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

1 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 from Hobbs et al. (1990) and images from Cox and Pietrzykowski (1986) and 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).

In this work, we consider a *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 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 work extends Juba's formulation of the abduction task to use partial examples and draw on declaratively specified background knowledge. We observe that by using a covering algorithm, it is possible to guarantee significantly better explanations when a small hypothesis (using relatively few terms) is adequate. 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, Mathew, and Juba (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).

2 Preliminaries

We work in a standard machine learning model in which the data consists of many *examples,* assigning Boolean values to a variety of *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.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2.1 Partial Observability

In the real world, it is hard to require each example to contain all of the attributes. So, we want to make inferences with incomplete data. *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 $[x_1 = 1,\ x_2 = *,\ \cdots,\ x_n = 0]$. (For convenience, we denote $\rho^{(i)}$ to be the ith example and ρ_i to be the ith coordinate of an example ρ.) In our abduction task, we say our partial examples are drawn from such a masked distribution $M(D)$.

2.2 Implicit Learning

The main tool to deal with partial observability is *implicit learning*. 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 1 (Witnessed Formula) *Given a partial example ρ, we say a formula ϕ is* witnessed *if $\phi|_\rho$ is 0 or 1, where $\phi|_\rho$ means the formula ϕ restricted to example ρ.*

Formally, a restricted formula $\phi|_\rho$ is defined recursively: we break down a formula at its logical connectives ($\{\neg, \vee, \wedge\}$) recursively up to single variables, where for the base cases, these singletons are set by the values from the example ρ. Restricted formulas' explicit expressions can be computed in linear time.

Informally speaking, we get the restricted formula by plugging in the observed value of the given example, resulting in a (shorter) formula of the unobserved variables. For instance, let $\phi = x_1 \vee x_2$, and in a partial example, $x_1 = 1; x_2 = *$. Then ϕ is witnessed (true) even though x_2 is not observed. But it could be hard to determine the value for more complex formulas; in general, this may be as hard as deciding if the formula is a tautology, which is NP-hard. Notice that each formula can be either witnessed true, witnessed false, or not witnessed.

Proof system Given a knowledge base KB (a set of formulas) and a query formula ϕ, for our purposes a proof system is an algorithm that can determine whether we can derive ϕ. If we can derive ϕ, we say ϕ is *provable* and denote it as $\vdash \phi$.

A proof system is *restriction-closed* if whenever there is a proof of a formula ϕ, there is also a proof of $\phi|_\rho$ for any partial assignment ρ. In general, if there is a proof of $\phi|_\rho$ from $KB|_\rho$, we say that ϕ *is provable from KB under ρ*. The formal language may be confusing, but the definition is indeed intuitive. Consider the following example: $\psi_1 = x_1 \wedge x_2$, $\psi_2 = x_3 \wedge x_4$, $\phi = x_1 \wedge x_2 \wedge x_3 \wedge x_4$, $\{\psi_1 \wedge \psi_2\} \vdash \phi$. 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, $\{\psi_1|_\rho \wedge \psi_2|_\rho\} \vdash \phi|_\rho$.

Notice that most common propositional proof systems such as Resolution, (Forward) Chaining, Cutting Planes, and Polynomial Calculus are indeed restriction closed.

2.3 DecidePAC Algorithm

Besides the information we directly witness from the examples, we want to know further what we can infer, given some knowledge base (a set of additional formulas). From the previous work by Juba (2013), we have an algorithm that can tell whether a formula is provable or not. 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 $M(D)$; otherwise, if [KB$\Rightarrow \phi$] is not true with probability at least $(1 - \epsilon - \gamma)$, then DecidePAC will $reject$ formula ϕ.

Notice that 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 if we assume in

KB we have $x_1 \Rightarrow \neg x_2$, then 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. We want to bridge from the witnessed values of examples to their ground truth, through logical inference.

DecidePAC was analyzed by Juba using an additive Chernoff bound. We can obtain an analogous *multiplicative* $(1 \pm \gamma)$ guarantee by instead using the multiplicative Chernoff bound:

Lemma 2 (Multiplicative Chernoff Bound) *Let $X_1, \ldots, X_m$ be independent random variables taking values in $[0,1]$, such that $E[\frac{1}{m}\sum_i X_i] = p$. Then for $\gamma \in [0,1]$,*

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

3 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.

Definition 3 (Partial Information Abduction) *For any fixed proof system, abduction is the following task: given any query formula c and independent partial examples $\{\rho^{(1)}, \ldots, \rho^{(m)}\}$ 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. $\Pr[\neg c \text{ provable under } \rho | \exists t \in h : t \text{ provable under } \rho] \leq \epsilon$ (Weak Entailment)

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 $\ell_{i_j} \in t_i$.

3.1 Choice of Formulation

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 the notion of *vivid knowledge* by 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.

We made the opposite decision for $c(x) = 1$, relaxing it to the condition that $\neg c$ is not provable. The main reason for this choice is that we wish to not penalize a good h if it is often impossible to check whether or not c holds. We use this liberal notion of entailment for our explanations because the intended semantics of the task is merely to propose *possible* causes given some tentative partial knowledge of the world, perhaps to guide further investigation. 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.

4 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.

We observe 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.

Algorithm 1: Implicit Abduction

input : Knowledge base KB, , 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 $m = \frac{6}{\mu\gamma^2} \log \frac{2|T|^r}{\delta} \log(\frac{3}{\gamma^2} \log(\frac{2|T|^r}{\delta}))$
1. **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.
2. Run **greedy algorithm** for set cover:
use terms in T to cover a μ-fraction of the examples. Get $\{t_1, \cdots, t_r\}$.
$h \leftarrow t_1 \vee \cdots \vee t_r$
return h.

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

1. Check all the 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 the terms that pass the test then satisfy weak entailment: the error condition $[\vdash t$ and $\vdash \neg c]$ has probability at most $\mu\epsilon(1 + \gamma)$.
2. Then use the greedy algorithm to choose an explanation. If the algorithm can find an explanation covering a μ-fraction of examples, then we can argue explanation has probability larger than μ by the Chernoff bound.

Thus, if there exists a good explanation, we can find an 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.

Theorem 4 (Implicit Abduction) *Given a query c, partial examples $\rho^{(1)}, \cdots, \rho^{(m)}$ from a masked distribution $M(D)$, and an efficient restriction-closed proof system with knowledge base KB, for constant k:*

If there exists a 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 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. $\Pr[\neg c$ *provable under* $\rho | \exists t \in h$ *provable under* $] < \tilde{O}(r(\log\log n + \log k + \log\log\frac{1}{\delta} + \log\frac{1}{\gamma})(1 + \gamma)\epsilon))$ *(Weak Entailment).*

4.1 Proof of the Main Theorem

Soundness. We first show that if the implicit abduction algorithm returns an explanation h, then h satisfies weak entailment. Plausibility will follow from the assumption that a good explanation exists, so we postpone its discussion to our discussion of completeness, below.

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 5 *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 first test, $\Pr[(\vdash t|_\rho) \wedge (\vdash (\neg c)|_\rho)] < \mu\epsilon(1+\gamma)$*

Proof of Claim 5 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 first 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 weak entailment with probability $1-\delta/2+2\delta'$. Now, we want to show if there is an optimal 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 plausibility and weak entailment.

Claim 6 *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 first test with probability $1-\delta'$.*

Proof of Claim 6 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 first test. ■

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

Claim 7 *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 7 Following Claims 5 and 6, with probability at least $1-\delta/2+\delta'$, all terms t^* in h^* can pass the first test, 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. For the greedy algorithm, 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 Slavík (1997).

Recall that $h = t_1 \vee \cdots \vee t_{r'}$. For each term, by Claim 5, $\Pr[(\vdash t|_\rho) \wedge (\vdash (\neg c)|_\rho)] < \mu\epsilon(1+\gamma)$, 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 8 *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 $\Pr[\exists t \in h \text{ provable under } \rho] > (1-\gamma)\mu$,

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

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. The test is run for each term of size at most k, of which there may be $\sim n^k$. And DecidePAC runs in time polynomial in n, $|\varphi|$, $|KB|$ (the running time of the underlying algorithm for the proof system), and $\frac{1}{\gamma^2}\log\frac{1}{\delta}$; the overall running time is also polynomial, as needed. ■

Acknowledgements

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

References

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.

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.

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

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

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.