragmatics*, pages 607–632. Blackwell, Oxford, UK, 2004.