Springer

Dear Author,

Here are the proofs of your article.

- You can submit your corrections **online**, via **e-mail** or by **fax**.
- For **online** submission please insert your corrections in the online correction form. Always indicate the line number to which the correction refers.
- You can also insert your corrections in the proof PDF and **email** the annotated PDF.
- For fax submission, please ensure that your corrections are clearly legible. Use a fine black pen and write the correction in the margin, not too close to the edge of the page.
- Remember to note the **journal title**, **article number**, and **your name** when sending your response via e-mail or fax.
- **Check** the metadata sheet to make sure that the header information, especially author names and the corresponding affiliations are correctly shown.
- **Check** the questions that may have arisen during copy editing and insert your answers/ corrections.
- **Check** that the text is complete and that all figures, tables and their legends are included. Also check the accuracy of special characters, equations, and electronic supplementary material if applicable. If necessary refer to the *Edited manuscript*.
- The publication of inaccurate data such as dosages and units can have serious consequences. Please take particular care that all such details are correct.
- Please **do not** make changes that involve only matters of style. We have generally introduced forms that follow the journal's style.
  Substantial changes in content, e.g., new results, corrected values, title and authorship are not allowed without the approval of the responsible editor. In such a case, please contact the Editorial Office and return his/her consent together with the proof.
- If we do not receive your corrections **within 48 hours**, we will send you a reminder.
- Your article will be published **Online First** approximately one week after receipt of your corrected proofs. This is the **official first publication** citable with the DOI. **Further changes are, therefore, not possible.**
- The **printed version** will follow in a forthcoming issue.

Please note

After online publication, subscribers (personal/institutional) to this journal will have access to the complete article via the DOI using the URL: http://dx.doi.org/[DOI].
If you would like to know when your article has been published online, take advantage of our free alert service. For registration and further information go to: http://www.link.springer.com.

Due to the electronic nature of the procedure, the manuscript and the original figures will only be returned to you on special request. When you return your corrections, please inform us if you would like to have these documents returned.

Metadata of the article that will be visualized in OnlineFirst

Please note: Images will appear in color online but will be printed in black and white.

ArticleTitle	Causal Learning with Occam's Razor	
Article Sub-Title		
Article CopyRight		
Journal Name	Studia Logica	
Corresponding Author	Family Name	**Schulte**
	Particle	
	Given Name	**Oliver**
	Suffix	
	Division	School of Computing Science
	Organization	Simon Fraser University
	Address	Burnaby, BC, V5A 1S6, Canada
	Phone	
	Fax	
	Email	oschulte@cs.sfu.ca
	URL	
	ORCID	http://orcid.org/0000-0002-2805-4313
Schedule	Received	
	Revised	
	Accepted	
Abstract	Occam's razor directs us to adopt the simplest hypothesis consistent with the evidence. Learning theory provides a precise definition of the inductive simplicity of a hypothesis for a given learning problem. This definition specifies a learning method that implements an inductive version of Occam's razor. As a case study, we apply Occam's inductive razor to causal learning. We consider two causal learning problems: learning a causal graph structure that presents global causal connections among a set of domain variables, and learning context-sensitive causal relationships that hold not globally, but only relative to a context. For causal graph learning, Occam's inductive razor directs us to adopt the model that explains the observed correlations with a minimum number of direct causal connections. For expanding a causal graph structure to include context-sensitive relationships, Occam's inductive razor directs us to adopt the expansion that explains the observed correlations with a minimum number of free parameters. This is equivalent to explaining the correlations with a minimum number of probabilistic logical rules. The paper provides a gentle introduction to the learning-theoretic definition of inductive simplicity and the application of Occam's razor for causal learning.	
Keywords (separated by '-')	Causal graph - Bayesian network - Formal learning theory - Mind change bounds - Probabilistic clauses	
Footnote Information		

Journal: 11225
Article: 9829

Author Query Form

Please ensure you fill out your response to the queries raised below and return this form along with your corrections

Dear Author

During the process of typesetting your article, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the 'Author's response' area provided below

Query	Details required	Author's response
1.	Please provide the 'Presented by editor' details.	
2.	Please check and confirm the corresponding author mail id is correctly identified and amend if necessary.	
3.	Please check and confirm the inserted citation of Figures 1, 5 is correct. If not, please suggest an alternative citation. Please note that figures should be cited in sequential order in the text.	

Oliver Schulte

Causal Learning with Occam's Razor

Abstract. Occam's razor directs us to adopt the simplest hypothesis consistent with the evidence. Learning theory provides a precise definition of the inductive simplicity of a hypothesis for a given learning problem. This definition specifies a learning method that implements an inductive version of Occam's razor. As a case study, we apply Occam's inductive razor to causal learning. We consider two causal learning problems: learning a causal graph structure that presents global causal connections among a set of domain variables, and learning context-sensitive causal relationships that hold not globally, but only relative to a context. For causal graph learning, Occam's inductive razor directs us to adopt the model that explains the observed correlations with a minimum number of direct causal connections. For expanding a causal graph structure to include context-sensitive relationships, Occam's inductive razor directs us to adopt the expansion that explains the observed correlations with a minimum number of free parameters. This is equivalent to explaining the correlations with a minimum number of probabilistic logical rules. The paper provides a gentle introduction to the learning-theoretic definition of inductive simplicity and the application of Occam's razor for causal learning.

Keywords: Causal graph, Bayesian network, Formal learning theory, Mind change bounds, Probabilistic clauses.

1. Introduction: Causal Learning and Inductive Simplicity

An inductive version of Occam's razor directs us to adopt the simplest hypothesis consistent with the evidence. This raises the question of how to define simplicity. In this paper we describe a learning-theoretic topological concept of inductive simplicity. As a case study, we apply the concept to an important and challenging inductive problem: inferring causal relationships from observed correlations among variables.

Learning Causal Graphs Causal graphs are a widely used model class for representing such relationships (also known as Bayesian networks [28, 29, 37]). We show that the inductive simplicity rank of a causal graph is measured exactly by the number of edges in the graph: The smaller this number is, the greater is the model's inductive simplicity. Thus the disconnected graph is simplest, and the fully connected graph the most complex. Occam's razor for learning causal graphs therefore directs us to adopt a graph that explains the observed correlations with a minimum number of direct causal links.

Studia Logica
https://doi.org/10.1007/s11225-018-9829-1

Journal: **11225** Article No.: **9829** ☐ TYPESET ☐ DISK ☐ LE ☐ CP Disp.:**2018/8/27** Pages: **33**

Learning Context-Sensitive Causal Relationships As a second study in causal learning, we examine context-sensitive causal relationships, that hold only in a given context. For example, different subpopulations may exhibit different causal connections. A common approach to learning context-sensitive causal relationships is to first learn a causal graph that describes general causal relationships, then expand the causal graph with a model of context-sensitive relationships. Context-sensitive causal relationships can be represented in graphs whose edges are labelled with specific values of variables. These graphical representations can be converted to probabilistic logical rules of the form "if the following causes obtain, then the effect obtains with probability p [27]. Therefore a method for learning context-sensitive relationships can also be used to learn probabilistic logical rules that represent statistical patterns in the data. We show that the inductive simplicity rank of a context-sensitive causal model is measured exactly by the number of free parameters in the model: The smaller this number is, the greater is the model's inductive simplicity. In this sense, the learning-theoretic concept agrees with widely used statistical model selection scores (e.g., BIC and AIC [4]), which also use the number of free parameters. While the methodological recommendations are similar, the justification is very different: A statistical model selection score penalizes complex models with many parameters to avoid overfitting, which improves out-of-sample generalization. The learning-theoretic justification is in terms of optimality criteria for a learning method: maximizing the simplicity of the selected models minimizes the worst-case number of times that a causal graph learner may have to change its model.

Learning Theory and Inductive Simplicity: General Concepts The paper presents a gentle introduction to the learning-theoretic definitions and theorems that lead to an inductive version of Occam's razor. We begin with the concept of a learning problem. A learning problem has three components: (i) A space of alternative hypotheses, (ii) a set of possible evidence items that learning uses to decide among different hypotheses, and (iii) a definition of which hypotheses are consistent with which evidence items. We discuss how causal learning can be framed as a learning problem in this sense.

We present a general definition of Occam's razor for a given learning problem with a finite space of alternative hypotheses. Finite hypothesis spaces are sufficient for our causal learning case study, and the finiteness assumption simplifies technical definitions and theorems. Learning theorists have developed the concept of inductive simplicity more generally for infinite hypothesis spaces [24]. In a finite hypothesis space, inductive simplicity is de-

Journal: 11225 Article No.: 9829 TYPESET DISK LE CP Disp.:2018/8/27 Pages: 33

fined in terms of the branching structure of the hypothesis space as follows. The inductive simplicity rank of a hypothesis H is the length of the longest chain of nested hypotheses, starting with the hypothesis H. One hypothesis H_1 is nested within another H_2 if all evidence items that are consistent with H_1 are also consistent with H_2, but not vice versa. This simplicity concept has several noteworthy features.

1. Although in practice a hypothesis is described within a hypothesis language, its inductive simplicity ranking is entirely a function of its observational content. Inductive simplicity therefore is completely *independent of the choice of language* for describing hypotheses.
2. Inductive simplicity depends on the learning context; it is *relative to the entire hypothesis space*. A hypothesis does not have an intrinsic inductive simplicity rank, but only a rank relative to alternatives under consideration. If the set of alternatives under consideration changes, the inductive complexity of a hypothesis can change as well [32].
3. A precise definition of inductive simplicity leads to a precise definition of Occam's razor for inductive inference. The inductive version of Occam's razor that corresponds to the topological definition can be rigorously justified in terms of learning performance: We show that Occam's inductive razor is the only learning method that achieves reliable, steady, and fast convergence to a correct hypothesis.

Applications of Inductive Simplicity This theorem provides a justification of Occam's razor in terms of learning performance that is independent of whether the inferences underwritten by Occam's inductive razor are intuitively plausible. Nonetheless, it is interesting to ask whether in cases of interest, Occam's inductive razor directs us towards plausible inferences. The answer is generally yes: We have already described the results for causal learning. Other case studies include Goodman's New Riddle of Induction, where Occam's inductive razor selects "all emeralds are green" over "all emeralds are grue" [33], and inferring conservation laws in particle physics, where the Occam method rediscovers the laws in the important Standard Model of particle physics [34]. This paper shows, using the example of causal learning, how learning theory can be applied to substantive real inductive problems.

Paper Organization We introduce the concept of a learning problem using an abstract toy example. Then we discuss how causal graph learning can be framed as a learning problem. The toy example serves to introduce performance criteria for a learning method. We give the general definition of

Journal: 11225 Article No.: 9829 TYPESET DISK LE CP Disp.:2018/8/27 Pages: 33

inductive simplicity for a finite hypothesis space, then prove the relationship between inductive simplicity and optimal learning performance. Causal learning illustrates these results in a concrete setting. We first apply the general theory to the problem of learning causal graph structures, then to the problem of learning context-senstive causal relationships.

2. Definition of Learning Problems

We first introduce general concepts from learning theory and present some general results concerning inductive simplicity. These concepts have been used with different nomenclature depending on the intended application [15,16,25]. We present a framework that is as simple as possible, while expressive enough to model causal learning. Our nomenclature is intended to be appropriate for discussing learning in general (rather than, e.g., language learning problems, in particular). After introducing the general concepts, we present in detail a model for applying them to causal graph learning.

2.1. Learning Problems

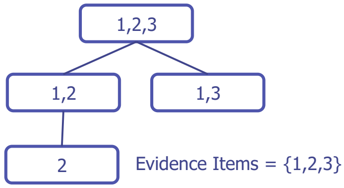
Learning begins with observations. The first part of a learning problem is therefore a set of **evidence items**. For learning-theoretic analysis, a sequence of evidence items is the basis for drawing inductive conclusions; more formally, a sequence of evidence items is the input to a learning algorithm.

Examples In modelling high-energy physics, an evidence item may be a single reaction that physicists report as having observed in a particle accelerator [34]. In cognitive psychology, an evidence item may be a reaction by an experimental subject to an input stimulus that is to be explained by a model of cognitive architecture [12]. In inductive problems often discussed in philosophy, an evidence item may be the color of a swan, the color of an emerald, or the observation of a sunrise on a given day. In language learning, an evidence item may be a single sentence heard by a child [13].

The second part of a learning problem is a space of **hypotheses**, possible explanations/models of the evidence. This space represents the background knowledge on which learning is based.

Examples In high-energy physics, a set of conservation laws defines which particle reactions are possible. In cognitive architectures, a specification of cognitive models and their interconnections can explain behavior. In language learning, a grammar defines a set of well-formed sentences.

Journal: 11225 Article No.: 9829 TYPESET DISK LE CP Disp.:2018/8/27 Pages: 33



Conjectures of a learner as evidence items accumulate

Evidence Item	2	1	3
Conjecture	{2}	{1,2} 1st mind change	{1,2,3} 2nd mind change
Simplest Alternatives	unique	unique	unique

Figure 1. Left: A set of abstract evidence items (three) and a simple abstract hypothesis space. A hypothesis is represented as the set of evidence items that are consistent with it. Links between hypotheses correspond to set inclusion. Right: A learning sequence for this learning problem. Three evidence items are presented in sequence (2,1,3). After each new evidence item, the output of the learner is shown

The third part of a learning problem is a specification of which evidence items are **consistent** with which hypotheses. This specifies the prediction made by a hypothesis, or its empirical content.

Examples In high-energy physics, a reaction is consistent with a set of conservation laws if it conserves all quantities posited by the conservation laws. A cognitive computational architecture is consistent with a subject's reaction to an input stimulus if the reaction is the same as the output computed by the hypothesized system. In language learning, a sentence is consistent with a grammar if it can be generated by the grammar.

In sum, we have the following definition of a learning problem.

DEFINITION 1. *A learning problem consists of the following three components.*

Evidence Items *A countable set of evidence items* E.

Hypotheses *A finite set of hypotheses* $\mathcal{H}$.

Consistency *A consistency relation between a hypothesis and an evidence item that specifies which hypotheses are consistent with which evidence items. Thus the consistency relation is a subset of the Cartesian product* $\mathrm{E} \times \mathcal{H}$.

Even more general definitions of a learning problem are possible (see e.g., [10]). The definition given is equivalent to the concept of a language learning problem used in formal learning theory [15]. Learning-theoretic analysis applies to infinite hypothesis sets as well as finite ones. We assume finiteness only to simplify the definitions (Figure 1).

Journal: 11225 Article No.: 9829 TYPESET DISK LE CP Disp.:2018/8/27 Pages: 33

2.2. Data Streams

A **data stream** is an *infinite* sequence of evidence items (or #). The content of a data stream is the set of evidence items that appear in the sequence. A hypothesis is **correct** for a data stream if the evidence items in the data stream comprise all and only those consistent with the hypothesis. A data stream is **possible** for a learning problem if some alternative hypothesis is correct for it. In other words, we assume that every possible *complete* sequence of evidence items can be explained by some hypothesis under consideration. It is however possible that for a finite evidence sequence, no hypothesis is consistent with all and only the evidence items observed. For instance in the toy problem above, if the first evidence sequence is (1), there is no hypothesis in the space is consistent with item 1 and only 1.

An important issue for both the philosophy and the practice of science is that different hypotheses may be **empirically equivalent**, meaning that they are correct for exactly the same data streams, and hence consistent with exactly the same evidence items. This occurs in many practical hypothesis spaces, because hypotheses are described using a hypothesis language, and just like natural language, a hypothesis language typically allows us to express the same content in different ways. For example one set of conservation laws is empirically equivalent to another if they both span the same linear subspace [34]. We discuss empirical equivalence for causal graphs in Section 3.3 below. This phenomenon has been described as *global underdetermination* [16]. In statistical terminology, it is referred to as the identifiability problem, where even an infinitely large sample does not entail a uniquely correct model and/or parameter values for the model. Global underdetermination raises the problem of defining which hypothesis is the correct one for a complete set of observations when there are empirically equivalent alternatives. A simple approach to global underdetermination is to have learners output an equivalence class of hypotheses. In effect, this changes the hypothesis space so that a hypothesis in the new space is an equivalence class of hypotheses in the original space. A benefit of focusing on empirical equivalence classes is that in many problems, determining which hypotheses are empirically equivalent is a problem of independent interest. In the following we assume that learners output equivalence classes of hypotheses. As we show in the causal graph example, an equivalence class of original hypotheses can be often be described compactly.

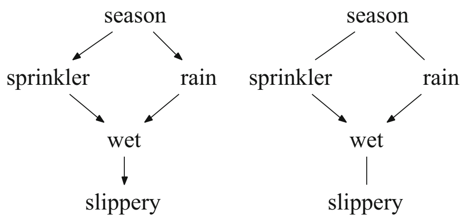


Figure 2. The sprinkler network (left) and its pattern (right). Sprinkler and Rain share Season as a common cause, and Wetness as a common effect. The wetness of the pavement is a direct cause of its being slippery

3. The Causal Graph Structure Learning Problem

We present a view of causal graph learning as a learning problem in the sense of Definition 1, following previous work [18,36]. In this paper, we use the term "causal graph" as essentially synonymous with "Bayesian network". The approaches for learning Bayesian networks are similar to those for causal graphs. The main difference is in the interpretation: causal graphs are viewed as representing the effects of actions or interventions. For further discussion see [29,37].

3.1. Hypothesis Space

A **causal graph structure** is a directed acyclic graph (DAG). The hypothesis space is the set of causal graphs that share a common fixed set of nodes $\mathbf{V} = \{X_1, X_2, \ldots, X_n\}$. The nodes are also called its **variables**. Every node X has a possible **domain of values**. In this paper we consider only discrete variables with finite domains. We write $X = x$ for an assignment of value to variable X, and use boldface vector notation such as $\mathbf{X} = \mathbf{x}$ for a set of variables with an assignment of values. Figure 2(left) shows a causal graph from [29, p. 15]. The graph represents causal relationships in an intuitive visual way, where parents are direct causes of their children. For example, a sprinkler running directly causes the pavement to be wet, and wetness directly causes the pavement to be slippery. Indirect causes can be read off by following the causal arrows. For example, a sprinkler running is an indirect cause of the pavement being slippery.

3.2. Evidence Items

An evidence item is a **conditional dependence statement** of the form

$$\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} | \mathbf{Z} \tag{1}$$

where $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ are disjoint sets of variables. Intuitively, a dependence statement can be read as "the variables in the set $\mathbf{X}$ are relevant to the variables in the set $\mathbf{Y}$, given an assignment of values to the variables in the set $\mathbf{Z}$."

Examples The most common interpretation of a dependence statement is as expressing a *probabilistic dependence*. Probabilistic dependencies are defined with respect to a **joint distribution** over the nodes in the graph. A joint distribution specifies a probability

$$P(\mathbf{V} = \mathbf{v})$$

for each complete assignment of values to the variables. From a joint distribution we obtain *marginal distributions* over any subset $\mathbf{X}$ of variables

$$P(\mathbf{Y} = \mathbf{y}) \equiv \sum_{\mathbf{x}} P(\mathbf{Y} = \mathbf{y}, \mathbf{X} = \mathbf{x})$$

where $\mathbf{X}$ is the set of variables $\mathbf{V} - \mathbf{Y}$ other than $\mathbf{Y}$.

A joint distribution also specifies *conditional distributions* via the definition

$$P(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}) \equiv \frac{P(\mathbf{Y} = \mathbf{y}, \mathbf{X} = \mathbf{x})}{P(\mathbf{X} = \mathbf{x})}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are disjoint, and $P(\mathbf{X} = \mathbf{x}) > 0$. In what follows we assume that all joint probabilities are positive so that conditional probabilities are well defined. For a discussion of learning causal graphs with 0 joint probabilities, which may occur with deterministic causal relationships, see [23].

The meaning of a conditional dependence statement can be defined in terms of a **probabilistic inequality** as follows:

$$\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} | \mathbf{Z} \equiv \exists \mathbf{x}, \mathbf{y}, \mathbf{z}. P(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) \neq P(\mathbf{Y} = \mathbf{y} | \mathbf{Z} = \mathbf{z}). \quad (2)$$

For example, in the graph of Figure 2, a joint distribution may specify a probability of 0.1 that it is summer, the sprinker is on, and all other variables are simultaneously true:

$$\begin{aligned} P(&\texttt{season} = \mathit{summer}, \texttt{sprinkler} = \mathit{on}, \texttt{rain} = T, \\ &\texttt{wet} = T, \texttt{slippery} = T) = 0.1. \end{aligned}$$

The joint distribution may entail the following claim: the probability that the pavement is wet is affected by the probability that the season is summer, even given that the sprinkler is on:

$$\begin{aligned} &P(\texttt{wet} = T | \texttt{season} = \mathit{summer}, \texttt{sprinkler} = \mathit{on}) \\ &\neq P(\texttt{wet} = T | \texttt{sprinkler} = \mathit{on}) \end{aligned}$$

which witnesses the dependence assertion that

$$\texttt{wet} \not\perp\!\!\!\perp \texttt{season}|\texttt{sprinkler}.$$

3.3. Consistency of a Causal Graph with Dependency Statements

For parametric models in general, a model structure is consistent with a set of evidence items if there exists an assignment of values to the model parameters that entails the observed evidence. This logic can be applied to causal graphs as follows. The parameters of a causal graph are conditional probabilities that specify, for each assignment of a value to a node, and for each assignment of values to the node's parents, the probability of the child node value, given the parent values. The combination of (graph structure + parameters) defines a joint distribution P over the nodes in the graph. Such a combination is consistent with a set of observed dependencies if the dependencies are entailed by the joint distribution P. A causal graph structure G by itself is then **consistent** with a set of dependence statements of the form 1 if *there exists* a parameter assignment for the graph G that is consistent with the observed dependencies. Figure 3 illustrates the logic of this definition.

The combination of (graph structure + parameters) defines a joint distribution via the product formula: multiply together all the conditional probabilities defined by each child-parent value assignment. For instance, for the graph in Figure 2(left), the joint probability that all variables above would be defined by the product

$$\begin{aligned}
&P(\texttt{season} = \mathit{summer}, \texttt{sprinkler} = \mathit{on}, \texttt{rain} = T, \texttt{wet} = T, \texttt{slippery} = T)\\
&\quad = P(\texttt{season} = \mathit{summer}) \times P(\texttt{sprinkler} = \mathit{on}|\texttt{season} = \mathit{summer})\\
&\qquad \times P(\texttt{rain} = T|\texttt{season} = \mathit{summer})\\
&\qquad \times P(\texttt{wet} = T|\texttt{sprinkler} = \mathit{on}, \texttt{rain} = T) \times P(\texttt{slippery} = T|\texttt{wet} = T)
\end{aligned}$$

where the conditional probabilities that appear in this expression are specified as parameter values for the causal graph.

Causal graphs allow us to compute joint probabilities that describe the effects of interventions. For example, suppose we turn on the sprinkler given the causal structure of Figure 2. The joint distribution given this intervention can be computed by removing the edge $\texttt{season} \rightarrow \texttt{sprinkler}$—which represents that the value $\texttt{sprinkler}$ has been determined exogenously outside the system—and using the conditional probabilities in the resulting truncated graph to compute joint probabilities. For more examples and details please see [29].

Journal: 11225 Article No.: 9829 TYPESET DISK LE CP Disp.:2018/8/27 Pages: 33

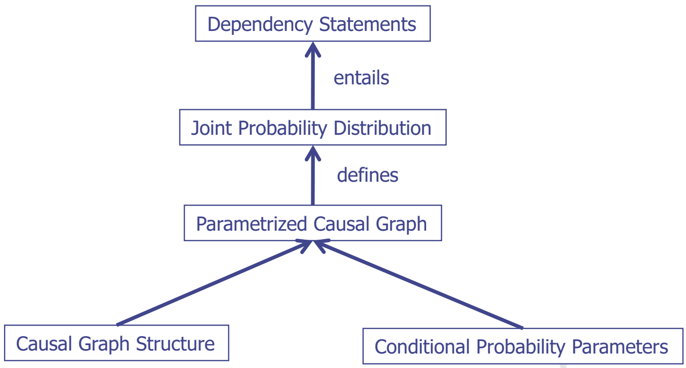


Figure 3. Consistency of causal graph structure with observed dependency statements. A parametrized causal graph is consistent with observed dependencies if its joint distribution entails the dependencies. A causal graph structure is consistent with observed dependencies if there is a parametrization of the structure that is consistent with the dependencies

Graph Dependencies and d-Separation It may appear that determining whether a graph structure G is consistent with a given set of dependencies is difficult because searching through the set of all possible parameter assignments is difficult. However, causal graph theory has developed a graph-based criterion, known as *d-separation*, that facilitates an efficient check whether a graph structure is consistent with given dependencies based on the links only, without reference to parameter values. For readers unfamiliar with this criterion, we provide a review in the appendix. In terms of d-separation, a causal graph structure is consistent with observed dependencies if it is an I-map of the given dependencies. This means that if any node set $\mathbf{X}$ is d-separated from another $\mathbf{Y}$ by the nodes $\mathbf{Z}$, then the dependency statement $\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}$ is not in the given dependencies.

The d-separation criterion also makes it possible to provide a graphical characterization of causal graphs that are empirically equivalent, that is, that are consistent with exactly the same set of dependency statements. Two nodes X, Y are **adjacent** in a graph G if G contains an edge $X \rightarrow Y$ or $Y \rightarrow X$. The **pattern** of DAG G is the partially directed graph that has the same adjacencies as G, and contains an arrowhead $X \rightarrow Y$ if and only if G contains a triple $X \rightarrow Y \leftarrow Z$ where X and Z are not adjacent. Figure 2 (right) illustrates the concept. Verma and Pearl proved that two graphs G_1 and G_2 are consistent with the same dependency statements if and only if they lead to the same pattern [40, Thm. 1]. Thus we can use a pattern as a syntactic representation of an empirical equivalence class of graphs.

This completes our description of causal graph learning as an instance of a learning problem: Evidence items are conditional dependence statements, hypotheses are patterns (equivalence classes of causal graphs), and a hypothesis is consistent with a set of dependence statements if the dependencies are entailed by applying d-separation to the graph. We next discuss the assumptions and limitations of this model (see also the predecessor paper by Schulte et al. [36]).

4. Discussion: Assumptions and Limitations

The key assumptions are characteristic of all learning-theoretic applications, concerning the correctness and completeness of the available evidence, as well as the adequacy of the hypotheses under consideration. We discuss how our assumptions relate to previous work on learning causal graphs.

Approaches to Learning Causal Graphs There are two well established general approaches to learning a causal graph, or Bayesian network structure. *Constraint-based* (CB) methods employ a statistical test to detect conditional (in)dependencies given a data sample, and then compute a graph structure that fits the (in)dependencies [5,37]. *Score-based* methods search for models that maximize a model selection score [14]. Statistical model selection scores typically balance model fit—measured by the likelihood of the data under the model—against model complexity—often measured by the number of model parameters. For example, the AIC score subtracts the number of model parameters from the data likelihood. A key difference is that a statistical model selection score measures the fit of a model to data as a continuous quantity that comes in degrees. As the name "constraint" suggest, CB methods evaluate graphs in a Boolean fashion against the data: a graph either satisfies the observed (in)dependencies or not. Our paper falls in the CB paradigm, because CB methods are based on a discrete notion of consistency that allows us to apply learning-theoretic analysis. An alternative recent approach to applying learning theory to statistical problem is to develop the theory directly for probabilistic models [10,19], where consistency is taken to be a matter of degree.

Obtaining Dependence Statements from Statistical Tests A dependence statement existentially quantifies over possible values of variables: it asserts that there exist specific values $\mathbf{x}, \mathbf{y}, \mathbf{z}$ such that if the conditioning variable set $\mathbf{Z}$ takes on the values $\mathbf{z}$, then the values $\mathbf{x}$ for $\mathbf{X}$ do not affect the probability that $\mathbf{Y}$ takes on value $\mathbf{y}$. This existentially quantified statement can be derived from basic unquantified probabilistic inequalities of the form

$$P(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) \neq P(\mathbf{Y} = \mathbf{y}|\mathbf{Z} = \mathbf{z}) \tag{3}$$

for fixed values $\mathbf{x}, \mathbf{y}, \mathbf{z}$. Such basic probabilistic inequalities can be ascertained using statistical tests. In practice, a Bayesian network structure learner obtains a random sample $\mathbf{d}$ drawn from the data generating joint distribution over the variables $\mathbf{V}$, and applies a suitable statistical criterion to decide if a dependency $X \not\perp\!\!\!\perp Y|\mathbf{S}$ holds [37], [39, Sec. 4]. Many constraint-based approaches to learning causal graphs use a statistical test as follows: given a query "Does $X \not\perp\!\!\!\perp Y|\mathbf{S}$ hold?", the system answers "yes" if the test rejects the hypothesis $X \perp\!\!\!\perp Y|\mathbf{S}$, and "no" otherwise. The assumption that this procedure yields correct results is called the assumption of valid statistical testing [5, Section 6.2]. Compared to this assumption, our model of learning from conditional dependencies (positive data) is more realistic in two respects. First, the model assumes only that *dependency information* is available, but does not rely on *independence* data. In fact, many statisticians hold that no independence conclusion should be drawn when a statistical significance test fails to reject an independence hypothesis [11], because there is no bound on the probability of falsely accepting an independence hypothesis after a failure to reject.[1] Second, the dependency learning model does not assume that the dependency information is supplied by an oracle all at once, but explicitly considers learning in a setting where more information becomes available as the sample size increases. Our model still assumes that a statistically significant correlation does not disappear as the sample size increases. The extent to which this assumption is plausible depends on the testing strategy that extracts correlations from the given samples. The most common approach in constraint-based methods is to employ a fixed conservative significance level (e.g., $\alpha = 0.1\%$ [37, Ch. 5], [6, 39]) for any sample size; with this kind of testing strategy, our assumption that the store of observed correlations grows monotonically is quite plausible.

Complete Data Enumeration This observation supports the standard assumption in learning theory that a complete infinite data stream enumerates exactly the true evidence items; in our case, the dependencies that are true in a domain.

Adequacy of Hypothesis Space In addition, the learning model model assumes that these true domain dependencies can be represented exactly by a causal graph. In causal graph theory, this assumption is known as *faithfulness*. For

[1] Schulte et al. [36] describe the Occam method for learning from independencies as evidence items.

Journal: 11225 Article No.: 9829 TYPESET DISK LE CP Disp.:2018/8/27 Pages: 33

discussions of the faithfulness assumption, see [29, Ch. 2.4], [41], [38, Ch. 8.1].

As the causal graph example shows, defining the three components of a learning problem for a realistic scenario can be a substantial task. Once this is accomplished, we can apply powerful results from general learning theory to determine optimal learning algorithms for the learning problem. We review some of these general results and apply them to causal graph structure learning in the next section.

5. Learning Methods and Optimal Learning

We generically denote learners by upper-case Greek letters such as Ψ, Φ. Intuitively, a **learner** takes as input a sequence of evidence items—called the **data sequence**, and produces as output a member of the hypothesis space. It is often convenient to slightly generalize this learning model: First, we allow the data sequence to contain the special non-evidence symbol $\{\#\}$ to model pauses in data presentation. Second, we allow the learner to output $\{?\}$, where ? corresponds to the vacuous output "no guess".

Figure 1 illustrates these concepts. There are three evidence sequences,

$$(2), (2, 1), (2, 1, 3).$$

The learner maps these to a sequence of conjectures outputs

$$\{2\}, \{1, 2\}, \{1, 2, 3\}.$$

Each output is consistent with exactly the observed items.

5.1. Performance Criteria for Learning

Learning theorists have studied a number of performance criteria for successful learning (often referred to as identification criteria [3]). In this paper we consider three: (i) reliable identification of a correct hypothesis in the limit, (ii) steady identification of a correct hypothesis, i.e., minimizing hypothesis changes, and (iii) fast identification of a correct hypothesis, i.e., minimizing time to convergence.

We say that a learner **identifies** a correct hypothesis on an infinite data stream if after some finite time, the learner outputs a hypothesis that is correct for the entire data stream. Identification requires induction in the sense of going beyond the data: although the learner typically has not observed all evidence items that may appear, the hypothesis it selects does concern future evidence items. A learner **reliably identifies** the correct hypothesis in

Journal: 11225 Article No.: 9829 TYPESET DISK LE CP Disp.:2018/8/27 Pages: 33

Occam Learner

Evidence Item	1	2	3
Conjecture	?	{1,2} no mind change	{1,2,3} 1st mind change
Simplest Alternatives	{1,2}, {1,3}	unique	

Not strongly mind-change optimal

Evidence Item	1	2	3
Conjecture	{1,3}	{1,2} 1st mind change	{1,2,3} 2nd mind change
Simplest Alternatives	{1,2}, {1,3}	unique	

Occam Learner

Evidence Item	3	1	2
Conjecture	{1,3}	{1,3} no mind change	{1,2,3} 1st mind change
Simplest Alternatives	unique	unique	unique

Not efficient for convergence time

Evidence Item	2	1	3
Conjecture	?	{1,2} no mind change	{1,2,3} 1st mind change
Simplest Alternatives	{2} unique	unique	unique

Figure 4. Examples of different learning methods. After each evidence is received, a learner outputs a hypothesis

a learning problem if it identifies a correct hypothesis on every possible data stream. To illustrate these concepts in the example of Figure 1, consider the infinite data stream

$$2, 1, 3, \#, \#, \ldots.$$

The set of evidence items for this data stream is $\{1, 2, 3\}$, so the learner converges to a correct hypothesis after three evidence items have been observed. Figure 4 provides more examples of evidence and hypothesis sequences.

Identifiability requires only eventual convergence. We can compare the performance of different learners with respect to convergence speed by using the decision-theoretic criterion of weak dominance: A learner Ψ is **faster than** a learner Φ on a data stream if Ψ converges to a hypothesis before Φ does. A learner Ψ is **uniformly faster** than Φ if Ψ is faster than Φ on some possible data stream, and Ψ is at least as fast as Φ on every possible data stream.

Our final criterion is steadiness of convergence: minimizing the number of times that a learner changes its hypothesis before convergence. A learner Ψ **changes its mind** at some nonempty finite sequence of evidence items $\langle e_1, \ldots, e_m, e_{m+1}\rangle$ if the output of Ψ after observing evidence items $\langle e_1, \ldots, e_m\rangle$ is not vacuous (i.e., is not ?) and differs from the output of Ψ after observing evidence items $\langle e_1, \ldots, e_m, e_{m+1}\rangle$ [15, Ch. 12.2], [18,31]. For any possible data stream, we can count the number of mind changes that a learner undergoes on that data stream. This is a finite number assuming that the learner eventually settles on a correct hypothesis.

Journal: 11225 Article No.: 9829 TYPESET DISK LE CP Disp.:2018/8/27 Pages: 33

We can assess the performance of a learner with respect to mind changes by using the decision-theoretic minimax criterion, which considers worst-case performance.[2] Say that a learner Ψ reliably identifies a correct hypothesis with at most k mind changes if it is reliable and changes its hypotheses at most k times on every possible data stream. In the toy problem of Figure 1, a learner can reliably identify a correct hypothesis with at most two mind changes. This is illustrated by the Occam learner shown in Figure 1 and in the left two examples of Figure 4. The Occam learner outputs a hypothesis that accommodates a minimum number of evidence items, if there is a unique such hypothesis. If there is not, the learner outputs?.

A natural alternative is to express uncertainty by a set (disjunction) of hypotheses as possible outputs rather than?. The meaning of an output is then that the true hypothesis is a member of the set. For example, a learner could output a set of languages, or a set of causal graphs. A mind change is then said to occur at $\langle e_1, \ldots, e_m, e_{m+1}\rangle$ if the set of hypotheses of Ψ after observing evidence items $\langle e_1, \ldots, e_m\rangle$ is not entailed by the output of Ψ after observing evidence items $\langle e_1, \ldots, e_m, e_{m+1}\rangle$. (In symbols, $\Psi(\langle e_1, \ldots, e_m\rangle) \not\supseteq \Psi(\langle e_1, \ldots, e_m, , e_{m+1}\rangle)$.) The results about problem complexity and learning optimality remain the same, so we keep with the simpler traditional? representation in formal learning theory [15, Ch. 12.2]. A global mind change bound is not quite good enough in many problems, because a learner may fail to take advantage of a lucky evidence sequence that makes it possible to learn with fewer mind changes than in the worst case. The top half of Figure 4 illustrates this possibility. After evidence item 1 is observed, it is possible to succeed with at most one further mind change: wait with? until further evidence decides between the hypotheses $\{1, 2\}$ or $\{1, 3\}$. Then at most one more mind change is required if the correct hypothesis turns out to be $\{1, 2, 3\}$. The learner in the top right box outputs $\{1, 3\}$ right away and undergoes two mind changes in case the item 2 is observed before 3. Since the problem requires two mind changes in the worst case (see Figure 1), this behavior is consistent with a global mind change bound of two. We therefore refine the mind change criterion as follows [36]. Say that a learning problem can be solved with at most k mind changes starting with evidence sequence $\langle e_1, \ldots, e_m\rangle$ if for every data stream extending the evidence sequence, there is a learner that reliably identifies a correct hypothesis, and uses at most k mind changes, where mind changes are counted starting at time m. That is, mind changes are counted with the output for $\langle e_1, \ldots, e_m\rangle$ as the initial hypothesis. A learner Ψ is **strongly mind change optimal** if for every finite

[2] Schulte shows that applying admissibility to mind changes is not fruitful [33].

Journal: 11225 Article No.: 9829 TYPESET DISK LE CP Disp.:2018/8/27 Pages: 33

evidence sequence $\langle e_1, \ldots, e_m \rangle$, the learner Ψ solves the learning problem with the best possible mind change bound starting with $\langle e_1, \ldots, e_m \rangle$.

5.2. Optimal Learning

Performance criteria can be used to select learning methods. We can picture a performance criterion as a filter that is applied to learning methods: applying multiple performance criteria is like applying successive filters to learning. The criteria we examine in the rest of the paper are as follows.

First, filter out all methods that do not reliably identify a correct hypothesis. Second, eliminate among the remaining ones those methods that are uniformly slower than another remaining method. Third, among these, filter out those that are not strongly mind change optimal. We refer to a method that meets these criteria simply as **optimal**. In decision-theoretic terms, the successive filters correspond to a lexicographic ordering of performance criteria: identifiability first, convergence time second, mind change optimality third. Other combinations of performance criteria lead to different concepts of optimality; the optimality concept of this paper is the most fruitful for applications [18,33].

The lexicographic concept is surprisingly powerful: we will show next that in many problems, including those with finite hypothesis spaces, there is only one optimal method. This optimal method guides us towards interesting and plausible inductive inferences in many domains. Determining the hypotheses selected by the optimal method is an investigation that leads to substantive insights into the methodological structure of a learning problem. We outline a general result that assists a theorist in determining the conjecture of the optimal method. Then we show how to apply this to causal graph search.

6. Optimal Learning and Inductive Simplicity

Our goals in this section are to prove the uniqueness of an optimal learner and to characterize optimal inferences in terms of the structure of the learning problem. This structure can be described in terms of a topological ranking of hypotheses in a given hypothesis space [24].

Definition 2. *Let $\mathcal{H}$ be a hypothesis space and H be a hypothesis in $\mathcal{H}$. We write $H \subset H'$ to denote that the evidence items consistent with hypothesis H are a proper subset of those consistent with H'. An **inclusion chain** of length k starting with H is a sequence of the form*

$$H \subset H_1 \subset \cdots \subset H_i \subset \cdots \subset H_k$$

*where each hypothesis in the chain is contained in the hypothesis space $\mathcal{H}$. The **inclusion depth** of H is the maximum length of an inclusion chain starting with H.*

Kevin Kelly has developed the view that inclusion depth, or closely related concepts, can be viewed as a simplicity ranking of hypotheses [17]. Accordingly, we define the **inductive simplicity rank** of a hypothesis H in a learning problem as the inclusion depth of H. Thus defined, simplicity increases with inclusion depth. Occam's razor directs us to adopt a maximally simple hypothesis that is consistent with the evidence. For a learning problem, this leads to the following definition of the **Occam learner**

$$\Psi_{occam}(\langle e_1, \ldots, e_m \rangle)$$
$$= \begin{cases} ? & \text{if there is no uniquely simplest hypothesis consistent with } \langle e_1, \ldots, e_m \rangle \\ H & \text{if } H \text{ is the uniquely simplest hypothesis consistent with } \langle e_1, \ldots, e_m \rangle \end{cases}$$

where $\Psi_{occam}(\langle e_1, \ldots, e_m \rangle)$ is the output of the Occam learner after receiving the evidence sequence $\langle e_1, \ldots, e_m \rangle$. This definition assumes that there are no empirically equivalent hypotheses; otherwise it should be modified so that the Occam learner outputs the maximally simple equivalence class if there is one.

The next proposition shows a strong connection between inductive simplicity and required mind changes: The worst case number of mind changes required is exactly the maximum of the inductive simplicity ranks in the hypothesis space.

Proposition 3. Luo and Schulte [24] *For any learning problem with a finite hypothesis space, there is a learner that reliably identifies a correct hypothesis with at most k mind changes $\iff$ the maximum inductive simplicity rank of any hypothesis is k.*

Proof. ($\Leftarrow$) Suppose that the maximum inductive simplicity rank of any hypothesis is k. Consider a mind change by the Occam learner Ψ_{occam} on a sequence of evidence items $\langle e_1, \ldots, e_m, e_{m+1} \rangle$. Let H be the output of Ψ_{occam} on the previous sequence $\langle e_1, \ldots, e_m \rangle$. Since a mind change occurred at stage $m + 1$, the hypothesis H is not vacuous. By the definition of the Occam learner, H is therefore the uniquely most simple consistent with the evidence $\langle e_1, \ldots, e_m \rangle$. Thus any other hypothesis consistent with the further evidence $\langle e_1, \ldots, e_m, e_{m+1} \rangle$ must have lower simplicity rank than H. Therefore any time that the Occam learner changes its mind, the maximum simplicity rank of the remaining hypotheses decreases by at least 1. Since the maximum rank over all is k, there can be at most k mind changes by the Occam learner.

($\Rightarrow$) Let Ψ be an arbitrary learner. Suppose that there exists an inclusion chain

$$H \subset H_1 \subset \cdots \subset H_i \subset \cdots \subset H_k.$$

There is a possible data stream that enumerates all and only evidence items that are consistent with H. On this data stream, the learner Ψ must output H at some finite stage m to converge to the correct hypothesis. At this point, there is a data stream that extends the finite evidence sequence and enumerates all and only evidence items that are consistent with H_1. Again the learner Ψ must output H_1 at some finite stage m_1 to converge to the correct hypothesis. This leads to at least one mind change by the learner. Repeating this argument, we can extend the data streams after the mind change occurs consistent with hypotheses $H_2, \ldots, H_i, \ldots, H_k$, in such a way that a mind change occurs for each hypothesis in the chain. The result is a data stream on which the learner Ψ requires at least k mind changes. Since Ψ is an arbitrary learner, the construction shows that in the worst case, every learner requires at least k mind changes. ■

This result can be extended to infinite hypothesis spaces using transfinite mind change bounds and a topological generalization of the concept of simplicity rank. The concept of inductive simplicity developed applies therefore in a wide class of problems. The main restriction is that learning with a mind change bound requires that some hypothesis be conclusively verifiable, in the sense that there is some evidence that is consistent only with that hypothesis. Kelly presents a generalization of mind change optimality that relaxes this assumption.

Proposition 3 characterizes the inductive complexity of an entire learning problem. The next proposition concerns the properties of optimal learners. The main result is that the Occam learner is the *only* learner that achieves reliable, steady, and fast convergence.

PROPOSITION 4. *For a finite hypothesis space, the Occam learner is the only learner that reliably identifies a correct hypothesis, is convergence-time efficient, and strongly mind change optimal.*

Proof Outline. Any reliable learner is strongly mind change optimal if and only if whenever it produces a nonvacuous hypothesis, the hypothesis is the uniquely simplest consistent with the evidence. For otherwise the learner may incur one more mind change than necessary given the evidence, as illustrated in the top right box of Figure 4. What distinguishes the Occam learner from other strongly mind change optimal learners is that the Occam learner does not wait: it immediately conjectures the uniquely simplest hypothesis as soon as there is one, whereas other strongly mind change optimal may output? instead. It is easy to see that the Occam learner is uniformly faster than all such learners: Since the output? is not correct for any data stream, whenever a non-Occam learner outputs? on an evidence sequence $\langle e_1, \ldots, e_m \rangle$, its convergence time is later than stage m. The Occam learner by contrast converges to the uniquely simplest hypothesis H by time m on every data stream extending the evidence $\langle e_1, \ldots, e_m \rangle$ for which H is correct. Therefore the Occam learner possibly converges sooner than the non-Occam learner, and never slower. ■

Luo and Schulte [24] provide a formal proof that includes the general case of infinite hypothesis spaces. Without the requirement of convergence-time efficiency, the conjectures of a strongly mind-change optimal learner are no longer uniquely determined, because the learner can always wait for more evidence to make a conjecture without selecting a hypothesis. This is a reasonable inductive strategy in many cases. However, as the argument above showed, it remains the case that when the learner does eventually change its mind, it must be to adopt a unique maximally simple hypothesis. Implementing the Occam learner requires determining the simplicity rank of a hypothesis. In the next section we determine the simplicity rank of causal graphs.

7. Inductive Simplicity for Learning Causal Graphs

In this section we characterize the inductive simplicity rank of a causal graph and the Occam learner for causal graphs.

A fundamental result in causal graph theory characterizes the inclusion relation between two causal graphs $G_1 \subset G_2$, meaning that G_2 is consistent with all dependency statements that are consistent with G_1. The Meek-Chickering theorem shows that the inclusion holds just in case G_2 can be transformed into G_1 by a sequence of two types of transformations: (i) deleting edges, and (ii) reversing a covered arc [26],[4, Thm. 4]. For definitions, please see [36] (Figure 5).

The Meek–Chickering theorem is the basis for the following characterization of inductive simplicity for causal graphs.

Proposition 5. *The inductive simplicity rank of a causal graph or pattern containing edges* $\mathbf{E}$ *is* $|\mathbf{V}| - |\mathbf{E}|$, *the number of edges that are* not *included in the graph.*

Journal: 11225 Article No.: 9829 TYPESET DISK LE CP Disp.:2018/8/27 Pages: 33

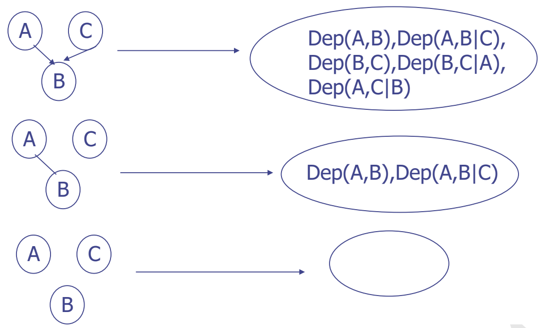


Figure 5. Left: An inclusion chain among causal graph patterns. The dependencies of the bottom pattern are included in those for the middle pattern which are included in those for the top patterns. Right: Representative dependency statements for each pattern. In figures we use the notation $Dep(A, B|C)$ for $A \not\perp\!\!\!\perp |C$

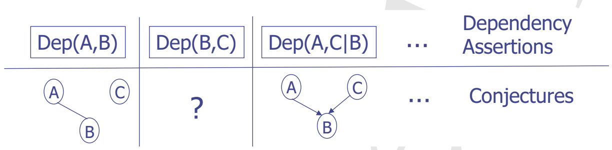


Figure 6. The hypotheses of the Occam learner on a sequence of three observed dependency statements

The formal proof can be found in [36]. So the simplest graph is the empty one, and the most complex graph is the complete one that contains all possible adjacencies. It is not surprising that the inductive complexity of a graph increases with the number of edges in the graph. The surprising aspect of the proposition is that the number of edges is all that matters: the direction of the adjacencies does not affect the simplicity rank.

The Occam learner for causal graph learning is therefore as folllows. Let $\mathcal{D}$ be a list of observed dependencies.

$$\Psi_{occam}(\mathcal{D})$$

$$= \begin{cases} ? & \text{if there is no uniquely simplest pattern consistent with the dependencies } \mathcal{D} \\ G & \text{if } G \text{ is the uniquely simplest pattern consistent with } \mathcal{D}. \end{cases}$$

Figure 6 illustrates some of the hypotheses of the Occam learner. Schulte et al. [36] provide a more elaborate example for a graph with four nodes.

As the number of variables increases, it becomes challenging to determine whether there is a uniquely optimal causal graph for a given list of observed correlations. Schulte et al. [36, Th. 23] show that computing the outputs of an Occam learner is NP-hard. Therefore we can conclude that there is no algorithm that implements the Occam learner exactly in reasonable (polynomial) computation time. Researchers in causal graph learning have developed a number of heuristic search algorithms that can be seen as approximating the Occam learner [35].

So far we have considered the problem of learning causal relationships among *variables*. However, some causal relationships involve specific values of variables. In the next section, we examine causal learning with Occam's razor for variable values.

8. The Occam Learner for Causal Context-Sensitive Causal Relationships

We provide a motivation for our approach and overview of our results, then go into the technical details.

8.1. Context-Sensitive Dependencies: Overview and Motivation

For discrete variables, it is common that a causal relationship holds not in general, but only conditional on the values of some variables [1,8,9]. These values establish a *context* that may reveal additional causal relationships. For a simple example, there may be a causal relationship between the intelligence of a student and their grade in a course. But this relationship holds only for students actually registered in the course, that is, conditional on Registration being true. Geiger and Heckerman discuss the following example [9]. Figure 7 shows a Bayesian network structure for this example.

> A guard of a secured building expects three types of persons to approach the building's entrance: workers in the building, approved visitors, and spies. As a person approaches the building, the guard can note its gender and whether or not the person wears a badge. Spies are mostly men. Spies always wear badges in an attempt to fool the guard. Visitors dont wear badges because they dont have one. Female workers tend to wear badges more often than do male workers. The task of the guard is to identify the type of person approaching the building.

Therefore for workers, gender is causally related to wearing a badge, but for spies and visitors, it is not. As Figure 7 illustrates, a single graph structure cannot represent this pattern explicitly, only implicitly through

Journal: 11225 Article No.: 9829 TYPESET DISK LE CP Disp.:2018/8/27 Pages: 33

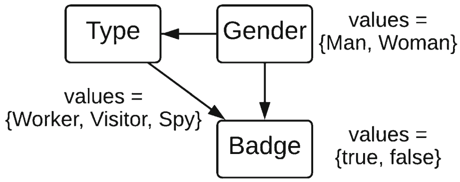


Figure 7. A Bayesian network for Geiger and Heckerman's security guard example. The type of a person predicts its gender (spies are mostly men). Conditional on *Type* = *Worker*, *Gender* predicts the wearing of a badge. However, conditional on *Type* = *Visitor* or on *Type* = *Spy*, *Gender* and *Badge* are independent. However, the graph structure is consistent with every dependency assertion, and fails to represent the two context-sensitive independencies

setting appropriate conditional probability parameters. An explicit context-sensitive representation of conditional probabilities conveys more information to the user, improves statistical efficiency, and facilitates faster inference to answer probabilistic queries [1,2,8]. Combining learning causal graphs and context-sensitive independencies suggests a two-part approach: first, employ an Occam learner to find a maximally simple graph with a minimum number of edges, then another Occam method to find a maximally simple representation of context-sensitive (in)dependencies between a child node and its parents.

One approach to representing context-sensitive (in)dependencies is to employ different causal graphs for different contexts, as in multinets or similarity networks [9]. Another is to employ a structured representation of the conditional probability parameters [1,8,21,30].

Probability estimation diagrams [2] are an intuitive formalism that provides a compact structured representation of conditional probabilities (see Figure 8 below for an example). In this section we examine *the mind-change optimal Occam learner for identifying a probability estimation diagram.*

Our main result is that the inclusion depth complexity of a PED is given by the number of its terminal nodes (with no children). The number of terminal nodes is equivalent to the number of parameters required to specify a joint distribution. Inductive simplicity as defined in terms of inclusion chain therefore agrees with standard statistical model selection criteria, such as BIC and AIC, in measuring the complexity of a PED by the number of its parameters, but offers a novel justification for minimizing the number of parameters: selecting the simplest PED consistent with the observed dependencies minimizes the number of worst-case mind changes a causal dependency learner may have to undergo. Statistical model selection criteria do

Journal: 11225 Article No.: 9829 TYPESET DISK LE CP Disp.:2018/8/27 Pages: 33

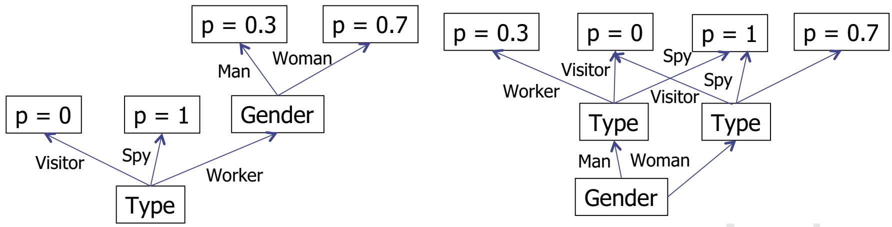


Figure 8. Two equivalent probability estimation diagrams for the security guard graph of Figure 7, for the child node *Badge*, and the parent nodes *Type* and *Gender*. We define $p = P(Badge = T)$. Both diagrams entail the same conditional dependencies and independencies

not usually take into account the number of edges in a causal graph.[3] However, because the number of parameters increases with the number of edges, for many graphs ranking by the number of parameters leads to very similar to results as our two-part scheme. We next give the formal definitions to develop our technical results.

8.2. Probability Estimation Diagrams

In what follows we consider a fixed graph structure G and a node Y in the graph. The parents of Y are denoted $X_1, \ldots, X_m$. In causal terminology, the child node Y represents a dependent or outcome variable, and the parent nodes X_i represent independent or treatment variables. An assignment of m values to each independent variable is denoted by boldface notation such as $\mathbf{x}, \mathbf{x}'$. The parameters of a causal graph specify conditional probabilities of the form

$$P(Y = y|\mathbf{X} = \mathbf{x}).$$

Often the required conditional probabilities are specified simply by enumerating them. The result is a flat table of conditional probabilities. A tabular representation fails to capture additional structure in the conditional distribution of the child variable given parent variable values; thus it fails to represent context-sensitive independencies. A **probability estimation diagram** is a DAG D whose nonterminal nodes are labelled with the parents of Y. Each terminal node (with no children) is labelled with a probability distribution over the possible values of Y. An edge from a node labelled X to a child is labelled with a possible value from the domain of X. The edges originating

[3]An exception is the minimum message length criterion, which explicitly includes both the number of edges and the number of parameters in an additive overall complexity measures [7].

from a node labelled X partition the domain of X, meaning that every possible value is assigned to one and only one edge. This entails that for any complete assignment of parent values, following the appropriate edges in the diagram leads to a unique terminal, which we denote as $\mathit{terminal}_D(\mathbf{x})$. Figure 8 illustrates two probability estimation diagrams for the security guard problem.

Given a probability estimation diagram, *it is easy to translate paths in the diagram into probabilistic clauses:* logical rules with probabilities attached. For instance, one of the paths in the left diagram of Figure 8 corresponds to the probabilistic clause

$$\mathit{Badge} = T \leftarrow \mathit{Type} = \mathit{Worker}, \mathit{Gender} = \mathit{Woman}; p = 0.7.$$

Such clauses provide a way to combine the formalism of first-order logic with probabilistic reasoning [27]. Khosravi *et al.* show that learning probability estimation trees (a special type of PEDs) to augment causal graph structures is a scalable way to discover probabilistic clauses that provide accurate predictions [20].

8.3. The Diagram Learning Problem

The learning problem is to identify a probability estimation diagram structure that is sufficiently complex to represent the true conditional probability distribution of the child variable conditional on the parent variables. The evidence items for this learning problem are conditional inequalities of the form

$$P(Y|\mathbf{X} = \mathbf{x}) \neq P(Y|\mathbf{X} = \mathbf{x}')$$

where $\mathbf{X}$ denotes the parent variables and $\mathbf{x}$ is an assignment of values to the parents. Thus an evidence item asserts that the distribution of the child variable, conditional on one assignment of values to the parent variables, differs from the distribution conditional on another assignment of values to the parent variables.

To complete the definition of the learning problem, we need to specify the set of evidence items are consistent with a diagram. A diagram D entails a conditional *equality* constraint

$$P(Y|\mathbf{X} = \mathbf{x}) = P(Y|\mathbf{X} = \mathbf{x}')$$

if $\mathit{terminal}_D(\mathbf{x}) = \mathit{terminal}_D(\mathbf{x}')$. A diagram is **consistent with** a conditional *in*equality constraint $P(Y|\mathbf{X} = \mathbf{x}) \neq P(Y|\mathbf{X} = \mathbf{x}')$ if the diagram does not entail its negation. Note that whether a diagram is consistent with an (in)equality constraint depends only on the qualitative graph structure of the diagram, not on the quantitative probability estimates in its terminals. In sum, the components of the learning problem in this section are as follows.

Hypothesis Space The set of probability estimation diagram structures for a child variable Y with parent variables $\mathbf{X}$.

Evidence Items Conditional distribution inequalities of the form

$$P(Y|\mathbf{X} = \mathbf{x}) \neq P(Y|\mathbf{X} = \mathbf{x}').$$

Consistency Relation A diagram structure is consistent with an inequality constraint $P(Y|\mathbf{X} = \mathbf{x}) \neq P(Y|\mathbf{X} = \mathbf{x}')$ if the assignments $\mathbf{x}, \mathbf{x}'$ are mapped to different terminals.

The Occam learner for this problem selects the inductively simplest diagram. In the next section we characterize the inductive simplicity rank of a probability estimation diagram.

9. The Inductive Simplicity of a Probability Estimation Diagram

Different diagram structures may be empirically equivalent in the sense of being consistent with exactly the same probabilistic inequalities. An equivalence class can be characterized by a partition of parent value assignments. More precisely, the terminals of a diagram induce a partition of the parent value assignments into equivalence classes, where two assignments are equivalent if they are mapped to the same terminal node:

$$\mathbf{x} \equiv \mathbf{x}' \iff terminal_D(\mathbf{x}) = terminal_D(\mathbf{x}'). \tag{4}$$

In general, if Eq. (4) holds for an equivalence partition $\equiv$ and a diagram D, we say that the partition **corresponds** to the diagram. It is easy to see that two diagram structures are consistent with exactly the same probabilistic inequalities just in case they corresond to the same partition. Table 1 shows the partition that corresponds to the diagrams for the security guard problem. Note that *the number of terminals in a diagram equals the size of any partition that corresponds to it.* Therefore equivalent diagrams have the same number of terminals, even if their internal structure is different.

It is easy to see that for every partition, there is a corresponding diagram: We can build a tree structure whose branches corresond to the possible assignment of parent values. If two assignments are equivalent, the corresponding branches end in the same terminal node. We thus have the following lemma.

Table 1. Each of the two diagrams in Figure 8 corresponds to the same 4-cell partition of parent assigments, shown in the table

Cell 1	Cell 2	Cell 3	Cell 4
Visitor = T, Gender = Man Visitor = T, Gender = Woman	Spy = T, Gender = Man Spy = T, Gender = Woman	Worker = T, Gender = Man	Worker = T, Gender = Woman

Lemma 6. *Consider the space of complete parent value assignments* $\mathbf{X} = \mathbf{x}$ *for a child node* Y.

1. *For every diagram structure* D, *there is a corresponding partition of the parent value assignments.*
2. *For every such partition* $\equiv$, *there is a corresponding diagram structure.*

A diagram semantically includes another if it is consistent with more inequality constraints. Therefore diagram D' includes D if and only if the partition of D' refines the partition of D; we denote this relationship by $D \leq D'$. The benefit of the partition representation is that inclusion depth in partition space is simply characterized by the size of the partition. Consider partitions of a finite set S elements. So the coarsest partition has size 1, and the maximally refined partition has size $|S|$. A refinement chain starting with partition $\equiv$ is a sequence

$$\equiv \leq \equiv_1 \leq \equiv_i \cdots \equiv_k,$$

where each element in the chain refines its predecessor but not vice versa. The **refinement depth** of partition $\equiv$ is the length k of a refinement chain starting with partition $\equiv$. The starting partition $\equiv$ is not included in the length count. The next proposition asserts that the refinement depth of a partition is simply the size $|S|$ of the most refined partition, minus the size of the partition. The basic reason for this is that to form a maximally long refinement chain, each partition in the chain should split exactly one cell of its predecessor into exactly two cells. Thus each partition in the chain contains exactly one more cell than its predecessor. Any other way to construct a refined partition is equivalent to merging two-way splits and therefore shortens the chain unnecessarily.

Proposition 7. *The refinement depth of a partition* $\equiv$ *on a finite set with* N *elements is* $N - |\equiv|$, *where* $|\equiv|$ *is the size of the partition.*

Journal: 11225 Article No.: 9829 TYPESET DISK LE CP Disp.:2018/8/27 Pages: 33

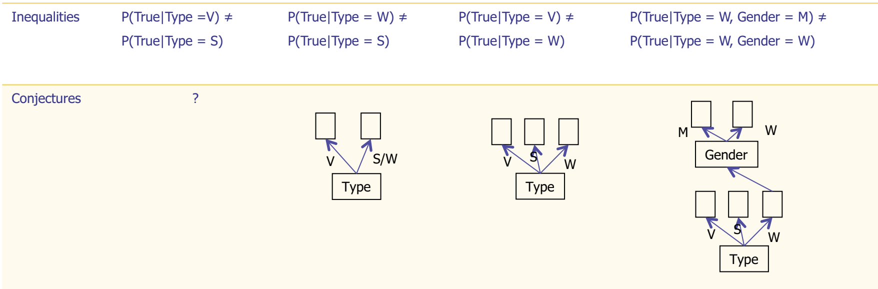


Figure 9. To illustrate the Occam learner for the security guard problem. We have used obvious abbreviations for variable values. The conditional probabilities are all of the form $P(Badge = T|\cdot)$, abbreviated as $P(T|\cdot)$

The proposition follows from the classical result [22] that the rank of a partition over a finite set is $N - |\equiv|$. To make the paper self-contained, we provide a direct proof in the appendix using our notation. By the correspondence Lemma 6, the inclusion depth of a diagram is the refinement depth of its corresponding partition. By Proposition 7, the refinement depth is the size of the partition, which equals the number of terminals in the diagram. Therefore the inclusion depth of a diagram equals the number of terminals in the diagram. We summarize our results in the following corollary.

Corollary 8. *The inclusion depth of a diagram is the number of its terminals. Therefore the Occam learner outputs a diagram with the minimum number of terminals, if all such diagram structures are equivalent (i.e., they are consistent with exactly the same probabilistic inequalities). Otherwise it outputs?.*

Figure 9 provides an example of the Occam learner. We leave for future work the analysis of the computational complexity of implementing the Occam learner.

10. Conclusion

An application of Occams razor to inductive inference directs us to choose the simplest hypothesis consistent with the evidence. The principle is plausible but vague to the extent that the concept of simplicity is undefined. Learning theorists have developed a precise definition of inductive simplicity based on the topology of the space of alternative hypotheses, which leads to a learning-theoretic version of Occam's razor. The learning-theory razor can be justified in terms of learning performance: it is the only learning

Journal: 11225 Article No.: 9829 TYPESET DISK LE CP Disp.:2018/8/27 Pages: 33

method that guarantees reliable, steady, and fast convergence to a correct hypothesis. As a case study, we applied Occam's inductive razor to learning causal relationships from observed correlations, an important and challenging practical problem. The Occam learner selects the causal graph that represents the observed dependencies among variables with a minimum number of edges. For context-sensitive dependencies, that may hold only in a given context, the Occam learner augments the causal graph with probability estimation diagrams that suffice to explain the observed correlations, with a minimum number of free parameters. Probability estimation diagrams are easily converted to probabilistic logical rules, so causal learning can be used to discover logical rules that represent statistical patterns in the data.

The learning-theoretic version of Occams razor has a clear justification in terms of learning performance guarantees. In many learning problems, including causal modelling, it leads to plausible inferences and to substantive insights into the methodological structure of the problem.

Acknowledgements. This research was supported by an NSERC discovery grant to the author. Preliminary results were presented at the Center for Formal Epistemology at Carnegie Mellon University. The author is grateful to the audience at the Center for helpful comments.

Proof of Proposition 7

PROPOSITION 9. *The refinement depth of a partition $\equiv$ on a finite set with N elements is $N - |\equiv|$, where $|\equiv|$ is the size of the partition.*

Proof Outline. The proof is by downward induction on partition size $|\equiv|$. Base case: $|\equiv| = N$. Then the partition is maximally refined and hence a maximal refinement chain contains only $\equiv$, so its length is counted as 0. Inductive Step: Assume the hypothesis for $n+1$ and consider a starting partition $\equiv$ of size n. We can split one cell of $\equiv$ into two subcells to obtain a partition $\equiv'$ that contains exactly one more cell than $\equiv$. So $\equiv'$ contains $n+1$ cells. By inductive hypothesis, there is a refinement chain

$$\equiv' \leq \equiv_1 \cdots \equiv_{N-(n+1)}$$

that extends $\equiv'$ with $N-(n+1)$ elements. Hence the refinement depth of the starting partition $\equiv$ is *at least* $N-(n+1)+1 = N-n$.

To show that the refinement depth if $\equiv$ is *at most* $N - n$, consider any maximal refinement chain

$$\equiv \leq \equiv_1 \leq \equiv_i \cdots \equiv^N \tag{5}$$

where $\equiv^N$ denotes the maximally refined partition. We argue that (*) the partition $\equiv_1$ splits exactly one cell of the starting partition $\equiv$ in exactly two subcells. For suppose otherwise for contradiction. If $\equiv_1$ splits two or more cells of the starting partition, form a partition $\equiv_{0.5}$ that contains the first cell as in the starting partition $\equiv$, but splits the other cells as in partition $\equiv_1$. Then $\equiv_{0.5}$ refines $\equiv$ and is refined by $\equiv_1$. Therefore the chain

$$\equiv \leq \equiv_{0.5} \leq \equiv_1 \leq \equiv_i \cdots \equiv^N$$

refines the starting partition and is longer than the chain (5). So the original chain is not maximally long, contrary to assumption. This establishes that $\equiv_1$ splits at most one cell of the starting partition $\equiv$. If $\equiv_1$ splits the starting partition cell into more than two members, we can again construct an intermediate partition $\equiv_{0.5}$ by following only one of the splits and not the others. Then there exists a longer refinement chain than that in Equation (5), contrary to assumption. This establishes the claim (*) that the partition $\equiv_1$ splits exactly one cell of the starting partition $\equiv$ in exactly two subcells. Therefore $\equiv_1$ contains exactly one more cell than the starting partition:

$$|\equiv_1| = |\equiv_1| + 1 = n + 1.$$

Applying the inductive hypothesis to $\equiv_1$, it follows that the chain

$$\equiv_1 \leq \equiv_i \cdots \equiv^N$$

has length at most $N-(n+1)$, so the chain (5) has length at most $N-(n+1)+1 = n$ members. Since this chain was chosen to have maximum length, the refinement depth of the starting partition $\equiv$ is n. ■

Appendix: d-separation

An (undirected) **path** in G is a sequence of nodes such that every two consecutive nodes in the sequence are adjacent in G and no node occurs more than once in the sequence. A node Y is a **collider on undirected path** p in DAG G if p contains a triple $X \rightarrow Y \leftarrow Z$. Thus a collider Y is a *common effect* of X and Z. If X and Z are adjacent in G, the collider Y is **shielded**, otherwise **unshielded**. Every Bayesian network structure defines a separability relation between a pair of nodes X, Y relative to a set of nodes $\mathbf{S}$, called **d-separation**: if X, Y are two variables and $\mathbf{S}$ is a set of variables disjoint from $\{X, Y\}$, then $\mathbf{S}$ d-separates X and Y if along every (undirected) path between X and Y there is a node W satisfying one of the following conditions:

1. W is a collider on the path and neither W nor any of its descendants is in $\mathbf{S}$, or
2. W is not a collider on the path and W is in $\mathbf{S}$.

We write $(X \perp\!\!\!\perp Y|\mathbf{S})_G$ if X and Y are d-separated by $\mathbf{S}$ in graph G. If two nodes X and Y are not d-separated by $\mathbf{S}$ in graph G, then X and Y are **d-connected** by $\mathbf{S}$ in G, written $(X \not\perp\!\!\!\perp Y|\mathbf{S})_G$.

Example. In the graph of Figure 2, the node `wet` is an unshielded collider on the path `sprinkler` − `wet` − `rain`; node `wet` is not a collider on the path `sprinkler` − `wet` − `slippery`. The pattern of the network has the same skeleton, but contains only two edges that induce the collider `wet`. The variables `sprinkler` and `rain` are d-separated given the set $\{\texttt{season}\}$, written $(\texttt{sprinkler} \perp\!\!\!\perp \texttt{rain}|\texttt{season})_G$, which can be seen as follows. There are two undirected paths from `sprinkler` to `rain`, namely `sprinkler` − `wet` − `rain` and `sprinkler` − `season` − `rain`. For the first path, clause (1) of the definition of d-separation applies, since `wet` is a collider on the path `sprinkler` − `wet` − `rain` and neither `wet` nor its descendant `slippery` is contained in the conditioning set $\{\texttt{season}\}$. For the second path, clause (2) applies, since `season` is not a collider on the path `sprinkler` − `season` − `rain` and `season` is a member of the conditioning set $\{\texttt{season}\}$. The variables `sprinkler` and `rain` are *not* d-separated given the set $\{\texttt{season}, \texttt{wet}\}$, written $(\texttt{sprinkler} \not\perp\!\!\!\perp \texttt{rain}|\texttt{season})_G$, because `wet` is a collider on the path `sprinkler` − `wet` − `rain` contained in the conditioning set, which violates clause (1) of the definition of d-separation.

A fundamental theorem of causal graph theory entails that a causal graph structure over discrete variables is consistent with a set of observed dependencies if and only if for each observed dependency, the corresponding d-connection relation holds in the causal graph structure.

References

[1] Boutilier, C., N. Friedman, M. Goldszmidt, and D. Koller, Context-specific independence in bayesian networks, in *UAI*, 1996 pp. 115–123.

[2] Boutilier, C., T. L. Dean, and S. Hanks, Decision-theoretic planning: Structural assumptions and computational leverage, *Journal of Artificial Intelligence Research (JAIR)* 11:1–94, 1999.

[3] Case, J., and C. Smith, Comparison of identification criteria for machine inductive inference, *Theoretical Computer Science* 25:193–220, 1983.

[4] Chickering, D., Optimal structure identification with greedy search, *Journal of Machine Learning Research* 3:507–554, 2003.

[5] Cooper, G., An overview of the representation and discovery of causal relationships using Bayesian networks, in C. Glymour and G. Cooper, (eds.), *Computation, Causation, and Discovery*, AAAI Press/The MIT Press, Cambridge, 1999, pp. 4–62.

[6] de Campos, L. M., A scoring function for learning Bayesian networks based on mutual information and conditional independence tests, *Journal of Machine Learning Research* 7:2149–2187, 2006.

[7] Dowe, D. L., MML, hybrid bayesian network graphical models, statistical consistency, invariance and uniqueness, in *Handbook of Philosophy of Science*, volume 7: Handbook of Philosophy of Statistics, Elsevier, 2011.

[8] Friedman, N., and M. Goldszmidt, Learning Bayesian networks with local structures, in *Proceedings of the NATO Advanced Study Institute on Learning in graphical models*, Norwell, MA, USA, Kluwer Academic Publishers, 1998, pp. 421–459.

[9] Geiger, D., and D. Heckerman, Knowledge representation and inference in similarity networks and Bayesian multinets, *Artificial Intelligence* 82(1-2):45–74, 1996.

[10] Genin, K., and K T. Kelly, The topology of statistical verifiability, in *Proceedings Conference on Theoretical Aspects of Rationality and Knowledge, TARK*, 2017, pp. 236–250.

[11] Giere, R. N., The significance test controversy, *The British Journal for the Philosophy of Science* 23(2):170–181, 1972.

[12] Glymour, C., On the methods of cognitive neuropsychology, *British Journal for the Philosophy of Science* 45:815–835, 1994.

[13] Gold, E. M., Language identification in the limit, *Information and Control* 10(5):447–474, 1967.

[14] Heckerman, D., A tutorial on learning with Bayesian networks, in *Proceedings of the NATO Advanced Study Institute on Learning in graphical models*, 1998, pp. 301–354.

[15] Jain, S., D. Osherson, J. S. Royer, and A. Sharma, *Systems that Learn*, 2 edition, MIT Press, Cambridge, 1999.

[16] Kelly, K., *The Logic of Reliable Inquiry*, Oxford University Press, Oxford, 1996.

[17] Kelly, K., Justification as truth-finding efficiency: How Ockham's razor works, *Minds and Machines* 14(4):485–505, 2004.

[18] Kelly, K., Why probability does not capture the logic of scientific justification, in C. Hitchcock, (ed.), *Contemporary Debates in the philosophy of Science*, Wiley-Blackwell, London, 2004, pp. 94–114.

[19] Kelly, K. T., and C. Mayo-Wilson, Causal conclusions that flip repeatedly and their justification, in *UAI*, 2010, pp. 277–285.

[20] Khosravi, H., O. Schulte, J. Hu, and T. Gao, Learning compact Markov logic networks with decision trees, *Machine Learning* 89(3):257–277, 2012.

[21] Lauritzen, S. L., and D. J. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems, *Journal of the Royal Statistics Society B* 50(2):157–194, 1988.

[22] Lucas, J. F., *Introduction to Abstract Mathematics*, Rowman & Littlefield, Lanham, 1990.

[23] Luo, W., Learning Bayesian networks in semi-deterministic systems, in *Canadian AI 2006*, number 4013 in LNAI, Springer-Verlag, 2006, pp. 230–241.

[24] Luo, W., and O. Schulte, Mind change efficient learning, *Information and Computation* 204:989–1011, 2006.

[25] Martin, E., and D. N. Osherson, *Elements of Scientific Inquiry*, The MIT Press, Cambridge, Massachusetts, 1998.

[26] Meek, C., *Graphical Models: Selecting causal and statistical models*, Ph.D. thesis, Carnegie Mellon University, 1997.

[27] Ngo, L., and P. Haddawy, Answering queries from context-sensitive probabilistic knowledge bases, *Theoretical Computer Science* 171(1-2):147–177, 1997.

[28] Pearl, J., *Probabilistic Reasoning in Intelligent Systems*, Morgan Kauffmann, San Mateo, CA, 1988.

[29] Pearl, J., *Causality: Models, Reasoning, and Inference*, Cambridge university press, Cambridge, 2000.

[30] Provost, F. J., and P. Domingos, Tree induction for probability-based ranking, *Machine Learning* 52(3):199–215, 2003.

[31] Putnam, H., Trial and error predicates and the solution to a problem of Mostowski, *The Journal of Symbolic Logic* 30(1):49–57, 1965.

[32] Schulte, O., Discussion. What to believe and what to take seriously: A reply to David Chart concerning the riddle of induction, *The British Journal for the Philosophy of Science* 51(1):151–153, 2000.

[33] Schulte, O., Means-ends epistemology epistemology, *The British Journal for the Philosophy of Science* 79(1):141–147, 1996.

[34] Schulte, O., The co-discovery of conservation laws and particle families, *Studies in the History and Philosophy of Modern Physics* 39(2):288–314, 2008.

[35] Schulte, O., G. Frigo, R. Greiner, and H. Khosravi, The IMAP hybrid method for learning Gaussian Bayes nets, in A. Farzindar, and V. Keselj, (eds.), *Canadian Conference on AI*, volume 6085 of *Lecture Notes in Computer Science*, Springer, 2010, pp. 123–134.

[36] Schulte, O., W. Luo, and R. Greiner, Mind-change optimal learning of Bayes net structure from dependency and independency data, *Information and Computation* 208:63–82, 2010.

[37] Spirtes, P., C. Glymour, and R. Scheines, *Causation, prediction, and search*, MIT Press, Cambridge, 2000.

[38] Studeny, M., *Probabilistic Conditional Independence Structures*, Springer, Berlin, 2005.

[39] Tsamardinos, I., L. E. Brown, and C. Aliferis, The max-min hill-climbing Bayesian network structure learning algorithm, *Machine Learning* 65(1):31–78, 2006.

[40] Verma, T. S., and J. Pearl, Equivalence and synthesis of causal models, in *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence (UAI 1990)*, 1990, pp. 220–227.

[41] Xiang, Y., S. K. Wong, and N. Cercone, Critical remarks on single link search in learning belief networks. in *Proceedings of the 12th Annual Conference on Uncertainty in Artificial Intelligence (UAI 1996)*, 1996, pp. 564–57.

O. Schulte
School of Computing Science
Simon Fraser University
Burnaby BCV5A 1S6
Canada
`oschulte@cs.sfu.ca`