The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Learning Abduction under Partial Observability

Brendan Juba, Zongyi Li, Evan Miller
Dept. of Computer Science and Engineering
Washington University in St. Louis
{bjuba, zli, evan.a.miller} @wustl.edu

Abstract

Juba recently proposed a formulation of learning abductive reasoning from examples, in which both the relative plausibility of various explanations, as well as which explanations are valid, are learned directly from data. The main shortcoming of this formulation of the task is that it assumes access to *full-information* (i.e., fully specified) examples; relatedly, it offers no role for declarative background knowledge, as such knowledge is rendered redundant in the abduction task by complete information. In this work we extend the formulation to utilize such partially specified examples, along with declarative background knowledge about the missing data. We show that it is possible to use implicitly learned rules together with the explicitly given declarative knowledge to support hypotheses in the course of abduction. We also show how to use knowledge in the form of graphical causal models to refine the proposed hypotheses. Finally, we observe that when a small explanation exists, it is possible to obtain a much-improved guarantee in the challenging *exception-tolerant* setting. Such small, human-understandable explanations are of particular interest for potential applications of the task.

Introduction

Abduction is the task of inferring a plausible hypothesis to explain an observed or hypothetical condition. Although it is most prominently observed in scientific inquiry as the step of proposing a hypothesis to be investigated, it is also an everyday mode of inference. Simple tasks such as understanding stories (Hobbs et al. 1990) and images (Cox and Pietrzykowski 1986; Poole 1990) involve a process of abduction to infer an interpretation of the larger events, context, and motivations that are only partially depicted. Its significance to AI was first recognized by Charniak and McDermott (1985).

Abduction has been formalized in several different ways. The oldest formulations simply tried to minimize syntactic criteria such as the number of literals (e.g., in ATMS (Reiter and de Kleer 1987)) or more generally, a weighted cost per literal (Hobbs et al. 1990). Another prominent approach is to assume that a prior distribution over potential explanations is given (Bylander et al. 1991; Pearl 1988; Poole 1993) and treat the problem essentially as one of MAP inference. McIlraith (1998) provides a critical review of this work: both approaches have shortcomings. Namely, the syntactic measures are tantamount to assuming that the attributes in question represent either independent or mutually exclusive events. As for the Bayesian methods, abduction relies to an unusual degree on the quality of the prior that is used. Note that unlike in standard, inductive inference, in abduction we are specifically interested in inference tasks in which the observed data *do not* essentially determine the best hypothesis. Therefore, such approaches must directly confront the problem of choosing a highly informative prior, which is hard.

In this work, we consider a *learning to reason* (Khardon and Roth 1997) or *PAC-learning* (Valiant 1984; 2000) formulation of the combined task of *learning to abduce*, introduced by Juba (2016). In this formulation, one is given a collection of examples drawn from the prior distribution (i.e., example jointly sampled values of attributes) together with a condition to explain, represented as a Boolean formula c on the attributes. The task is then to propose a formula h, which essentially must be a k-DNF ($k = O(1)$) for computational reasons, satisfying the following two criteria:

1. *Plausibility:* the probability that h is satisfied on the prior distribution must be at least some (given) minimum value $\mu > 0$
2. *Entailment:* the probability that the condition to explain, c, is satisfied conditioned on the hypothesis h holding, is at least $1 - \epsilon$ for some given error tolerance $\epsilon > 0$.

By casting the task as operating directly on examples, Juba avoids the problem of explicitly learning and representing the prior distribution. The main shortcoming of this formulation is that it assumes access to *complete information,* so any attributes to be invoked in the explanation must be recorded in *all* of the examples. This is a problem, for example, when we wish to infer the intentions of characters in stories, which are frequently either left ambiguous or are assumed to be clear from the given context. It is also a problem if, for example, we would like to use the abduced hypothesis to guide further exploration that may include attributes that we previously were not measuring.

Our Results

In this work, we make the following contributions:

1. We extend Juba's formulation of the abduction task to

use partial examples and draw on declaratively specified background knowledge.

2. We consider how to incorporate the knowledge encoded in a tentative, partial causal model of a domain in the abduction task.
3. We observe that by using a covering algorithm, it is possible to guarantee significantly better explanations when exceptions and counterexamples are inevitable, but a small hypothesis (using relatively few terms) is adequate. Encouragingly, this is precisely the case of most interest, when the k-DNF is human-readable. Concretely, when some r-term k-DNF explanation on n attributes has an error rate of ϵ^*, we obtain an error rate of $\tilde{O}(r(\log\log n + \log k)\epsilon^*)$, in contrast to the bound obtained for the state-of-the-art algorithm of Zhang et al. (2017), which gave an error rate of $\tilde{O}(\sqrt{n^k}\epsilon^*)$ (but does not consider the effect of the size of the hypothesis).

Our extension of the abduction task to a partial-information formulation is analogous to Juba's (2013) extension of Khardon and Roth's (1997) deductive reasoning task, based on Michael's (2010) model of learning from partial examples. We also provide a guarantee that any formula that is (implicitly) observed true sufficiently often can be implicitly used along with a base of explicit, declarative knowledge to support a hypothesis entailing the conclusion. We note that this does not follow immediately from Juba's work on learning deduction; we must address foundational issues that arise due to the interplay between partial information and the conditional probability formulation of the abduction task.

Our incorporation of causal knowledge in the abduction task is a significant refinement of Juba's (2016) framework. It was possible there to incorporate some causal knowledge by restricting the set of terms that could be used in an explanation; for example, to capture a time-series notion of causality, one could require that the explanation only use attributes with earlier time indices than appear in the condition to be explained. Here, we consider how to use a partial (i.e., more permissive) graphical causal model (i.e., in Pearl's (2009) framework) to identify hypotheses that are *(i)* consistent with our knowledge of potential causes of the condition and *(ii)* minimal in a desirable sense, that they contain no literals that are *d-separated* by the rest from the condition to be explained. We stress that our formulation is consistent with the role of abduction in scientific discovery, in which we do not yet have firm knowledge of the underlying causal processes and are instead seeking plausible candidate hypotheses for further, rigorous investigation of a potential, more definite causal link.

Finally, our observations regarding the behavior of covering when the hypothesis is small are related to, but distinct from the observation that greedy covering algorithms achieve vastly reduced *sample complexity*, by Haussler (1988). By contrast, here, we observe that the "blow-up" of the *error rate* of the hypothesis we find (relative to the best-fit) is vastly reduced, as compared to the state-of-the-art algorithm for this task, due to Zhang et al. (2017). Actually, our result here is partially inspired by the use of a sophisticated covering algorithm by Zhang et al.; indeed, we anticipate that their algorithm also achieves a similar error bound for small k-DNFs, although they did not consider this. But, we observe here that even a very simple set-covering algorithm provides a good bound when the k-DNF is small. Indeed, this bound only depends logarithmically (rather than polynomially) on the number of attributes, and is thus much better than the bound they claimed in such a case. Again, we stress that encouragingly, these results hold for the cases of the most practical interest, in which we are seeking a small, human-readable formula.

Preliminaries

In this paper, we show how to perform abduction using partially observed examples. First, we will describe the model of partial observations we use, and then introduce implicit learning, the main tool to use partial observations.

Partial Observability

We work in a standard machine learning model in which the data consists of many *examples,* assigning values to a variety of *attributes.* In this work, we will only consider Boolean attributes. For example, if our data is about birds, each bird may correspond to an example and then there can be attributes such as: whether the bird has feathers or not, whether it eats bugs or not, and other properties. We denote the number of attributes by n and we denote the number of examples by m.

Partial observability means that some attributes of examples may be unknown. We represent this by allowing the value of each attribute to be 1 (true), 0 (false), or $*$ (unobserved). For instance, an example $\rho^{(i)}$ could be $[x_1 = 1,\ x_2 = *,\ \cdots,\ x_n = 0]$. (We take the convention of denoting the ith example by $\rho^{(i)}$ and the ith coordinate of an example ρ by ρ_i.) In the real world, it is hard to require each example to contain all of the attributes. Indeed, in data analysis, we are often interested in inferring the values of attributes that are not recorded as part of the data. Or, in some examples, one subset of the attributes may have been recorded, and another subset may have been recorded in another example. Both motivate relaxing the requirement of complete examples to partial examples.

We will work in a PAC-learning style framework in which the (complete) examples are drawn i.i.d. from an unknown distribution on "ground truth" examples. The partial examples are then produced from these complete examples by a separate random process we refer to as a *masking process*; for brevity, we refer to the distribution over partial examples induced by these two processes as *partial distributions*. This learning framework was introduced by Michael (2010), and the model of partial observability is essentially a variant of Rubin's model (Rubin 1976).

Definition 1 (Masking Process (Michael 2010)) *We say a partial example ρ is* consistent *with a completely observed example x if for every ith attribute, whenever $\rho_i \neq *$, $\rho_i = x_i$, which means, whenever an attribute is observed, its values in the complete example and the partial example are the same.*

A masking process M is a random function $M : \{0,1\}^n \rightarrow \{0,1,*\}^n$ that maps completely observed examples to consistent partial examples.

When we apply a masking process M to a distribution D, we get a masked distribution we denote by $M(D)$ that is consistent with D. In our abduction task, our partial examples are drawn from such a masked distribution $M(D)$.

Implicit Learning

The main tool to deal with partial observability is *implicit learning*. In this section we will explain how to perform implicit learning in service of abduction. Implicit learning means learning without producing explicit representations. Given a knowledge base (a set of formulas), and a query formula, we want to know if the knowledge base can derive the query formula. The main theorem of implicit learning says, as long as the formulas in a knowledge base are sufficiently observed in partial examples, we can determine whether the knowledge base can derive the query *without* explicitly constructing or representing the knowledge base.

Definition 2 ($(1-\epsilon)$-valid) *Given a ("ground truth") distribution D, we denote a formula ϕ to be* $(1-\epsilon)$-valid *if $\Pr_{x \in D}[\phi(x) = 1] \geq 1-\epsilon$.*

In other words, a formula is $(1-\epsilon)$-valid if it is correct with error up to ϵ. If we take a formula being true as an event of the distribution D, $(1-\epsilon)$-valid also means this event has probability larger than $(1-\epsilon)$.

Definition 3 (Restricted Formulas) *For a formula ϕ and partial example ρ, the* restricted formula *$\phi|_\rho$ is defined recursively from the base case where variables set to $*$ are unaffected, and others are replaced by ρ_i (We will define it for the de Morgan basis, $\{\vee, \wedge, \neg\}$, but it can be easily extended to other kinds of connectives.)*

- *If ϕ is $\neg\psi$, then if $\psi|_\rho \in \{0,1\}$, then $\phi|_\rho$ is the negation of $\psi|_\rho$, and otherwise it is $\neg(\psi|_\rho)$.*
- *If ϕ is $\psi \vee \xi$, then if either $\psi|_\rho$ or $\xi|_\rho$ is* 1*, $\phi|_\rho$ is also* 1*; if both are* 0*, then $\phi|_\rho$ is* 0*; if just one, say WLOG $\psi|_\rho$, is* 0*, then $\phi|_\rho$ is $\xi|_\rho$; and otherwise, $\phi|_\rho$ is $(\psi|_\rho) \vee (\xi|_\rho)$*
- *If ϕ is $\psi \wedge \xi$, then $\phi|_\rho$ is defined similarly to $\psi \vee \xi$, except of course that it is* 0 *if either are* 0*,* 1 *if both are* 1*, equal to the non-*1 *formula if the other simplifies to* 1*, and otherwise is equal to $(\psi|_\rho) \wedge (\xi|_\rho)$.*

Note that it is hard in general to say what the value of a formula should be under partial information. For example, tautologies always evaluate to 1 without any information, but it is intractable to detect this in general. By contrast, the local evaluation provided by a restriction is a linear-time operation that generalizes standard formula evaluation and sometimes evaluates the formula if enough information is provided.

Given an observation ρ, the value of $\phi|_\rho$ and ϕ should be the same. If ϕ has been observed true (or false) in ρ, then $\phi|_\rho$ is simultaneously true (or false). For example, if $\phi = x_1 \wedge x_2$, and in partial example ρ, x_1 is observed true, while x_2 is unobserved, then $\phi|_\rho$ is just x_2. On the other hand, if in ρ', x_1 is unobserved, while x_2 is observed false, then $\phi|_{\rho'}$ is just false, since ϕ is false in ρ'.

Definition 4 (Witnessed Formulas) *Given a partial example ρ, we say a formula ϕ is* witnessed *if $\phi|_\rho$ is 0 or 1.*

For a basic example, let $\phi = x_1 \vee x_2$. In a partial example, $x_1 = 1; x_2 = *$. Then ϕ is witnessed (true) even if x_2 is not observed. Notice that each formula can be either witnessed true, witnessed false, or not witnessed.

Definition 5 (Proof System) *Given a knowledge base KB (a set of formulas), and a query formula ϕ, a* proof *is a finite sequence of formulas $\psi_1, \cdots, \psi_k$, such that:*
1. $\{\psi_1, \cdots, \psi_k\} \vdash \phi$.
2. $\forall i \in [1,k]$, either $\psi_i \in KB$ or $\{\psi_1, \cdots, \psi_{i-1}\} \vdash \psi_i$.
*Where "$\vdash$" means "can prove" or "*provable*".*

Each step of the proof $\{\psi_1, \cdots, \psi_{i-1}\} \vdash \psi_i$ corresponds to a relation $R_j(\psi_1, \cdots, \psi_{i-1}, \psi_i)$. A proof system *is a set of such relations $\{R_j\}_{j=0}^{\infty}$, i.e., such that whenever $R_j(\psi_1, \cdots, \psi_{i-1}, \psi_i)$ holds, $\{\psi_1, \cdots, \psi_{i-1}\} \vdash \psi_i$.*

Definition 6 (Restriction-closed Proof System) *A proof system is* restriction-closed *if, for any step of the proof $\{\psi_1, \cdots, \psi_{i-1}\} \vdash \psi_i$, and any partial example ρ, $\{\psi_1|_\rho, \cdots, \psi_{i-1}|_\rho\} \vdash \psi_i|_\rho$. More generally, if there is a proof of $\phi|_\rho$ from $KB|_\rho$, we say that ϕ* is provable from *KB* under *ρ (and we may omit mention of KB when it is clear from context).*

The formal language may be confusing, but the definition is indeed intuitive. Consider the following example: $\psi_1 = x_1 \wedge x_2$, $\psi_2 = x_3 \wedge x_4$, $\phi = x_1 \wedge x_2 \wedge x_3 \wedge x_4$, $\{\psi_1, \psi_2\} \vdash \phi$. If in ρ, x_1, x_3 are observed true and x_2, x_4 are unobserved, then $\psi_1|_\rho = x_2$, $\psi_2|_\rho = x_4$, $\phi|_\rho = x_2 \wedge x_4$, We thus anticipate, $\{\psi_1|_\rho, \psi_2|_\rho\} \vdash \phi|_\rho$.

DecidePAC Algorithm Given knowledge base KB and partial examples $\{\rho^{(1)}, \cdots, \rho^{(m)}\}$ drawn from $M(D)$, for a query formula ϕ, DecidePAC can tell whether there is a proof of ϕ if the knowledge we need is witnessed sufficiently often: DecidePAC will *Accept* if there exists a proof of ϕ in from KB and formulas $\psi_1, \psi_2, \cdots$ that are simultaneously witnessed true with probability at least $1-\epsilon+\gamma$ on $M(D)$; or, if $[KB \Rightarrow \phi]$ is not $(1-\epsilon-\gamma)$-valid, then DecidePAC will $reject$ formula ϕ. Otherwise, $[KB \Rightarrow \phi]$ is $(1-\epsilon-\gamma)$-valid, but no adequate proof exists, and DecidePAC may accept or may reject. (There is no strict guarantee in this final case.)

Algorithm 1: DecidePAC

```
input : Formula $\phi, \epsilon, \delta, \gamma \in (0,1)$, partial examples
        $\rho^{(1)}, \cdots, \rho^{(m)}$ from $M(D)$ for $m(\delta, \gamma)$ as
        given in Lemma 7, hypothesis formulas $KB$
begin
    $FAILED \leftarrow 0$.
    foreach partial example $\rho^{(i)}$ in the list do
        if $KB|_{\rho^{(i)}} \not\vdash \phi|_{\rho^{(i)}}$ then
            Increment $FAILED$.
            if $FAILED > \lfloor \epsilon \cdot m \rfloor$ then
                return Reject
    return Accept
```

In the following, we will let $|\phi|$ denote the *size* of ϕ (in symbols) and $|KB|$ similarly denote the total size of the formulas in KB.

Lemma 7 (Implicit Learning (Juba 2013)) *Suppose that whether or not there exists a proof of ϕ from KB can be decided in time $T(n, |\phi|, |KB|)$ on input ϕ and KB over n variables. Let D be a distribution over examples, M be any masking process, and KB be any set of formulas. Then DecidePAC, on input query ϕ, KB, confidence parameter δ, accuracy parameter γ, and error tolerance ϵ, uses $O(1/\gamma^2 \log 1/\delta)$ examples, runs in time $O(T(n, |\phi|, |KB|)\frac{1}{\gamma^2}\log\frac{1}{\delta})$, and given that either*

- *$[KB \Rightarrow \phi]$ is not $(1-\epsilon-\gamma)$-valid w.r.t. D or*
- *there exists a proof of ϕ from $\{\psi_1, \cdots, \psi_k\} \cup KB$ s.t. $\psi_1, \cdots, \psi_k$ are simultaneously witnessed to evaluate to true with probability $1-\epsilon+\gamma$ over $M(D)$*

decides which case holds with probability $1-\delta$.

As noted by Juba (2013), DecidePAC may be applied to essentially all standard fragments of proof systems for which efficient proof search algorithms are known, e.g., width-bounded or treelike resolution and degree-bounded polynomial calculus. In particular, each of these fragments is a restriction-closed proof system.

Remark Juba (2013) uses the additive Chernoff bound. If we use the multiplicative Chernoff bound instead (Lemma 8, below), we find that DecidePAC will be able to distinguish whether $[KB \Rightarrow \phi]$ is not $(1-\epsilon(1+\gamma))$-valid, or is provable from an implicit knowledge base that is witnessed with probability $(1-\epsilon(1-\gamma))$, given $\frac{3}{\epsilon\gamma^2}\ln\frac{1}{\delta}$ examples.

Lemma 8 (Multiplicative Chernoff Bound) *Let $X_1, \cdots, X_m$ be independent random variables taking values in $[0,1]$, such that $E[\frac{1}{m}\sum_i X_i] = p$. Then for $\gamma \in [0,1]$,*

$$\Pr\left[\frac{1}{m}\sum_i X_i > (1+\gamma)p\right] \leq e^{-mp\gamma^2/3}$$

$$\text{and } \Pr\left[\frac{1}{m}\sum_i X_i < (1-\gamma)p\right] \leq e^{-mp\gamma^2/2}$$

Abduction

Here we will formulate a partial information version of the learning abduction task, and explain how to incorporate causal models in this task.

Abduction under Partial Observability

Given a query or an event, abduction is the task of finding an explanation for the query or event. An explanation is a combination of some conditions that may have caused the query. For example, when the query is "Engine does not run," an explanation can be "No gas, or key is not turned."

We require the resulting explanation to satisfy two conditions, *"plausibility"* and *"entailment."* Entailment means that when the conditions in the explanation are true, the query should also often be true, or at least rarely false. Thus, the explanation is a (potential) cause of the query. Plausibility means the explanation is often true. In other words, for many examples, these conditions are observed. This suppresses unlikely explanations such as "A comet hits the car," which is a valid entailment, but not plausible.

Juba (2016) defined a complete information form of this task as follows. Given a query condition c, we wish to find a hypothesis h that is plausible in the sense that $\Pr[h] \geq \mu$ for some minimum μ, and h entails c in the sense that $\Pr[c|h] \geq 1-\epsilon$. These definitions are inadequate when we only have partial information and cannot directly evaluate h and c. We propose the following extension of the abduction task:

Definition 9 (Partial Information Abduction) *Abduction is the following task: given any query formula c and independent partial examples $\{\rho^{(1)}, \cdots, \rho^{(m)}\}$ over a masked distribution $M(D)$, we want to find a k-DNF explanation h, such that the explanation h satisfies:*

1. $\Pr[\exists t \in h : t \text{ provable under } \rho] \geq \mu$ (Plausibility)

2. $\Pr\left[\begin{array}{r} \neg c \text{ provable} \\ \text{under } \rho \end{array} \middle| \begin{array}{r} \exists t \in h : t \text{ provable} \\ \text{under } \rho \end{array}\right] \leq \epsilon$ (Weak Entailment)

Here "provable" is again (c.f. the discussion following Lemma 7) with respect to the proof system we are using, which could be any restriction-closed fragment with an efficient algorithm for proof search. For example, in practice it could be tree-like or regular resolution.

Recall, a k-DNF explanation h with r terms is in the following form: $h = t_1 \vee t_2 \vee \cdots \vee t_r$ where each *term* $t_i = \ell_{i_1} \wedge \ell_{i_2} \wedge \cdots \wedge \ell_{i_k}$. For convenience, we say $t_i \in h$ and $\ell_{i_j} \in t_i$. We will assume k is a constant throughout.

We use k-DNFs primarily because prior work by Juba (2016) established that it is essentially the most expressive natural class of formulas for which this task is tractable; in particular, finding conjunctive explanations is likely intractable, even given complete information.

Example. Suppose you are lecturing for a large class, and you are interested in finding an explanation for why some students don't attend. Students' attendance is not directly observed since your class has an enrollment of more than 300, and you do not take attendance. The partial examples $\{\rho^{(1)}, \cdots, \rho^{(m)}\}$ correspond to your knowledge of students sampled uniformly at random from the class (with replacement). The examples' attributes $x_1, \cdots, x_n$ are Boolean, propositional attributes such as $x_1 :=$ "the student is a freshman," $x_2 :=$ "the student turned in Homework 2," $x_3 :=$ "the student likes me" and so on. $\rho^{(j)} = (0, 1, *, \cdots)$ then means that student j is not a freshman, has turned in Homework 2, and you do not (directly) observe whether the student likes you or not. The number of such attributes, n could be large (say, 10^3). A knowledge base KB may consist of rules such as (If "the student likes you", then "the student attends") and (If not "the student attends," and not "the student is a genius", then "the student will fail the class,") etc.

Let x_4 be "the student attends." Then our query formula is just the literal $c = \neg x_4$. A potential 2-DNF explanation formula h could be ("the student has dropped," (x_5)) or ("the student does very well," (x_6) and not "the student is interested," $(\neg x_7)$) or ("this class meets at 8 A.M." (x_8) and "the student sleeps in until 11 A.M." (x_9)). Formally, this would be the formula $h = (x_5) \vee (x_6 \wedge \neg x_7) \vee (x_8 \wedge x_9)$, and (x_5), $(x_6 \wedge \neg x_7)$, and $(x_8 \wedge x_9)$ are the three terms of h. To be a valid solution to our task, we require h to have the properties of plausibility and weak entailment. If you estimate that only 100 out of the 300 students enrolled attend lecture, you may want the hypothesis to be able to potentially explain at least why 150 out of the 200 students don't show up. This would correspond to setting $\mu = 50\%$. The condition $\Pr[\exists t \in h : t \text{ provable under } \rho] \geq \mu$ then means that at least 50% of the student population can be inferred (proved) to satisfy at least one of the three conditions (terms) of h. So, h addresses at least half of the population. Meanwhile, you want h to be a potential cause of c. If you are either told that a student has dropped the class, or you can somehow infer that a student sleeps in until 11 A.M., then you should generally not see him or her in your class. Of course, exceptional circumstances may arise, as when a student shows up to colllect a handout for another student or when a student stays up until past 8 A.M. and decides to attend your lecture for once to see what it is like. Such exceptional circumstances should comprise no more than an ϵ fraction of the population, and thus h will satisfy weak entailment.

Discussion. There are three different concepts of being true: 1. $observed$ (or $witnessed$), 2. $provable$, and 3. $true$. For example, let $t = x_1 \wedge \neg x_2$. In example $\rho^{(1)}$, it is observed that $x_1 = 1, x_2 = 0$, so t is observed to be true in $\rho^{(1)}$; in example $\rho^{(2)}$, $x_1 = 1$ while x_2 is unobserved, but assume in KB we have a rule $x_1 \Rightarrow \neg x_2$, then $\neg x_2$ is provable, so t is provable; in example $\rho^{(3)}$, nothing is observed and we know nothing, but in fact, t can be true. Notice that being observed can imply being provable, and being provable can imply truth. Each possible outcome of a formula is an event. When we talk about its probability, properly it is the event with respect to the joint distribution (M, D).

Plausibility. We have chosen to relax the condition that $h(x) = 1$ in Juba's complete information abduction task to the condition that some term of h is provable under ρ. This is of intermediate strength between h being observed and h being provable. Provability captures whether or not an agent "knows" t is true of a given partial example ρ. Our choice is somewhat like Levesque's notion of *vivid knowledge* (Levesque 1986), that the individual literals of some definite t should be known. The weaker condition that merely h is provable is also interesting, but seems much harder to work with; we leave it as a direction for future work. We could also have relaxed this to cases where $\neg h$ is not provable, but observe that this includes the cases where h is unknown in its favor. Note that this may "mix" many cases where h was actually false into our estimate of the effect of h occurring, which is not desirable, and we anticipate that it would harm the quality of the inferences we can draw.

Weak Entailment. We made the opposite decision for $c(x) = 1$, relaxing it to the condition that $\neg c$ is not provable (weak entailment), rather than requiring c to be provable (strong entailment). Notice that weak entailment and strong entailment are equivalent under complete information, since then "provability" collapses to evaluating to true, and $c = 1$ if and only if $\neg c = 0$.

Note that the classical logical formulation of abduction takes the stance that the given model completely captures the behavior of the system. So, in the classical formulation, whether or not a hypothesis entails the observation relative to the KB is a complete characterization of whether or not the hypothesis is a satisfactory explanation. Hence, the classical abduction task is, in this sense, closer to the complete information setting: the world model gives enough information to, in principle, decide whether or not each possible hypothesis is a good explanation (in the propositional case). By contrast, here we are seeking to learn the rules describing this world model. If the partial information we observe on the training data is inadequate, then we do not have access to this complete characterization, even given unlimited computation time.

So, we could consider "credulous" or "skeptical" standards in the face of this lack of information. Given our intended characterization of abduction as proposing "plausible" hypotheses given some tentative, partial knowledge of the world, perhaps to guide some further investigation, we are lead to prefer the "credulous" interpretation, that is, weak entailment. More precisely, the main reason for our choice of weak entailment is that we wish to not penalize a good h if it is often impossible to check whether or not c holds. At the same time, we would like to take ϵ to be very small, so that we can aggressively rule out h's for which c is frequently known to fail to occur. But, if we are including the outcome of c being unknown as a "failure" of h, then this suggests that in the cases where c is indeed often unknown, then ϵ must be large, even for a good h.

In any case, we stress that if we have enough information to learn a complete implicit knowledge base (as assumed in the logical formulation of the task), then provability with respect to the knowledge base again captures the formula values. Hence, the criteria use in our formulation will coincide with the properties used in the complete information task in such a case. The distinction between weak and strong entailment will have disappeared.

Pruning with Causal Models

Ideally, our explanation should be a potential cause of the query, not just something that correlates with it. Causal models capture precisely such relationships and so can be used to refine our explanations. If we know some causal relation such as: c and x_1 are conditionally independent given x_2 and x_3, i.e. $[c \perp x_1 | x_2, x_3]$. Then we say x_1 is redundant given x_2 and x_3. When we choose terms, we don't want a term like $t = x_1 \wedge x_2 \wedge x_3$ containing redundant attributes. So we prune the terms where some of their attributes are independent given the rest of the terms.

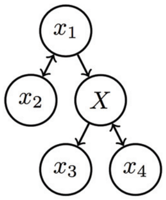

Figure 1: A Causal Model Where x_1 d-separates X and x_2

Partial Causal Models A *partial causal model* is given by a graph $G(E, V)$, where $V = \{x_i\}_{i=1}^n$ is the set of all attributes. When there is *no* directed edge from x_i to x_j, this represents that we know x_i has no (direct) causal effect on x_j (the converse is not necessarily true). Thus, the "default" (uninformed) causal model is a complete, bi-directed graph, meaning that any two nodes may have causal relation.

Definition 10 (d-separation (Pearl 2009)) *We say x and y are* d-connected *by z if and only if there exists some undirected path between x and y such that for every* collider *c on the path (i.e both of the edges to c on the path point to it), either c or a descendent of c is z, and no non-collider on the path is z. x and y are* d-separated *by z if and only if x and y are not d-connected by z.*

Notice that this definition is also valid for x, y and z as sets of nodes. The following lemma establishes that separation implies conditional independence.

Lemma 11 (Theorem 1.2.4 of (Pearl 2009)) *If x and y are d-separated by z, then $[x \perp y|z]$ in every distribution compatible with the graph.*

Continuing our example from earlier, suppose we are given a causal model such as in Figure 1, with the query $c = X =$ "the student attends lecture," and attributes $x_1 =$ "the student has learned similar material before," $x_2 =$ "the student majors in math," $x_3 =$ "the student will pass the class," and $x_4 =$ "the student is interested in the class." In the graph, there is only a directed edge from X to x_3, which means x_3 is not a good explanation for X. These are intuitively correlated, but it is the student's attendance (or lack of it) that has a causal effect on their likelihood of passing, and not the other way around. In the graph, x_1 d-separates X and x_2. That is, whether or not the student majors in math and whether or not he or she attends lecture are independent given that we know whether the student has learned the material before or not.

If $t = x_1 \wedge x_2$ and $t' = x_2$, then we should prefer t' over t. Because (i) whenever t is true, t' is true, and (ii) they have same causal effects on c. In such a case, where t is a subterm of t' and the event "t" is also a subset of "t'" in the probability space, we say t' *covers* t. We stress that although we also consider a term as the set of literals, we won't use "covering" to refer to this notion.

Corollary 12 *If a set of attributes x is d-separated by a set of attributes t′ from c, then for $t := x \wedge t'$:*
(i) $t \Rightarrow t'$ *and*
(ii) $\Pr[c|t] = \Pr[c|t']$.

In other words, for any t' obtained from t by removing one of its redundant attribute, t' covers t.

Proof of Corollary 12

i. Since t is a conjunction of literals, deleting literals of t cannot make it switch from true to false on a given assignment. So, $t = [x \wedge t'] \Rightarrow t'$.

ii. By Lemma 11, if t' d-separates x and c, then $[c \perp x|t']$. That is, $\Pr[c|t'] = \Pr[c|t' \wedge x] = \Pr[c|t]$. ■

We call terms that contain no redundant attributes *parsimonious terms*:

Definition 13 (Parsimony) *We define a term t to be* parsimonious *(with respect to a causal model CM) if there does not exist a set of attributes $s \subset t$, such that $t \setminus s$ separates s from all attributes of the query c in that CM.*

We conclude that the collection of parsimonious terms can fully cover the original set of terms:

Corollary 14 (Parsimonious Terms) *$\forall t$, $\exists t^*$ s.t. t^* is parsimonious, $t \Rightarrow t^*$, and* $\Pr[c|t^*] = \Pr[c|t]$

Implicit Abduction Algorithm

A k-DNF explanation is actually a disjunction of terms, $h = t_1 \vee t_2 \vee \cdots \vee t_r$. Each term represents a condition, or a possibility. Our goal is to find a formula that covers as many such conditions as possible while still being a potential cause of the query c. The problem, cast this way, is similar to the set cover problem, so we can use greedy algorithms to choose terms. In this section, we will develop the connection to the set cover problem, present our algorithm, and prove our main theorem.

Set Cover

There is a natural correspondence between our k-DNF abduction task and set cover: each example of abduction is an element of the set cover problem, and each term is a set. We say a term covers an example when the term is provable in that example. The number of examples from the distribution is equivalent to its frequency or empirical probability with respect to the distribution $M(D)$. If the resulting explanation consists of terms that are provable in most of examples, then we can conclude that our explanation is provable with high probability, and so this explanation satisfies "plausibility."

Definition 15 (Partial set cover problem) *Given a universe $U = \{x_1, \cdots, x_m\}$ of m elements, and collection $S = \{S_1, \cdots, S_N\}$ of subsets of U, the task is to find a subcollection T of S, such that T covers a μ-fraction of U, i.e., $|\bigcup_{S_i \in T} S_i| \geq \mu|U|$, and $|T|$ is minimized.*

The *greedy algorithm* just selects the set that can pick up most new items. Let S_j denote the j^{th} set that the algorithm picks: $S_j = \arg\max_{S \setminus S_1, \cdots, S_{j-1}} |S_j \cap (U \setminus \cup_{i=1}^{j-1} S_i)|$.

Lemma 16 (Greedy Algorithm (Slavík 1997)) *For m examples and letting $H(i)$ denote the ith harmonic number, the greedy algorithm returns a solution to the unweighted partial set cover problem of cost $H(\mu m) \cdot OPT \sim \log(\mu m) \cdot OPT$ where OPT is the minimum size of a family covering μm elements.*

Furthermore, unless NP has quasipolynomial time algorithms, no algorithm achieves a $(1-\epsilon)\log(m)$ approximation (Feige 1998), so the greedy algorithm achieves the optimal approximation ratio for set cover.

In our abduction task, we want to find an explanation satisfying the plausibility condition, i.e. $\Pr[\exists t \in h : \vdash t|_\rho] \geq \mu$; empirically, this means that for μ-fraction of the partial examples drawn from the distribution, some term of h should be provable under that example. Thus, we will use a slightly more restrictive notion of covering in our algorithm: we will consider a term t to cover the example ρ if t is *provable* in ρ. Therefore, if $\{t_1, \cdots, t_r\}$ cover a μ-fraction of examples (in this sense), then at least empirically, $\Pr[\exists t \in h : \vdash t|_\rho] \geq \mu$. We thus use set cover to perform abduction: examples are the elements of our universe, and terms are sets.

Applying the greedy algorithm, we find that if there exists an optimal explanation h^* that covers a μ-fraction of m examples with r terms, then the greedy algorithm will return an explanation h of size $r\log(\mu m)$, which also covers a μ-fraction of the m examples.

Algorithm 2: Implicit Abduction

```
input  : Knowledge base $KB$, Causal model $CM$,
         query $c$ and parameters $\mu, \epsilon, \delta, \gamma \in [0,1]$
output: A $k$-DNF explanation $h$
begin
    Initialize $T$ to be the set of all terms of at most $k$
    literals. Draw partial examples $\{\rho^{(1)}, \cdots, \rho^{(m)}\}$
    from $M(D)$ for
    $m = \frac{6r}{\mu\gamma^2}\left(\log\frac{3r}{\gamma^2} + \log\log\frac{2|T|}{\delta}\right)\log\frac{2|T|+4}{\delta}$
    1. forall the $t' \in T$ of size $\leq k-1$ do
        forall the $x \notin t'$ s.t. $t'$ d-separates $x$ from $c$ in
        $CM$ do Delete $t = x \wedge t'$ from $T$.
    2. forall the $t \in T$ s.t. $\#\{\rho :$
    $t$ provable under $\rho \wedge \neg c$ provable under $\rho\} > \mu\epsilon m$
    do Delete $t$ from $T$.
    3. Run greedy algorithm for set cover: use terms in
    T to cover a $\mu$-fraction of the examples. Get
    $\{t_1, \cdots, t_r\}$.
    $h \leftarrow t_1 \vee \cdots \vee t_r$
    return $h$.
```

Implicit Abduction Algorithm

In the implicit abduction algorithm, we first enumerate through all possible k literal terms.

1. The first step is to use our causal model to prune terms. Using an algorithm due to Geiger et al. (1989), we test whether or not a term contains a redundant literal. The algorithm then simply deletes such terms, so that we can choose an explanation from only parsimonious terms.
2. The second step is to check the rest of terms using the same technique underlying DecidePAC: We count the number of bad examples where $\neg c$ and t are both provable. If the bad examples are more than a $\mu\epsilon$-fraction, then we delete this term.

   By the Chernoff bound, all terms that pass the test then satisfy weak entailment: the error condition $[\vdash t|_\rho$ and $\vdash \neg c|_\rho]$ has probability at most $\mu\epsilon(1+\gamma)$.
3. The third step is to use the greedy algorithm to choose an explanation. If the algorithm can find an explanation covering a μ-fraction of examples, then the Chernoff bound guarantees that the explanation has probability at least μ.

Thus, if there exists a good explanation, we can find a parsimonious explanation satisfying entailment and plausibility.

Remark If μ^* is the optimal probability that the terms of a potential explanation h^* can be provable, Juba (2016) showed that a multiplicative approximation to μ^* can be easily found by binary search. We assume that such an estimate μ is given as input.

Let's return to our example of proposing possible explanations for why students don't attend lecture. In the first step, we prune out terms with redundant literals such as ("the student turns in Homework1," and "the student turns in any homework"). In the second step, we check if this term could (approximately) entail the query. Specifically, we count the number of example students who can be inferred to satisfy the term but can also be inferred to attend lecture. If there are too many such counterexamples, we discard this term. In the third step, we use a greedy algorithm to find a collection of terms that can describe at least μ-fraction of the total sample, so that our explanation could empirically explain the behavior of half of the class if $\mu = 0.5$. Finally, we return the explanation as an OR of these terms.

Theorem 17 (Implicit Abduction) *Given a query c, a causal model CM, partial examples $\rho^{(1)}, \cdots, \rho^{(m)}$ from a masked distribution $M(D)$, and a restriction-closed proof system with knowledge base KB, for constant k:*

If there exists a parsimonious r-term k-DNF $h^* = t_1^* \vee \cdots \vee t_r^*$ satisfying:

1. *With probability at least $(1+\gamma)\mu$ over ρ from $M(D)$, $\exists t_i^* \in h^*$, such that t_i^* is provable from KB under ρ (Plausibility).*
2. *Under ρ drawn from $M(D)$, if some term t^* of h^* is provable, then $\neg c$ is only provable with probability at most $(1-\gamma)\epsilon$. (Weak Entailment)*

Then, we can find a parsimonious k-DNF h in polynomial time, such that with probability $1-\delta$,

1. $\Pr[\exists t \in h$ *provable under* $\rho] \geq (1-\gamma)\mu$ *(Plausibility)*
2. $\Pr\left[\begin{matrix}\neg c \text{ provable} \\ \text{under } \rho\end{matrix} \,\middle|\, \begin{matrix}\exists t \in h \text{ provable} \\ \text{under } \rho\end{matrix}\right] < \tilde{O}(r(\log\log n + \log k + \log\log\frac{1}{\delta} + \log\frac{1}{\gamma})(1+\gamma)\epsilon))$ *(Weak Entailment).*

Proof of the Main Theorem

Soundness. We first show that if the implicit abduction algorithm returns an explanation h, then h satisfies parsimony and weak entailment. Plausibility will follow from the assumption that a good explanation exists, so we postpone its discussion to our discussion of completeness, below. We first observe that in the algorithm, we delete all terms with redundant literals, so all outputs are automatically parsimonious.

Each term of the explanation is checked by Implicit Learning, so all terms have low error rates: for $\delta' = \frac{\delta}{2\binom{2n}{\leq k}+4}$,

Claim 18 *For our choice of* $m \geq \frac{12}{\mu\gamma^2} \log \frac{1}{\delta'}$ *we can guarantee that with probability* $1 - \delta/2 + 2\delta'$*, for all terms* t *that pass the second test,* $\Pr[(\vdash t|_\rho) \wedge (\vdash (\neg c)|_\rho)] < \mu\epsilon(1+\gamma)$

Proof of Claim 18 In the Implicit Learning Algorithm, we enumerate through all possible k-DNF terms over n attributes, so there are at most $\binom{2n}{\leq k}$ possible terms. In the algorithm, for every term t that passes the second test $[(\vdash t|_\rho) \wedge (\vdash (\neg c)|_\rho)]$ happens in less than a $\mu\epsilon$-fraction of the examples. By the multiplicative Chernoff bound, when we take enough examples, we will be able to guarantee that $\Pr[\#\{\rho : (\vdash t|_\rho) \wedge (\vdash (\neg c)|_\rho)\} < (1-\gamma/2)(1+\gamma)\mu\epsilon] < \delta'$, i.e., any term with at most $\mu\epsilon$ bad examples has error at most $(1+\gamma)\mu\epsilon$ with high probability. For each term, the Chernoff bound requires $\frac{12}{\mu\gamma^2}\log(\frac{1}{\delta'})$ examples to be correct with probability $1 - \delta'$. We have chosen δ' so that after a union bound over the terms we get $\delta/2 - 2\delta' = \binom{2n}{\leq k}\delta'$. Thus, $m \geq \frac{12}{\mu\gamma^2} \log \frac{1}{\delta'}$ examples suffice. ■

Completeness. We just proved that every output satisfies parsimony and weak entailment with probability $1 - \delta/2 + 2\delta'$. Now, we want to show that if there is an optimal, parsimonious r-term k-DNF explanation h^* satisfying

1. (Plausibility) for a $(1+\gamma)\mu$-fraction of examples, some term $t \in h^*$ is provable, and
2. (Weak Entailment) if some $t \in h^*$ is provable, then with high probability $\neg c$ is not provable

then we are able to find a good solution that satisfies parsimony, plausibility, and weak entailment.

First, pruning with the causal model doesn't compromise completeness. Since the optimal explanation h^* is parsimonious, each of its term contains no redundant literals, so these terms can all pass the causal model pruning. Next, we show the second and third steps also guarantee completeness.

Claim 19 *If there exists a solution* $h^* = t_1^* \vee t_2^* \vee \cdots \vee t_r^*$ *such that [*$\neg c$ *is provable when some* t_i^* *is provable] has probability at most* $(1-\gamma)\mu\epsilon$*, then all these terms* t^* *can pass the second test with probability* $1 - \delta'$.

Proof of Claim 19 We are given that $\Pr[\ [(\vdash t_1^*|_\rho) \wedge (\vdash (\neg c)|_\rho)] \vee \cdots \vee [(\vdash t_r^*|_\rho) \wedge (\vdash (\neg c)|_\rho)]\ ] < (1-\gamma)\mu\epsilon$. By a Chernoff bound, for our choice of m, $[(\vdash t^*|_\rho) \wedge (\vdash (\neg c)|_\rho)]$ happens for any t^* in h^* in less than $\mu\epsilon$-fraction of examples with probability, $1 - \delta'$ so all these terms t^* pass the second test. ■

Next, we show the number of terms r' is controlled, since r' depends upon the solution of the set cover problem.

Claim 20 *If there exists a solution* $h^* = t_1^* \vee t_2^* \vee \cdots \vee t_r^*$ *that satisfies*

- $\Pr[\exists t \in h^* : \vdash t|_\rho] \geq (1+\gamma)\mu$
- $\Pr[\vdash (\neg c)|_\rho \mid \exists t \in h^* : \vdash t|_\rho\ ] < (1-\gamma)\epsilon$

then Implicit Abduction finds an h *using at most* $r' = r\log(\mu m)$ *terms such that* $\#\{\rho : \exists t \in h, \vdash t|_\rho\} > \mu m$*. Furthermore, by a union bound on the error of each term,* $\Pr[(\exists t \in h : \vdash t|_\rho) \wedge (\vdash (\neg c)|_\rho)] < \mu\epsilon(1+\gamma)$*. We thus find that with probability at least* $1 - \delta$ h *satisfies plausibility with* $(1-\gamma)\mu$ *and weak entailment.*

Proof of Claim 20 Following Claims 18 and 19, with probability at least $1 - \delta/2 + \delta'$, all terms t^* in h^* can pass the first and second tests, so they are available for set cover. Moreover, by another Chernoff bound, since at least one of the terms of h^* is provable with probability $(1+\gamma)\mu$ in each example, with probability $1 - \delta'$ at least one of the terms is provable in at least a μ-fraction of the m examples. Thus, there is a set of r terms (the terms of t^*) that pass these tests and indeed cover a μm examples. By Lemma 16, if Opt (h^*) covers μm examples using r sets, then our greedy algorithm can find a cover using $r' = r\log(\mu m)$ sets that also covers μm examples.

Recall that $h = t_1 \vee \cdots \vee t_{r'}$. For each term, by Claim 18, $\Pr[(\vdash t|_\rho) \wedge (\vdash (\neg c)|_\rho)] < \mu\epsilon(1+\gamma)$, so if take an union bound over the terms of h, the error, $\Pr[\exists t \in h (\vdash t|_\rho) \wedge (\vdash (\neg c)|_\rho)]$, is at most $r'\mu\epsilon(1+\gamma)$ in total. If we plug in $r' = r\log(\mu m)$, the resulting error is $O(r\log(\mu m)(1+\gamma)\mu\epsilon)$.

To see that the returned h satisfies plausibility, we consider a Chernoff bound for the fraction of examples in which each possible r'-term k-DNF has a provable term with $\hat{\delta} = \delta/2|T|^{r'}$. So when we take a union bound on all k-DNF explanations, any r'-term explanation found will actually have plausibility $(1-\gamma)\mu$ with probability $1 - \delta/2$. Therefore, it suffices to have

$$m \geq \frac{3}{\mu\gamma^2} \log \frac{2|T|^{r'}}{\delta} \quad \text{or} \quad m \geq \frac{3r\log(\mu m)}{\mu\gamma^2} \log \frac{2|T|}{\delta}.$$

Here we apply the inequality

Lemma 21 *For* $a \geq 1$*, if* $x \geq 2a\log a$*, then* $x \geq a\log x$.

By plugging in $x = \mu m$ and $a = \frac{3r}{\gamma^2}\log\frac{2|T|}{\delta}$, we get $m \geq \frac{6r}{\gamma^2\mu}\log(\frac{2|T|}{\delta})\log(a)$ examples suffice. Here, $\log a$ is dominated by other terms, so we get $m = \tilde{O}(\frac{r}{\gamma^2\mu}\log\frac{n^k}{\delta})$.

Since we condition on some $t \in h$ provable and $\Pr[\exists t \in h$ provable under $\rho] > (1-\gamma)\mu$,

$$\begin{aligned}
&\Pr[\vdash (\neg c)|_\rho \mid \exists t \in h : \vdash t|_\rho\ ] \\
&\quad = \Pr[(\exists t \in h \vdash t|_\rho) \wedge (\vdash (\neg c)|_\rho)] / \Pr[\exists t \in h : \vdash t|_\rho] \\
&\quad < O(r\log(\mu m)(1+\gamma)\mu\epsilon/\mu) \\
&\quad = O(r\log(\mu m)(1+\gamma)\epsilon)
\end{aligned}$$

and thus, we indeed find an h satisfying weak entailment with the claimed error rate with probability $1 - \delta$. ■

Finally, when we plug in $m = \tilde{O}(\frac{r'}{\mu\gamma^2}\log\frac{3(2n)^k}{\delta})$,

$$\begin{aligned}
O(r\log(\mu m)(1+\gamma)\epsilon) &= \tilde{O}(r\log(\frac{\mu r}{\mu\gamma^2}\log\frac{n^k}{\delta})(1+\gamma)\epsilon) \\
&= \tilde{O}(r(\log\log n + \log k + \log\log\frac{1}{\delta} + \log\frac{1}{\gamma})(1+\gamma)\epsilon)
\end{aligned}$$

We conclude that $\Pr[\vdash (\neg c)|_\rho \mid \exists t \in h : \vdash t|_\rho\ ] < \tilde{O}(r(\log\log n + \log k + \log\log\frac{1}{\delta} + \log\frac{1}{\gamma})(1+\gamma)\epsilon))$ with probability $1 - \delta$.

Running Time. We note that the algorithm of Geiger et al. runs in time proportional to the number of edges of the causal model, which is $O(n^2)$, and identifies all of the attributes that are d-separated from c by t'. We run this algorithm for all of the $O(n^{k-1})$ terms t' of at most $k-1$ literals. Thus, overall the pruning step runs in time $O(n^{k+1})$. The second test is run for each of the surviving terms of size at most k, of which there may be $\sim n^k$. For each such term, DecidePAC runs in time $O(T(n, |\varphi|, |KB|)\frac{1}{\gamma^2}\log\frac{1}{\delta})$; as we are given that we have chosen our proof system so that this is a polynomial, the overall running time is also polynomial, as needed. ■

Extensions and Directions for Future Work

We note that our algorithm can be parallelized. Although the greedy algorithm at the heart of our algorithm is sequential, the computation of the greedy choice can certainly be parallelized well, and in the case of most interest, the size of the cover is small (so there are few rounds). Alternatively, we note that Bateni et al. (2016) considered parallel approximation algorithms for the partial set cover problem running in four MapReduce rounds and achieving an approximation ratio of $(1+\epsilon)\log 1/(1-\mu)$. This could be used to obtain a good parallel algorithm overall.

There are two main problems left untouched by this work. The first is that we still do not know how our error rate should depend on either the sparsity or the number of sets. It could be that the improvement due to exploiting sparsity of the unknown k-DNF is not inherent. But, without some kind of inapproximability bounds, we cannot resolve this.

The second concerns the formulation of the task. We required that for the ideal explanation, some *individual term* of h^* should be provable. A more relaxed condition that could be reasonable would be to only require that h^* is provable in its entirety. The challenge is now to identify such a hypothesis h when we cannot even rely on knowledge of its terms. Certainly the kind of greedy covering technique we used here does not work, but it is again consistent with the state of our understanding that such an algorithm could exist. This would be of interest, as it would allow the algorithms to discover hypotheses in cases where the kind of algorithms we have proposed here will fail on account of having insufficient information.

Acknowledgements

B. Juba and E. Miller were partially supported by an AFOSR Young Investigator Award.

References

Bateni, M.; Esfandiari, H.; and Mirrokni, V. S. 2016. Distributed coverage maximization via sketching. *arXiv:1612.02327 [cs.DS]*.

Bylander, T.; Allemang, D.; Tanner, M. C.; and Josephson, J. R. 1991. The computational complexity of abduction. *Artificial Intelligence* 49:25–60.

Charniak, E., and McDermott, D. 1985. *Introduction to Artificial Intelligence*. Reading, MA: Addison-Wesley.

Cox, P., and Pietrzykowski, T. 1986. Causes for events: their computation and applications. In *Proc. 8th Int'l Conf. Automated Deduction*, 608–621.

Feige, U. 1998. A threshold of ln n for approximating set cover. *Journal of the ACM (JACM)* 45(4):634–652.

Geiger, D.; Verma, T. S.; and Pearl, J. 1989. d-Separation: from theorems to algorithms. In *Proc. 5th UAI*, 118–124.

Haussler, D. 1988. Quantifying inductive bias: AI learning algorithms and Valiant's learning framework. *Artificial Intelligence* 36:177–221.

Hobbs, J.; Stickel, M.; Appelt, D.; and Martin, P. 1990. Interpretation as abduction. Technical Report 499, SRI, Menlo Park, CA.

Juba, B. 2013. Implicit learning of common sense for reasoning. In *Proc. 23rd IJCAI*, 939–946.

Juba, B. 2016. Learning abductive reasoning using random examples. In *Proc. 30th AAAI*, 999–1007.

Khardon, R., and Roth, D. 1997. Learning to reason. *J. ACM* 44(5):697–725.

Levesque, H. J. 1986. Making believers out of computers. *Artificial Intelligence* 30(1):81–108.

McIlraith, S. A. 1998. Logic-based abductive inference. Technical Report KSL-98-19, Knowledge Systems Laboratory.

Michael, L. 2010. Partial observability and learnability. *Artificial Intelligence* 174(11):639–669.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, J. 2009. *Causality*. Cambridge University Press, second edition.

Poole, D. 1990. A methodology for using a default and abductive reasoning system. *Int'l J. Intelligent Sys.* 5:521–548.

Poole, D. 1993. Probabilistic Horn abduction and Bayesian networks. *Artificial Intelligence* 64(1):81–129.

Reiter, R., and de Kleer, J. 1987. Foundations for assumption-based truth maintenance systems: Preliminary report. In *Proc. AAAI-87*, 183–188.

Rubin, D. B. 1976. Inference and missing data. *Biometrika* 63(3):581–592.

Slavík, P. 1997. Improved performance of the greedy cover algorithm for partial cover. *Information Processing Letters* 64(5):251–254.

Valiant, L. G. 1984. A theory of the learnable. *Communications of the ACM* 18(11):1134–1142.

Valiant, L. G. 2000. Robust logics. *Artificial Intelligence* 117:231–253.

Zhang, M.; Mathew, T.; and Juba, B. 2017. An improved algorithm for learning to perform abduction. In *Proc. 31st AAAI*, 1257–1265.