Learning abductive reasoning using random examples

Brendan Juba*
Washington University in St. Louis
bjuba@wustl.edu

Abstract

We consider a new formulation of *abduction*. Our formulation differs from the existing approaches in that it does not cast the "plausibility" of explanations in terms of either syntactic minimality or an explicitly given prior distribution. Instead, "plausibility," along with the rules of the domain, is *learned* from concrete examples (settings of attributes). Our version of abduction thus falls in the *"learning to reason"* framework of Khardon and Roth. Such approaches enable us to capture a natural notion of "plausibility" in a domain while avoiding the problem of specifying an explicit representation of what is "plausible," a task that humans find extremely difficult.

In this work, we specifically consider the question of which syntactic classes of formulas have efficient algorithms for abduction. It turns out that while the representation of the *query* is irrelevant to the computational complexity of our problem, the representation of the *explanation* critically affects its tractability. We find that the class of k-DNF explanations can be found in polynomial time for any fixed k; but, we also find evidence that even very weak versions of our abduction task are intractable for the usual class of *conjunctive* explanations. This evidence is provided by a connection to the usual, inductive PAC-learning model proposed by Valiant. We also briefly consider an exception-tolerant variant of abduction. We observe that it is possible for polynomial-time algorithms to tolerate a few adversarially chosen exceptions, again for the class of k-DNF explanations. All of the algorithms we study are particularly simple, and indeed are variants of a rule proposed by Mill.

1 Introduction

Abduction is the process of passing from an observation to a plausible explanation or diagnosis. For example, to understand a story in which a man is holding a gun in a bank, one must "abduce" that (perhaps) the man wishes to rob the bank. This is not a sound inference, of course – the man could be a guard, the man could be seeking to place the firearm in a safe deposit box, etc. – but it represents at least a highly plausible explanation for the given facts. Unlike the usual forms of inference of deduction and induction, abduction as a form of inference was only brought to prominence relatively recently, by Pierce [1931]. It was then promoted as a core task in AI by Charniak and McDermott [1985]. It has since been observed that problems as diverse as diagnosis [Reggia, 1983; Reiter, 1987], image understanding [Cox and Pietrzykowski, 1986; Poole, 1990], natural language understanding [Hobbs *et al.*, 1990], planning [Eshghi, 1988; Missiaen *et al.*, 1995], and plan recognition [Charniak and McDermott, 1985] all involve abduction. We will discuss applications slightly further in Section 5.

As we will review, the task of abduction itself has already been formalized in (at least) three distinct ways, and we propose a new formalization of abduction using *examples*; the examples here consist of settings of the various attributes, e.g., encoding a concrete "scene" or "episode" in the image units of Valiant's neuroidal model [2000a], or more abstractly, as in the entries in a database. We assume these examples to have been drawn independently from a common unknown distribution D, modeling the frequency with which such scenes occur in a domain. We formulate the task as searching for a *conditional distribution*: in addition to this data, we are given a Boolean query that we wish to "explain" in a sense we will elaborate on below, and a class of Boolean formulas $\mathcal{C}$ over a distinguished set of attributes of the data A. A indicates the attributes that we wish to allow in our explanations, for example, attributes that took values prior to the condition to be explained, and might therefore be predictive. We then seek to find a formula c in the class $\mathcal{C}$ using only the attributes in A such that

(i) *Approximate validity:* the query is (almost) always true on examples drawn from the conditional distribution $D|c$, i.e., the distribution over assignments induced by D given that c is satisfied, and

(ii) *Plausibility:* the probability of c being true under D is at least some minimum value μ. We will often seek to find a c attaining a maximum bound μ on its plausibility.

*Work partially performed while the author was affiliated with Harvard University and supported by ONR grant number N000141210358. Currently supported by an AFOSR Young Investigator award.

So, c is an "explanation" in the sense that the query empirically follows from c, and c holds sufficiently often. For example, in our model, the query might indicate whether or not the key facts of the story – the gun in the bank – are present in the examples, and the property c to be found represents the desired "explanation." We can refer directly to notions of (approximate) "validity" and "entailment" in this model because we assume that we have completely specified examples, on which the various formulas can be evaluated directly, much as in model-based reasoning [Kautz *et al.*, 1995].[1] We stress that all of the entailment constraints of the domain are thus properties of the distribution D, and are therefore (only) represented implicitly in the problem formulation by the examples drawn from D.

In other words, this entailment relation underlying our abductive reasoning task must be *learned* from the examples drawn from D. We seek computationally efficient algorithms with a "PAC" guarantee: that with high *probability* (over the examples) both conditions are *approximately* satisfied. (We will define the task more formally in Section 2.) Our model thus belongs to the *learning to reason* framework of Khardon and Roth [1997b]. Indeed, Khardon and Roth suggested in that work that such a learning formulation of abduction was possible, although they did not actually present a formalization of abduction in their framework. In a later work, Roth [1996] also briefly indicated that various aspects of an abduction task could be carried out in neural networks (and thus learned) but again did not elaborate on the semantics of the task. Similarly, in their work on model-based reasoning, Kautz et al. [1995] likewise both considered model-based abduction and indicated that model-based reasoning could use random examples, but again did not actually formally specify the semantics of the task. Our work is thus (perhaps surprisingly) the first to explicitly consider this formulation of abduction.

It is already widely appreciated that learning is a highly effective alternative to explicit knowledge engineering. Indeed, machine learning (e.g., from examples) has been far more effective than traditional knowledge engineering at acquiring robust representations across a variety of domains and tasks. Relatedly, Valiant [1994; 2000a] argued that learned representations should enable systems to better cope with an open world, and thus learning should be used as a basis for robust cognition. But, the main motivation for treating abduction itself as a learning task as we do is that it provides a means of efficiently capturing natural, domain-specific notions of the "plausibility" of explanations.

In particular here, we avoid the need to explicitly represent a (prior) distribution. Both the probabilistic version of the "set covering" model of abduction presented by Bylander et al. [1991] and the models of abduction based on probabilistic graphical models [Pearl, 1988; Poole, 1993] were Bayesian models that depended on assigning some such *prior probabilities* to the various *explanations.* In these models, "plausibility" of the proposed conditions is evaluated in terms of these probabilities. The need to estimate such a prior was deemed to be one of the main drawbacks of these previous probabilistic models of abduction [McIlraith, 1998], as good priors are hard to estimate. In this way, our formulation thus differs crucially from these previous models. This is also how our formulation differs from, for example that of Hobbs et al. [1990], in which explicit weights or costs (without necessarily having a probabilistic interpretation) are attached to the various literals to be used in an explanation.[2]

Moreover, for the non-probabilistic logic-based or logic programming [Denecker and Kakas, 2002] approaches, the usual syntactic criteria, such as minimizing the number of literals as done in ATMS [Reiter and de Kleer, 1987], appears to serve essentially a proxy for some other kind of unspecified domain-specific "plausibility" notion, by appealing to something like Occam's razor. McIlraith [1998] nicely discusses the problems that may arise with these approaches, for example they are highly representation-dependent. Previous works on combining learning and abduction simply used such syntactic minimization criteria for plausibility [Thompson and Mooney, 1994; Flach and Kakas, 2000]. Another probabilistic formulation, proposed by Bacchus et al. [1996] proposed to use a maximum-entropy distribution over the attributes as a prior, which is essentially similar to these syntactic criteria. In particular, it is also representation language dependent and may be simply inappropriate.

1.1 Our results

The main question we consider in this work is, for which classes of formulas $\mathcal{C}$ do efficient algorithms abduce explanations in our new model?[3] We find that a particularly simple algorithm abduces disjunctive explanations. This simple algorithm is easily generalized to k-DNF explanations (for constant k). We further generalize this algorithm to provide some weak *"exception tolerance"*: if the best disjunctive explanation only gives the query conditional probability $1-\epsilon$ for some $\epsilon > 0$, then the exception-tolerant algorithm finds a disjunction that gives the query conditional probability $1-O(n\epsilon)$ when there are n attributes in the vocabulary (or probability $1-O(n^k\epsilon)$ for k-DNF explanations by an analogous modification). That is, the probability of counterexamples to the "explanation" we find may be $O(n)$ times greater than that of the best possible explanation.[4] Thus, we see that this new abductive reasoning task is feasible for some natural classes of explanations, even in the presence of some noise.

[1] We note that the role of *proofs* in such models is that they serve as a means to decide entailment under incomplete information. We leave the extension of this model to partial information as a subject for future work.

[2] Hobbs et al. briefly suggest that an interpretation in terms of conditional probabilities might be used to obtain such weights, but the assignment of specific weights to the literals is problematic unless they refer to events that are for example either disjoint or uncorrelated.

[3] We do not limit the class of representations that the query is drawn from, apart from assuming that it can be evaluated efficiently on an example. The complexity of the query representation appears to be largely irrelevant in this model.

[4] Although this $O(n)$ increase is indeed somewhat large, we stress that it should be contrasted with the state-of-the-art in such exception-tolerant supervised learning of disjunctions, which similarly suffers a $O(n^{1/3})$ blow-up of the error [Awasthi *et al.*, 2010].

We also stress that both the algorithms and the query representations we use in these results are particularly simple, and interpretable by humans. In particular, the algorithms, which essentially eliminate terms when they encounter (too many) bad examples for these terms, follow a classical human strategy for identifying hypotheses proposed by Mill [1843, Book III, Chapter 8]. We therefore view our algorithms and the representations we produce as being cognitively plausible, although our model did not strictly require it.

On the other hand, we find that abducing the usual, *conjunctive* explanations is likely to be intractable, even if we allow the explanation to be expressed by a richer representation: Any algorithm that finds explanations whenever a conjunctive explanation exists would yield an algorithm for PAC-learning DNF formulas in the standard PAC-learning model [Valiant, 1984]. This has been the central open problem in PAC-learning since the model was first proposed by Valiant [1984], and recent results by Daniely et al. [2014] and Daniely and Shalev-Shwartz [2014] show that the problem may be intractable, given a new assumption about the hardness of refuting random k-SAT instances (stronger than Feige's assumption [2002]). Since most of the usual classes of representations can either be expressed by k-DNF formulas or can themselves express conjunctions, this result together with our algorithms essentially settles the question of which representations are tractable in our model.

2 Abduction for disjunctive explanations

In this work, we are seeking to find explicit representations of "explanations" of possibly low, but non-negligible probability for which conditioned on the corresponding event, some given *query* property to be *explained* or *diagnosed* is (almost) always satisfied. For example, we may have various attributes about the state of a car, and the query may be something like $\mathsf{key\_turned} \wedge \neg\mathsf{engine\_running}$, for which we may seek an explanation such as $\mathsf{key\_turned} \wedge \neg\mathsf{gas\_in\_tank}$: although $\Pr[\mathsf{key\_turned} \wedge \neg\mathsf{gas\_in\_tank}]$ may be low, it occasionally may happen, and surely $\Pr[\mathsf{key\_turned} \wedge \neg\mathsf{engine\_running} | \mathsf{key\_turned} \wedge \neg\mathsf{gas\_in\_tank}] = 1$. Our probability distributions will not be given explicitly, but instead will be represented by examples drawn from the distributions in question. Formally, we focus on the following class of problems:

Definition 1 (Abduction) *For a* representation class $\mathcal{C}$ *of Boolean formulas over propositional attributes* $x_1, \ldots, x_n$*, the (proper)* abduction *task is as follows. We are given as input* m *independent examples* $x^{(1)}, \ldots, x^{(m)}$ *from an arbitrary distribution* D *over* $\{0,1\}^n$ *(assignments to the* n *attributes), a* query *formula* ψ *over* $x_1, \ldots, x_n$*, and an* alphabet $A \subseteq \{x_1, \ldots, x_n\}$*, for which there exists* $c^* \in \mathcal{C}$ *only using attributes in* A *such that* $\Pr[\psi(x) = 1 | c^*(x) = 1] = 1$ *and* $\Pr[c^*(x) = 1] \geq \mu$*. Then, with probability* $1 - \delta$*, find some* explanation $c \in \mathcal{C}$ *only using attributes in* A *such that*

1. $\Pr[\psi(x) = 1 | c(x) = 1] \geq 1 - \epsilon$ *and*
2. $\Pr[c(x) = 1] \geq 1/p(1/\mu, n, 1/(1-\epsilon))$ *for a polynomial* p*,*

in time polynomial in $n, 1/\mu, 1/\epsilon$*, and* $1/\delta$*.*

As with PAC-learning, we will usually obtain running times that only depend polynomially on $\log 1/\delta$ rather than $1/\delta$ (but in general we might be satisfied with the latter). Furthermore, we could consider an *"improper"* version of the problem, finding representations from some larger, possibly more expressive class than the $\mathcal{C}$ containing the "optimal" condition c^*. The form of the representation is naturally important for some applications (we will discuss an example in Section 5), though, and so in this work we focus primarily on the proper version of the problem.

We formulated the second condition to include a notion of approximation. Our positive results, for conjunctions and k-DNFs *do not* require such a notion of approximation – whenever a condition c^* that is satisfied with probability at least μ exists, our algorithms will actually find a condition c that is also satisfied with probability at least μ. But, the value of such an abstract definition is largely in the power it grants to establish *negative* results, and here we would like the broadest possible definition. Indeed, we will see in the sequel that for some extremely simple representations – specifically, conjunctions – the abduction task is unfortunately unlikely to have efficient algorithms, even in this very liberal sense.

Valiant [1984] gave an analysis showing that the "elimination algorithm" (Algorithm 1), a simple method essentially proposed by Mill [1843, Book III, Chapter 8], is a PAC-learning algorithm for disjunctions. We note that the same algorithm also can be used to identify a disjunctive explanation (in our sense) quite efficiently. As disjunctions are themselves a rather natural, simple kind of knowledge representation, this result establishes (moreover) that our new model captures cognitively plausible algorithms for finding cognitively natural representations of explanations.

input : Examples $x^{(1)}, \ldots, x^{(m)}$, query ψ, alphabet A.
output: A disjunction of literals over A.
begin
  Initialize c to be the disjunction over all literals on the alphabet A.
  for $i = 1, \ldots, m$ **do**
    if $\psi(x^{(i)}) = 0$ **then**
      forall the $\ell \in c$ **do**
        if $\ell(x^{(i)}) = 1$ **then**
          Remove ℓ from c.
        end
      end
    end
  end
  return c.
end

Algorithm 1: Elimination algorithm

Theorem 2 *The abduction task for disjunctive explanations can be solved by the elimination algorithm using* $O(\frac{1}{\mu\epsilon}(n + \log 1/\delta))$ *examples, obtaining a disjunction* c *with* $\Pr[c] \geq \mu$*.*

Proof: For the given sample size, it is immediate that the elimination algorithm runs in polynomial time. We thus need only establish correctness.

Initially every literal on A is contained in c, so c contains all of the literals of c^*. We claim that this invariant is maintained: since $\Pr[\psi|c^*] = 1$, whenever $\psi(x^{(i)}) = 0$, it must be that every literal of c^* is falsified on $x^{(i)}$. Thus, these literals are not removed during any iteration, so they are included in the final disjunction c. It therefore follows that since $\Pr[c^*] \geq \mu$ and $\{x : c^*(x) = 1\} \subseteq \{x : c(x) = 1\}$, $\Pr[c] \geq \mu$.

We now bound the probability that the algorithm terminates with a c such that $\Pr[\psi|c] < 1-\epsilon$. We note that for such c, $\Pr[\neg\psi \wedge c] > \epsilon \Pr[c]$. For any c with $\Pr[c] \geq \mu$, it follows that $\Pr[\neg\psi\wedge c] > \mu\epsilon$. Thus, after $m = \frac{1}{\mu\epsilon}(n \ln 3 + \ln 1/\delta)$ examples, the probability that such a c is always falsified when $\psi(x) = 0$ for all of them is at most $(1-\mu\epsilon)^m \leq e^{-\ln(3^n/\delta)} = \delta/3^n$ so by a union bound over the (at most) 3^n such disjunctions, we find that the probability of any of them surviving is at most δ. Since the c we output is guaranteed by construction to be false for every example where $\psi(x) = 0$, we find that with probability $1 - \delta$ $\Pr[\psi|c] \geq 1 - \epsilon$ as needed. ■

We remark that the algorithm extends to solve the abduction task for k-DNF formulas with sample complexity $O(\frac{1}{\mu\epsilon}(n^k + \log 1/\delta))$ in the usual way, by replacing the individual literals in the elimination algorithm with all possible conjunctions of size up to k on A.

For our example problem of generating a k-DNF diagnosis for key_turned $\wedge$ $\neg$engine_running given a set of example situations, the Elimination algorithm simply rules out all of the terms which were true in examples where either key_turned was false or engine_running was true. For $k \geq 2$, this will yield a disjunction of terms that includes the term key_turned $\wedge$ $\neg$gas_in_tank possibly among some others that are either possible explanations or have never occurred in our examples. The disjunction of such terms is our maximally plausible condition.

We note that we could modify the algorithm to also eliminate such suprious terms that never occurred in any example; so, in our example scenario, this will only leave those terms that have been observed to be satisfied when both key_turned and $\neg$engine_running hold. Then as long as this modified algorithm is provided with $O(\frac{n^k}{\mu\epsilon}(k \log n + \log 1/\delta))$ examples, it still returns explanations with plausibility at least $(1-\epsilon)\mu$. Indeed, letting N ($\leq (ne/k)^k$) denote the number of terms of size at most k, if a term is true with probability greater than $\mu\epsilon/N$, then $\frac{N}{\mu\epsilon}(\log N + \log 2/\delta)$ examples only fail to include an example of such a term being satisfied with probability at most $\delta/2N$. Thus, with probability $1-\delta/2$, the only terms of the optimal c^* that may be deleted are only individually satisfied with probability $\epsilon\mu/N$—in aggregate, with probability at most $\epsilon\mu$. Hence with probability at least $(1-\epsilon)\mu$ c^* must be satisfied by some other term that is included in the returned c. The rest of the analysis is the same. We note that the algorithm is still rather simple – we merely now disregard the terms that never occur on any examples – and so it, like the basic elimination algorithm, is natural and easy for humans to grasp.

3 Exception-tolerant abduction

In this section, we consider a variant of our basic model in which no explanation entails the target condition with conditional probability 1. For example, we may be looking for an explanation for the observation that "Tweety flies," but no representation in our class may possibly perfectly explain "flying"—we may have that 99.99% of the birds we have observed fly, but as this is less than 100%, our earlier method cannot propose flying if we have seen enough examples of birds, since an encounter with a counterexample causes terms of the desired explanation to be deleted. We would like an alternative framework that allows such slightly imperfect explanations.

We will *not* assume any additional simple structure on the errors, e.g., that they are the result of "independent noise." This "agnostic" formalization captures the philosophy that the world may actually be described precisely by very complicated rules, but we wish to merely find an explanation that is often sufficient. This is one possible PAC-learning style approach to solutions to the qualification problem in common sense reasoning [McCarthy, 1980] (see works by Roth [1995] and Valiant [1994; 1995; 2006] for more on this aspect).

We observe that a simple variant of the elimination algorithm achieves a weak kind of exception tolerance: suppose that we only delete the literals that are true on more than $8\mu\epsilon m$ examples where ψ is false out of the m total examples. (8 is a convenient constant larger than 1.) Then:

Theorem 3 *If there is a disjunction c^* with probability at least $\mu/4$ and at most 4μ that gives ψ conditional probability $1-\epsilon$, then given $\Omega(\frac{1}{\mu\epsilon}(\log \frac{n}{\delta}))$ examples, we find that the above ϵ-tolerant elimination algorithm obtains a disjunctive explanation with probability at least that of c^* of being satisfied, under which ψ is true with probability $1 - O(n\epsilon)$.*

So, for example if only $\sim 10^{-4}$ of the birds we see do not fly and we have a vocabulary of $\sim 10^3$ attributes, then we can obtain an explanation such as "Tweety is a bird" for "Tweety flies" that is good except with probability $\sim 10^{-1}$. Of course we would like a better dependence on the size of our vocabulary, for example matching the dependence of $n^{1/3}$ achieved by Awasthi et al. [2010] for agnostic PAC-learning of disjunctions; we suggest this as a natural direction for future work. Nevertheless, we again find that in our new formulation of abduction, a rather simple algorithm finds the same simple kind of explanations, but now moreover featuring some robustness to rare counterexamples, as might occur in a complex open world.

The proof will require the Chernoff bound:

Theorem 4 (Chernoff bound) *Let $X_1, \ldots, X_m$ be independent random variables taking values in $[0,1]$, such that $\mathbb{E}[\frac{1}{m}\sum_i X_i] = p$. Then for $\gamma \in [0,1]$,*

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

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

Proof of Theorem 3: We first observe that for the "ideal" disjunction c^* for which $\Pr[\psi|c^*] \geq 1-\epsilon$ and $4\mu \geq \Pr[c^*] \geq$

$\mu/4$, $\Pr[\neg\psi \wedge c^*] \leq \Pr[c^*]\epsilon \leq 4\mu\epsilon$. So, no literal of c^* may be true when ψ is false with probability greater than $4\mu\epsilon$. Given more than $\frac{3}{4\mu\epsilon} \ln \frac{2n}{\delta}$ examples, it follows from the Chernoff bound that the probability that any one of these (at most n) literals is true when ψ is false in more than a $8\mu\epsilon$ fraction of the examples is at most $\frac{\delta}{2n}$.

At the same time, taking $\gamma = 1/2$, we find that the probability that any literal ℓ for which $\Pr[\neg\psi \wedge \ell] \geq 16\mu\epsilon$ remains in our disjunction after $\frac{8}{16\mu\epsilon} \ln \frac{2n}{\delta}$ examples is also at most $\frac{\delta}{2n}$. Noting that no literal ℓ can give $\neg\psi \wedge \ell$ probability both greater than $16\mu\epsilon$ and less than $4\mu\epsilon$ simultaneously, we can simply take a union bound over the appropriate event for all $2n$ literals to find that the overall probability of any occurring is at most δ. When none occur, the algorithm obtains a disjunction c that contains all of the literals of c^* – and so $\Pr[c] \geq \Pr[c^*] \geq \mu/4$ – and moreover (by a union bound over the literals) gives $\Pr[\neg\psi \wedge c] \leq 16n\mu\epsilon$. Hence, for this c, $\Pr[\neg\psi|c] \leq 64n\epsilon$. Thus, $\Pr[\psi|c] \geq 1 - 64n\epsilon$. ■

Obtaining a value for μ. Although we have placed a stronger condition on c^* – we now require an *upper* bound on its probability in addition to a lower bound – we now argue that we can find a satisfactory μ by repeatedly running the algorithm as follows. We can also use the Chernoff bound to *verify* that μ was sufficiently small that the condition c we have found satisfies, e.g., $\mu/4 \leq \Pr[c]$ whenever the stronger condition $\mu \leq \Pr[c]$ actually holds. We achieve this by checking to see that c satisfies at least a $\mu/2$-fraction out of $m \geq \frac{16}{\mu} \ln \frac{1}{\delta}$ examples. So, running the algorithm when there is a c^* with $\Pr[c^*] \geq \mu$ is guaranteed to yield a c that passes this test except with probability at most δ, since our algorithm guarantees $\Pr[c] \geq \Pr[c^*]$. Moreover, if c does not pass, therefore, $\Pr[c^*] \leq \mu$ (for our current estimate of μ), and we can safely reduce our estimate of μ by a factor of four and attempt to run the algorithm again; if we also reduce δ by a factor of 2 each time, we obtain a failure probability of at most 4δ overall, and find a satisfactory c after $\log_4 \frac{1}{\Pr[c^*]}$ iterations of the algorithm.

As before, essentially the same algorithm can find k-DNF by replacing the literals with conjunctions of at most k literals. In this case, the factor of n increase in the error probability unfortunately becomes a n^k increase.

4 Abduction for conjunctive representations solves hard learning problems

Abduction is frequently cast as the task of finding explanations given as a *conjunction* of literals. It is therefore natural to ask whether our abduction task for conjunctions is tractable. Unfortunately, we obtain evidence that it is not:

Theorem 5 *If the abduction task for conjunctions can be solved (even improperly), then DNF is PAC-learnable in polynomial time.*

We remind the reader of the definition of PAC-learning:

Definition 6 (PAC-learning [Valiant, 1984]) *A class of representations of Boolean predicates $\mathcal{C}$ is said to be* (improperly) PAC-learnable *if there is a polynomial time algorithm that, given access to examples drawn independently from an unknown distribution D, together with the evaluation of some unknown $c \in \mathcal{C}$ on the examples, for some input parameters ϵ and δ returns an efficiently evaluable hypothesis h such that with probability $1 - \delta$ over the examples, for x drawn from D, $\Pr[h(x) = c(x)] \geq 1 - \epsilon$.*

The proof is actually quite similar to the analogous result for agnostic learning of conjunctions by Kearns et al. [1994]: Recall that a *weak learning* algorithm is a PAC-learning algorithm that merely produces a hypothesis h such that $\Pr[h = c] \geq 1/2 + 1/\mathrm{poly}(n, |c|)$ (i.e., for some arbitrary polynomial in the size of the examples and representation c). A famous result by Schapire [1990] showed how to efficiently reduce PAC-learning to weak learning by *"boosting."*

Proof of Theorem 5: Suppose our examples are of the form (x, b) where $b = \varphi(x)$ for some DNF φ. We will show how to obtain a weak learner for φ from an algorithm for abducing conjunctions, thus obtaining a PAC-learning algorithm by boosting.

Suppose φ has size q. If φ is satisfied with probability greater than $1/2 + 1/q$ or less than $1/2 - 1/q$, then the constant functions will do as weak learners, so assume φ is satisfied with probability $1/2 \pm 1/q$. Then, we see that some term T in φ is satisfied with probability at least $1/2q - 1/q^2 \stackrel{\text{def}}{=} \mu$. We note that $\Pr[\varphi|T] = 1$. An algorithm for abducing conjunctions (with $\epsilon = 1/4$) for $\psi = b$ and $A = \{x_1, \ldots, x_n\}$ therefore finds a hypothesis $\tilde{T}$ such that $\Pr[\tilde{T}] \geq 1/p(1/\mu, n, 4/3) \geq 1/p'(n, q)$ for some polynomial $p'(n, q)$ and $\Pr[\varphi|\tilde{T}] \geq 3/4$.

Our weak learner is now as follows: if $\Pr[\varphi|\neg\tilde{T}] \geq 1/2$, we use the constant 1, and otherwise we use $\tilde{T}$. Note that this is equivalent to a hypothesis that predicts according to a majority vote on $\neg\tilde{T}$ (and predicts 1 on $\tilde{T}$). We note that it is correct with probability at least

$$\begin{aligned}\frac{1}{2}(1 - \Pr[\tilde{T}]) + \Pr[\varphi|\tilde{T}] \Pr[\tilde{T}] &\geq \frac{1}{2} + (\frac{3}{4} - \frac{1}{2}) \Pr[\tilde{T}] \\ &\geq \frac{1}{2} + \frac{1}{4p'(n, q)}\end{aligned}$$

which is sufficient for weak learning. ■

As a consequence, we obtain a hardness result for conjunctive conditions: recent work by Daniely et al. [2014] and Daniely and Shalev-Shwartz [2014] have established the following hardness of learning DNF.

Theorem 7 ([Daniely and Shalev-Shwartz, 2014]) *If there is some $f : \mathbb{N} \to \mathbb{N}$ such that $f \to \infty$ for which no polynomial-time algorithm can refute random k-SAT instances with $n^{f(k)}$ clauses, then there is no polynomial-time PAC-learning algorithm for DNF.*

The premise of the theorem is a strengthening of Feige's hypothesis [Feige, 2002], which was that a linear number of constraints are hard to refute. (Note that the state of the art requires $n^{k/2}$ clauses [Coja-Oghlan *et al.*, 2010]; as we add more constraints, it becomes *easier* to find a refutation.) Thus, as a corollary of Theorem 5, we find:

Corollary 8 *If there is some $f : \mathbb{N} \to \mathbb{N}$ such that $f \to \infty$ for which no polynomial-time algorithm can refute random*

k-SAT instances with $n^{f(k)}$ constraints, then there is no algorithm that solves the abduction task for conjunctions.

So, although this hypothesis is new and largely untested, it still provides some complexity-theoretic evidence that we should not expect to find a tractable algorithm for conjunctive explanations in our new model. This result essentially settles the question of which (natural) representations can be produced as explanations in our new model: Most natural knowledge representations are either expressible by a k-DNF, and thus fall within the scope of our earlier algorithms, or can themselves express conjunctions, and thus seem to be outside the scope of any algorithm on account of Theorem 5.

5 Applications and further motivations

Our use of k-DNF representations for abduction is a bit unusual. To our knowledge, only Inoue [2012] has previously considered abducing DNF representations. We therefore conclude with a brief discussion of some notable potential applications.

Goal formulation. Our results seem particularly suitable for the following application in planning: consider an essentially propositional formulation of planning such as in STRIPS or a factored (PO)MDP. Suppose we are given a collection of example traces and a (possibly complex) goal predicate. Then our algorithm, given the traces together with the goal as the query ψ (and an alphabet comprising, say, all of the state attributes), can identify a k-DNF condition that is satisfied moderately often that entails the goal is satisfied (when such a k-DNF exists). This k-DNF can then be provided as a representation of the goal for planning. Since standard techniques in planning based on SAT-solvers/resolution theorem-proving only handle DNF goals and are generally more efficient for DNFs of low width, this representation is a good fit: If no DNF subgoal exists, then the planning techniques cannot be applied, and a low-width DNF should be preferred if it can be found.

Selection of preconditions. The problems we consider here arise in Valiant's Robust Logic framework [Valiant, 2000b], which was a proposal to capture certain kinds of common sense reasoning [Valiant, 1994; 1995; 2006]; the problem also arises in a similar probabilistic formalization of common sense reasoning by Roth [1995], developed further by Khardon and Roth [1997a]. Roughly, the issue is that Valiant and Roth show that the famous non-monotonic effects of common sense reasoning can be captured naturally by a probabilistic semantics, in which the incorporation of new knowledge is captured by filtering the examples used for reasoning, referred to by Valiant as "applying a precondition." In these works, the precondition is simply assumed to be given; *how* it would be selected is not explicitly considered. Another work by Valiant [1994, p.164] informally suggests that such "context" might be simply given by the attributes that are currently firing in the neuroidal model. But, this view leads to problems: the specific, irrelevant details of the scene may never have been encountered before, leaving no examples to perform the common sense reasoning. The problems we formalize here might be viewed as the problem of *proposing* a candidate precondition (relative to some desired property) that has enough data to offer meaningful predictions.

We can formalize such a problem for a class of representations $\mathcal{C}$ as, *given* an observation x^* and a query ψ, finding a property $c \in \mathcal{C}$ of x^* – a *precondition* for x^* relative to ψ – of maximum probability that (approximately) entails ψ. Notice, the disjunction of all $c \in \mathcal{C}$ in which ψ is entailed has probability equal to at least that of the $c^* \in \mathcal{C}$ that maximizes $\Pr[c^*]$. So, for the cases we consider where $\mathcal{C}$ is either the class of disjunctions or k-DNFs, c^* must be equal to the disjunction over all such $c \in \mathcal{C}$. In particular, if x^* satisfies any $c \in \mathcal{C}$, x^* must satisfy c^*. Moreover, our analysis of the elimination algorithm also shows that whenever x^* satisfies c^*, it also satisfies the c we obtain, as c contains all of the terms of c^*. So, the elimination algorithm is also solving this precondition selection problem relative to a query for disjunctive classes.

In a standard example, we wish to select a precondition with respect to which we will make judgments about whether or not a bird Tweety flies in some particular example scene. We suppose that are told (or otherwise fix our attention somehow) specifically on the fact that Tweety is a penguin. Then, we might seek a precondition for the query penguin. This abduced condition may be more specific than merely satisfying penguin, and yet is as general as possible with respect to $\mathcal{C}$, satisfying the specific instance x whenever possible. Reasoning only using the examples that satisfy the abduced condition for penguinhood, we may thereby make the judgment $\neg$can_fly.

This application is related to the selection of a *"reference class"* for estimation, a problem that featured prominently in Reichenbach's theory of probability [1949]. Our use of disjunctive representations here does not follow Reichenbach's suggestion or its refinements by Kyburg [1974] or Pollock [1990], to find a *most specific* reference class, and relatedly, to *disallow disjunctive classes*; see Bacchus et al. [1996] for a discussion of this approach and some problems it encounters. Alas, the most natural representations for reference classes are arguably conjunctions, which Theorem 5 suggests may be out of reach.

Acknowledgements

We thank Leslie Valiant, Madhu Sudan, Or Sheffet, Sam Wiseman, Kilian Weinberger, and Sanmay Das for conversations that helped direct this work, and Loizos Michael and the anonymous reviewers for their feedback on an earlier draft.

References

[Awasthi *et al.*, 2010] P. Awasthi, A. Blum, and O. Sheffet. Improved guarantees for agnostic learning of disjunctions. In *COLT*, 2010.

[Bacchus *et al.*, 1996] F. Bacchus, A. J. Grove, J. Y. Halpern, and D. Koller. From statistical knowledge bases to degrees of belief. *AIJ*, 87:75–143, 1996.

[Bylander *et al.*, 1991] T. Bylander, D. Allemang, M. C. Tanner, and J. R. Josephson. The computational complexity of abduction. *AIJ*, 49:25–60, 1991.

[Charniak and McDermott, 1985] E. Charniak and D. McDermott. *Introduction to Artificial Intelligence*. Addison-Wesley, 1985.

[Coja-Oghlan *et al.*, 2010] A. Coja-Oghlan, C. Cooper, and A. Frieze. An efficient sparse regularity concept. *SIAM J. Discrete Math.*, 23(4):2000–2034, 2010.

[Cox and Pietrzykowski, 1986] P. Cox and T. Pietrzykowski. Causes for events: their computation and applications. In *8th CADE*, pages 608–621, 1986.

[Daniely and Shalev-Shwartz, 2014] A. Daniely and S. Shalev-Shwartz. Complexity theoretic limtations on learning DNF's. arXiv:1404.3378, 2014.

[Daniely *et al.*, 2014] A. Daniely, N. Linial, and S. Shalev-Shwartz. From average case complexity to improper learning complexity. In *46th STOC*, pages 441–448, 2014.

[Denecker and Kakas, 2002] M. Denecker and A. Kakas. Abduction in logic programming. In *Computational Logic: Logic Programming and Beyond*, volume 2407 of *LNAI*, pages 402–437. Springer, 2002.

[Eshghi, 1988] K. Eshghi. Abductive planning with event calculus. In *5th Int'l Logic Programming Conf.*, pages 562–579, 1988.

[Feige, 2002] U. Feige. Relations between average case complexity and approximation complexity. In *34th STOC*, pages 534–543, 2002.

[Flach and Kakas, 2000] P. A. Flach and A. C. Kakas. *Abduction and Induction: Essays on their relation and integration*. Springer, 2000.

[Hobbs *et al.*, 1990] J. Hobbs, M. Stickel, D. Appelt, and P. Martin. Interpretation as abduction. Technical Report 499, SRI, Menlo Park, CA, 1990.

[Inoue, 2012] K. Inoue. DNF hypotheses in explanatory induction. In S. H. Muggleton, A. Tamaddoni-Nezhad, and F. A. Lisi, editors, *ILP 2011*, volume 7207 of *LNCS*, pages 173–188. Springer, 2012.

[Kautz *et al.*, 1995] H. Kautz, M. Kearns, and B. Selman. Horn approximations of empirical data. *AIJ*, 74(1):129–145, 1995.

[Kearns *et al.*, 1994] M. J. Kearns, R. E. Schapire, and L. M. Sellie. Towards efficient agnostic learning. *Mach. Learn.*, 17(2-3):115–141, 1994.

[Khardon and Roth, 1997a] R. Khardon and D. Roth. Defaults and relevance in model based reasoning. *AIJ*, 97(1-2):169–193, 1997.

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

[Kyburg, 1974] H. E. Kyburg. *The Logical Foundations of Statistical Inference*. Reidel, 1974.

[McCarthy, 1980] J. McCarthy. Circumscription – a form of non-monotonic reasoning. *AIJ*, 13(1–2):27–39, 1980.

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

[Mill, 1843] J. S. Mill. *A System of Logic*, volume 1. John W. Parker, London, 1843.

[Missiaen *et al.*, 1995] L. Missiaen, M. Bruynooghe, and M. Denecker. CHICA, a planning system based on event calculus. *J. Logic and Computation*, 5(5), 1995.

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

[Pierce, 1931] C. S. Pierce. Elements of logic. In C. Hartshorn et al., editor, *Collected Papers of Charles Sanders Pierce*. Harvard University Press, 1931.

[Pollock, 1990] J. L. Pollock. *Nomic Probabilities and the Foundations of Induction*. Oxford University Press, 1990.

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

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

[Reggia, 1983] J. Reggia. Diagnostic expert systems based on a set-covering model. *Int'l J. Man-Machine Studies*, 19(5):437–460, 1983.

[Reichenbach, 1949] H. Reichenbach. *Theory of Probability*. University of California Press, 1949.

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

[Reiter, 1987] R. Reiter. A theory of diagnosis from first principles. *AIJ*, 32:57–95, 1987.

[Roth, 1995] D. Roth. Learning to reason: the non-monotonic case. In *14th IJCAI*, volume 2, pages 1178–1184, 1995.

[Roth, 1996] D. Roth. A connectionist framework for reasoning: Reasoning with examples. In *AAAI-96*, pages 1256–1261, 1996.

[Schapire, 1990] R. E. Schapire. The strength of weak learnability. *Mach. Learn.*, 5(2):197–227, 1990.

[Thompson and Mooney, 1994] C. A. Thompson and R. J. Mooney. Inductive learning for abductive diagnosis. In *AAAI'94*, pages 664–669, 1994.

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

[Valiant, 1994] L. G. Valiant. *Circuits of the Mind*. Oxford University Press, 1994.

[Valiant, 1995] L. G. Valiant. Rationality. In *8th COLT*, pages 3–14, 1995.

[Valiant, 2000a] L. G. Valiant. A neuroidal architecture for cognitive computation. *J. ACM*, 47(5):854–882, 2000.

[Valiant, 2000b] L. G. Valiant. Robust logics. *AIJ*, 117:231–253, 2000.

[Valiant, 2006] L. G. Valiant. Knowledge infusion. In *AAAI'06*, pages 1546–1551, 2006.