Conditional distribution search*

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

August 24, 2014

Abstract

We propose the problem of identifying explicit representations of events of interest in a given probability distribution using random examples. That is, we seek a representation of an event that has some given (moderately small but non-negligible) minimum probability, and gives a property of interest conditional probability near 1. Specifically, each example is of the form (x, b) where $b \in \{0, 1\}$ indicates whether or not the target property holds, and we seek a Boolean function $C(x)$ such that $\Pr[C(x) = 1] \geq \mu$ and $\Pr[b = 1|C(x) = 1] \geq 1 - \epsilon$ for some given μ and ϵ. We argue that this problem is of interest for several reasons. First, we believe that knowledge of such moderately rare events captures the power granted by *"big data,"* and as such our formal problems provide a framework in which one can demonstrate the value of "big data." Second, the problem we propose presents a "PAC" formalization of *abductive reasoning,* a basic AI task in which one seeks plausible (but not necessarily logically entailed) causes or explanations for an observation or desired outcome. Third and finally, such problems point the way to a new family of problems, *learning of special cases.*

We present four main theorems about the model we propose:

1. Disjunctive events (and more generally k-DNF events) can be found in time $O(n^2)$ (resp. $O(n^{2k})$).
2. By contrast, an efficient algorithm for identifying conjunctive events would yield an algorithm for distribution-free PAC-learning of DNF.
3. The tractability of search for a class of condition representations in this model is polynomial-time equivalent to the learnability of the same concept representation in the *positive-reliable* learning model proposed by Kalai et al. [9].
4. If the best disjunctive event gives the target property conditional probability $1-\epsilon$, we give an efficient algorithm that finds a disjunctive event for which the conditional probability of the target property is at least $1 - O(\frac{n}{\mu}\epsilon)$; that is, we give an algorithm for an *agnostic* variant of our conditional distribution search problem.

1 Introduction

What is *"big data"* good for? Standard answers, that more data allows richer models to be fit, don't hold up to scrutiny. The standard statistical analysis of machine learning algorithms (e.g., dating back to Vapnik and Chervonenkis [27]) shows that the amount of (i.i.d.) data required to fit a

*Work performed while the author was affiliated with Harvard University and supported by ONR grant number N000141210358.

discrete model is essentially proportional to the representation size. So, if the data requirements of the model were really the issue, then we would expect "big data" to just mean "big models." This is of course contrary to the usual wisdom in machine learning theory that smaller models are to be preferred, since the provide better generalization ability (cf. "Occam's razor" style theorems [3, 2]). More generally, this view of affairs disregards the fact that *computational power* rather than *sample size* tends to be the barrier to fitting richer models.

Nevertheless, there *is* a substantial benefit to "big data": Lambert [13] notes that much of Google's power at translation, spell correction, and so on is derived from a very simple model – looking at "n-grams," n-tuples of words or characters – that have vast coverage. Merely adding enough billions of n-grams yields state-of-the-art performance, even though the model is rather simplistic. A perhaps more accurate characterization of the power granted by big data in these cases is that it provides *knowledge of relatively rare events.* Typically in the analysis of machine learning, rare, exceptional events can be ignored or even should be ignored, in the case of noise-tolerant learning. By contrast, in this work, we focus on the question of *how to find a relatively rare event of interest* in a data set. In particular, as with the usual machine learning tasks, we would like to find descriptions of events that generalize well.

We formulate the question as searching for a *conditional distribution*: we assume that the data, drawn i.i.d. from a distribution D, is labeled true or false indicating on whether or not it is of interest (perhaps by a database query), and that we are given a Boolean representation class $\mathcal{C}$. We then seek to find a predicate C in the representation class such that

(i) the labels of $D|C$ are (almost) always true and

(ii) the probability of C under D is at least some minimum value μ.

(We will restate the definition more formally in Section 2.) We seek computationally efficient algorithms with a "PAC" guarantee: that with high probability (e.g., over the examples) both conditions are approximately satisfied.

We note that this problem is a new formalization of *"abductive reasoning,"* a basic task in AI applications, in which one seeks plausible circumstances or causes for a known fact or desired result. For example, to understand a story in which a man is holding a gun in a bank, one must "abductively infer" that the man wishes to rob the bank. This is not a sound inference, of course – 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. In our model, the labels might indicate whether or not the key facts of the story – the gun in the bank – are present in the examples, and the predicate C to be found represents the desired "explanation." We will discuss such applications slightly further in Section 5. The model we propose here is not the first probabilistic model of abduction, but it differs crucially from previous models, which were Bayesian models that depended on assigning some prior probabilities to the various explanations. In particular, both the probabilistic version of the "set covering" model of abduction presented by Bylander et al. [4] and the models of abduction based on probabilistic graphical models [16, 18] were of this form. In these models, "plausibility" of the proposed conditions is evaluated in terms of prior probabilities. These methods therefore require specifying a prior somehow, which is deemed to be one of their main drawbacks [14], as good priors are hard to estimate. Another method, proposed by Bacchus et al. [1] simply uses a maximum-entropy distribution over the attributes as a prior, which is representation language dependent and may be simply inappropriate. By contrast, the model we propose here is an analogue of the "distribution-free" PAC model of machine learning proposed by Valiant [22] that avoids such problems by avoiding the explicit representation of priors of any kind. We believe that this model

features many of the further attractive properties of PAC learning, specifically that it is a relatively simple model that is (i) appropriate to many application domains, (ii) well suited to the design and analysis of algorithms, and (iii) features strong connections to other areas.

This last point is illustrated by our results: In addition to two motivating positive results, we present two results connecting our framework to other problems in machine learning. The positive results are algorithms for finding disjunctive events, one simple algorithm for the *"realizable"* (error-free) case in which some disjunctive event C gives the target label conditional probability 1 of being true, and one approximation algorithm for an *"agnostic"* (adversarial) setting, in which the best disjunctive event may only give the target label conditional probability $1 - \epsilon$ for some $\epsilon > 0$. The first connection is that an algorithm for the realizable case for *conjunctive* events, by contrast, would yield an algorithm for PAC-learning DNF formulas in the standard distribution-free model. This has been the central open problem in PAC-learning since the model was first proposed by Valiant [22], and recent results by Daniely, Linial, and Shalev-Shwartz [5] and Daniely and Shalev-Shwartz [6] show that the problem may be intractable, given a strong assumption about the hardness of refuting random k-SAT (stronger than Feige's assumption [7]). The second connection is that algorithms for the realizable version of our problem are polynomial-time equivalent to the *"positive-reliable agnostic learning"* model of Kalai et al. [9]. In this model, one must guarantee that false-positives are below a given bound, and false-negatives are to be minimized subject to this first, hard constraint. Although the positive-reliable learnability of disjunctions was recently established using a similar algorithm by Kanade and Thaler [10][1], the rest of the results we present are also new in the positive-reliable model. At the same time, Kanade and Thaler's main results for positive-reliable learnability also carry back to our conditional distribution search model. In particular, their subexponential-time algorithms for positive-reliable learning of DNFs provide subexponential-time algorithms for finding DNF events in our model. Of course, we view all of these results as secondary contributions that serve to validate the formulation of the problem, which we claim as our main contribution.

We also note that the model suggests a more general class of questions than we address here: can we identify conditions under which prediction is possible? Our results may be viewed as finding a condition in which a trivial, constant predictor is successful. It is often the case in science (or life in general) that simple rules only succeed at making predictions under carefully controlled conditions or in simple examples: For example, the basic laws of mechanics are best illustrated on low-friction surfaces such as sheets of ice. We often discover such rules by first observing the simplified case, and then explicitly apply the rules so learned to more complex cases in which multiple effects are present. Given the large, rich data sets available to us, it would be desirable to be able to automatically identify such special cases in order to generate hypotheses. We will discuss this in more detail along with other questions raised by this work in Section 6

2 Disjunctive versus conjunctive conditions

In this work, we are seeking to find explicit representations of events of low but non-negligible probability in which some target property is (almost) always satisfied. Formally, we focus on the following class of problems:

[1]We note that the direct analysis we present for our model is slightly tighter than one would obtain by using Kanade and Thaler's algorithm with the reduction we give between the models.

Definition 1 (Conditional distribution search) *For a* representation class $\mathcal{C}$ *of* $c : \{0,1\}^n \to \{0,1\}$*, the (proper)* conditional distribution search problem *is as follows. Given as input* m *i.i.d. examples* $(x^{(1)}, b^{(1)}), \ldots, (x^{(m)}, b^{(m)})$ *from an arbitrary distribution* D *over* $\{0,1\}^n \times \{0,1\}$ *for which there exists* $c^* \in \mathcal{C}$ *such that* $\Pr[b = 1 | c^*(x) = 1] = 1$ *and* $\Pr[c^*(x) = 1] \geq \mu$*, with probability* $1 - \delta$*, find some* $c \in \mathcal{C}$ *such that*

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

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

As with machine learning more generally, 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 as usual consider an *"improper"* version of the problem, finding representations from some larger class than the $\mathcal{C}$ containing the "optimal" condition c^*. The form of the representation is important for some of the AI applications we discuss 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. Although we might ideally like to maximize the probability of the event we are conditioning on (and we will actually achieve this for disjunctions below), in general we don't have much control over the measure μ of the target event. Granting the algorithm some slack, so long as it identifies an event that is guaranteed to have non-negligible probability, seems reasonable. In particular, our results connecting conditional distribution search to other models tolerate such slack, and some of our results require a small amount of slack. Specifically, Theorem 7 inherits a bit of slack from the positive-reliable model; the algorithm for the agnostic variant of our setting in Section 4 also features such slack.

Valiant [22] gave an analysis showing that the "elimination algorithm" (Algorithm 1) is a PAC-learning algorithm for disjunctions. We note that essentially the same algorithm also can be used to identify a disjunctive condition quite efficiently.

```
input  : Examples (x^(1), b^(1)), ..., (x^(m), b^(m))
output : A disjunction of literals over x_1, ..., x_n.
begin
    Initialize c to be the (trivial) disjunction over all 2n literals.
    for i = 1, ..., m do
        if b^(i) = 0 then
            forall the ℓ ∈ c do
                if ℓ(x^(i)) = 1 then
                    Remove ℓ from c.
                end
            end
        end
    end
    return c.
end
```

Algorithm 1: Elimination algorithm

Theorem 2 *The conditional distribution search problem for disjunctive conditions 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 is contained in c, so c contains all of the literals of c^*. We claim that this invariant is maintained: since $\Pr[b|c^*] = 1$, whenever $b^{(i)} = 0$, 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 $c^* \Rightarrow c$, $\Pr[c] \geq \mu$.

We now bound the probability that the algorithm terminates with a c such that $\Pr[b|c] < 1 - \epsilon$. We note that for such c, $\Pr[b = 0 \wedge c] > \epsilon \Pr[c]$. For any c with $\Pr[c] \geq \mu$, it follows that $\Pr[b = 0 \wedge c] > \mu\epsilon$. Thus, after $m = \frac{1}{\mu\epsilon}(n \ln 3 + \ln 1/\delta)$ examples, the probability that such a c is always falsified when $b = 0$ for all of them is at most

$$(1 - \mu\epsilon)^m \leq e^{-\ln(3^n/\delta)} = \frac{\delta}{3^n}$$

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 $b = 0$, we find that with probability $1 - \delta$ $\Pr[b|c] \geq 1 - \epsilon$ as needed. ■

We remark that the algorithm extends to solve the conditional distribution search problem for k-DNF formulas with sample complexity $O(\frac{1}{\mu\epsilon}(n^k + \log 1/\delta))$ in the usual way.

Comparison to PAC-learning of disjunctions. The algorithm is identical to the standard algorithm for learning disjuncitons. But, we stress that our analysis differs from the PAC-learning analysis (by Valiant [22]) in a few ways. First, in PAC-learning, by hypothesis one can never encounter an example where $b^{(i)} = 1$ and $c(x^{(i)}) = 0$ since it is assumed that $b^{(i)} = c^*(x^{(i)})$ and $c^* \Rightarrow c$. Here, we may encounter such examples, but we are free to ignore them (as c^* does). Second, PAC-learning did not need to establish a bound on $\Pr[c]$—instead, it requried agreement with c^* when $c^* = 1$. Of course, both conditions follow from the same property, that $c^* \Rightarrow c$. Third and finally, in PAC-learning, one just needs to bound $\Pr[b = 0 \wedge c]$ (under D) whereas here we need to bound the quantity under $D|c$. Since c is at least μ-dense in D, we can achieve this at the cost of the extra $1/\mu$-factor in the sample complexity.

The above distinctions may seem like trivial technicalities. Our next result shows that this is not so, and conditional distribution search is apparently quite different from PAC-learning. We recall that PAC-learning is "closed under complementation," and so the elimination algorithm solves the problem of learning conjunctions by simply switching the labels of the attributes and examples. By contrast, an algorithm for the *conditional distribution search* problem for conjunctions would be a major breakthrough:

Theorem 3 *If the conditional distribution search problem for conjunctions can be solved (even improperly), then DNF is PAC-learnable in polynomial time.*

The proof is actually quite similar to the analogous result for agnostic learning of conjunctions by Kearns et al. [11]: Recall that a *weak learning* algorithm is a PAC-learning algorithm that merely

produces a hypothesis h such that $\Pr[h = c] \geq 1/2 + 1/\text{poly}(n)$. A famous result by Schapire [21] showed how to efficiently reduce PAC-learning to weak learning by *"boosting."*

Proof: Suppose $b = \varphi(x)$ for some DNF φ. We will show how to obtain a weak learner for φ from an algorithm for conjunctive conditions, thus obtaining a PAC-learning algorithm by boosting. Suppose φ has size $q(n)$. If φ is satisfied with probability greater than $1/2 + 1/q(n)$ or less than $1/2 - 1/q(n)$, then the constant functions will do as weak learners, so assume φ is satisfied with probability $1/2 \pm 1/q(n)$. Then, we see that some term T in φ is satisfied with probabilty at least $1/2q(n) - 1/q(n)^2 = \mu$. We note that $\Pr[\varphi|T] = 1$. An algorithm for conjunctive condition search (with $\epsilon = 1/4$) therefore finds a hypothesis $\tilde{T}$ such that $\Pr[\tilde{T}] \geq p(1/\mu, n, 3/4) = 1/p'(n)$ for some polynomial $p'(n)$ 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

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

which is sufficient for weak learning. ■

As a consequence, we obtain a hardness result for conjunctive conditions: recent work by Daniely et al. [5] and Daniely and Shalev-Shwartz [6] have established the following hardness of learning DNF.

Theorem 4 (Daniely et al. [5]; Daniely and Shalev-Shwartz [6]) *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 polynomial-time PAC-learning algorithm for DNF of size n^ϵ for any $\epsilon > 0$.*

The premise of the theorem is a strengthening of Feige's hypothesis [7], which was that a linear number of constraints are hard to refute.[2] Thus, as a corollary of Theorem 3, we find:

Corollary 5 *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 conditional distribution search problem for conjunctions.*

3 Condition search is equivalent to reliable agnostic learning

Kalai et al. [9] proposed models for *"reliable" agnostic learning* in which false positives and false negatives are treated as fundamentally different. In particular, in *positive-reliable learning,* one seeks to guarantee that the false positive rate of the classifier is below a given tolerance ϵ. The false negative rate is then minimized, subject to this hard constraint on the false positive rate. Formally:

Definition 6 (Positive-reliable learning [9, 10]) *For a* representation class $\mathcal{C}$ *of* $c : \{0,1\}^n \to \{0,1\}$ *the* distribution independent positive-reliable learning *task is as follows. One is given access to i.i.d. examples from an arbitrary distribution D over $\{0,1\}^n \times \{0,1\}$, and input parameters $\epsilon, \delta \in (0,1]$. With probability at least $1 - \delta$, the algorithm should produce a hypothesis $h : \{0,1\}^n \to \{0,1\}$ such that*

[2] Note that as we add more constraints, it becomes *easier* to find a refutation.

1. $\Pr_{(x,b)\in D}[h(x) = 1 \wedge b = 0] \leq \epsilon$ *and*
2. $\Pr_{(x,b)\in D}[h(x) = 0 \wedge b = 1] \leq \min_{c\in\mathcal{C}:\Pr[c(x)=1\wedge b=0]=0} \Pr_{(x,b)\in D}[c(x) = 0 \wedge b = 1] + \epsilon$.

When an algorithm solves this task by producing $h \in \mathcal{C}$, we say that the algortihm is proper.

This definition was intended to capture applications such as spam filtering where (i) no simple classifier can achieve 100% accuracy and (ii) we are more willing to tolerate a few spam e-mails than we are the loss of legitimate e-mails to the filter.

It turns out that this asymmetric treatment of false positives and false negatives is essentially what we require for conditional distribution search:

Theorem 7 *If positive-reliable learning can be solved for a concept class $\mathcal{C}$ in time $T(n, 1/\epsilon, 1/\delta)$, conditional distribution search for $\mathcal{C}$ can be solved with $1/p(1/\mu, n, 1/(1-\epsilon)) = (1-\epsilon)\mu$ in time $T(n, \frac{2}{\mu\epsilon}, \frac{1}{\delta})$. Conversely, If conditional distribution search can be solved in time $T(n, 1/\mu, 1/\epsilon, 1/\delta)$, with $1/p(1/\mu, n, 1/(1-\epsilon)) = (1-\epsilon)\mu$, then positive reliable learning can be achieved in time*

$$O\left(T\left(n, \frac{1}{\mu}, \frac{2}{\epsilon}, \frac{1}{\delta}\right)\log\frac{1}{\epsilon} + \frac{1}{\epsilon^3}\log\frac{1}{\epsilon}\log\frac{1}{\delta}\right).$$

The same results hold for both improper and proper versions of the problem as long as the all-false concept is in $\mathcal{C}$.

Proof:

Conditional distribution search from positive reliable learning. Suppose there exists $c^* \in \mathcal{C}$ such that $\Pr[c^*(x) = 1 \wedge b = 0] = 0$ and $\Pr[c^*(x) = 1] \geq \mu$ for our given μ. We will call the positive reliable learning algorithm on $\epsilon' = \epsilon\mu/2$ and the desired δ. Then with probability $1-\delta$, the algorithm returns a hypothesis h such that

$$\Pr[h(x) = 1 \wedge b = 0] \leq \epsilon' = \epsilon\mu/2$$

and

$$\begin{aligned}
\Pr[h(x) = 0 \wedge b = 1] &\leq \min_{c\in\mathcal{C}:\Pr[c(x)=1\wedge b=0]=0} \Pr[c(x) = 0 \wedge b = 1] + \epsilon' \\
&= \Pr[b = 1] - \max_{c\in\mathcal{C}:\Pr[c(x)=1\wedge b=0]=0} \Pr[c(x) = 1] + \epsilon' \\
&\leq \Pr[b = 1] - \mu + \epsilon' \\
&= \Pr[b = 1] - (1 - \epsilon/2)\mu
\end{aligned}$$

We therefore find

$$\begin{aligned}
\Pr[h(x) = 1] &= \Pr[h(x) = 1 \wedge b = 0] + \Pr[h(x) = 1 \wedge b = 1] \\
&= \Pr[h(x) = 1 \wedge b = 0] + \Pr[b = 1] - \Pr[h(x) = 0 \wedge b = 1] \\
&\geq 0 + (1 - \epsilon/2)\mu
\end{aligned}$$

so the second condition is satisfied. Returning to the first condition, since $\mu \leq \frac{1}{1-\epsilon/2}\Pr[h(x) = 1]$,

$$\begin{aligned}
\Pr[h(x) = 1 \wedge b = 0] &\leq \frac{\epsilon}{2}\frac{1}{1-\epsilon/2}\Pr[h(x) = 1] \\
&\leq \epsilon\Pr[h(x) = 1]
\end{aligned}$$

so our first condition is also satisfied. Thus, h is a solution to the conditional distribution search problem for μ and ϵ; and, if the positive reliable learning algorithm is proper, then $h \in \mathcal{C}$ as well.

Positive reliable learning from conditional distribution search. We first note that we can use binary search to optimize the value of μ for which the distribution search algorithm succeeds at finding an event at a cost of $O(\log \frac{1}{\mu})$ calls to the algorithm. After each call, we can test if the returned event has at probability at least $(1-\epsilon)\mu$ using $O(\frac{1}{\epsilon^2 \mu} \log \frac{1}{\delta})$ time and examples. If the largest μ is at most ϵ, then we will argue that the all-false concept is a solution to the positive-reliable learning problem, and hence this overhead is at most $O(\log \frac{1}{\epsilon})$ calls to the distribution search algorithm and $O(\frac{1}{\epsilon^3} \log \frac{1}{\delta} \log \frac{1}{\epsilon})$ time for testing.

Consider the hypothesis h produced by conditional distribution search on input parameters $\epsilon/2$ and δ. With probability $1-\delta$, its false positive rate is

$$\Pr[h(x) = 1 \wedge b = 0] \leq (\epsilon/2) \Pr[h(x) = 1] \leq \epsilon/2.$$

Now, note that

$$\Pr[h(x) = 1] = \Pr[h(x) = 1 \wedge b = 0] + \Pr[h(x) = 1 \wedge b = 1] \leq \Pr[h(x) = 1 \wedge b = 1] + \epsilon/2.$$

Since, assuming the optimal $\mu \geq \epsilon$, also

$$\begin{aligned} \Pr[h(x) = 1] &\geq (1 - \epsilon/2) \max_{c \in \mathcal{C}: \Pr[c(x)=1 \wedge b=0]=0} \Pr[c(x) = 1] \\ &\geq \max_{c \in \mathcal{C}: \Pr[c(x)=1 \wedge b=0]=0} \Pr[c(x) = 1 \wedge b = 1] - \epsilon/2 \end{aligned}$$

we have

$$\Pr[h(x) = 1 \wedge b = 1] \geq \max_{c \in \mathcal{C}: \Pr[c(x)=1 \wedge b=0]=0} \Pr[c(x) = 1 \wedge b = 1] - \epsilon.$$

Therefore, since $\Pr[b = 1]$ is conserved, i.e.,

$$\begin{aligned} \Pr[h(x) = 0 \wedge b = 1] + \Pr[h(x) = 1 \wedge b = 1] &= \Pr[b = 1] \\ &= \Pr[c(x) = 0 \wedge b = 1] + \Pr[c(x) = 1 \wedge b = 1] \end{aligned}$$

holds for any c,

$$\Pr[h(x) = 0 \wedge b = 1] \leq \min_{c \in \mathcal{C}: \Pr[c(x)=1 \wedge b=0]=0} \Pr[c(x) = 0 \wedge b = 1] + \epsilon$$

as needed. We now furthermore note that if $\max_{c \in \mathcal{C}: \Pr[c(x)=1 \wedge b=0]} \Pr[c(x) = 1] \leq \epsilon$, then the false-positive rate of the all-false concept is zero, and the optimal false-negative rate is at most ϵ. Therefore, the all-false concept is a solution. So, if the conditional distribution search algorithm was proper and the all-false concept is in $\mathcal{C}$, either way we also have $h \in \mathcal{C}$. ■

As a consequence, Theorem 2 can essentially be recovered from the analysis of the elimination algorithm for positive reliable learning by Kanade and Thaler [10], albeit with slightly worse parameters than obtained by our direct analysis. The rest of their results on learning with dual polynomials carry over to give, for example, subexponential algorithms for DNF condition search in our model. (We refer the interested reader to their paper for details.) The rest of our results, e.g., on the hardness of learning conjunctions, were not previously known in the positive reliable learning model. (In contrast to the standard agnostic learning model.)

Discussion. The proof of Theorem 2 highlights both the close connection between the two models; the essential differences are also present, but less prominently so. The main distinction between the models is that the all-false concept may make sense for positive-reliable learning, but it is never a legitimate output for conditional distribution search: After all, the all-false concept corresponds to an event of probability zero, and our probability space is discrete. Relatedly, there is a need to "terminate the search early" when the probability of positive examples is very small and the all-zero concept is a good approximator. This is because we require at least $1/\mu$ examples to detect an event of probability μ with constant confidence: The running time of the optimization version of conditional distribution search (used in the reduction of positive-reliable learning to conditional distribution search) therefore inherently depends on $1/\mu$, whereas positive-reliable learning should not depend on such a μ parameter.

4 Agnostic condition search for disjunctive conditions

We now consider a variant of our basic model in which no event of reasonably large probability gives the label conditional probability 1. This is the natural analogue of agnostic learning for conditional events. We will give a LP-based approximation algorithm for disjunctive conditions, inspired by the algorithm for MAX SAT of Goemans and Williamson [8].

For i.i.d. examples $(x^{(1)}, b^{(1)}), \ldots, (x^{(m)}, b^{(m)})$ drawn from D, we first solve the following linear feasibility problem; it is a relaxation of an integer linear program in which the $w(\ell)$ variable indicates whether the literal ℓ is included in the selected disjunction, and the w_j variable indicates whether or not the example x_j is true under the disjunction selected by the $w(\ell)$ variables.

$$
\begin{aligned}
w_j &\geq w(\ell) && \forall \ell : \ell(x^{(j)}) = 1, j = 1, \ldots, m \\
w_j &\leq \sum_{\ell : \ell(x^{(j)})=1} w(\ell) && j = 1, \ldots, m \\
\sum_{j : b^{(j)}=0} w_j &\leq \varepsilon \sum_{j=1}^{m} w_j \\
\frac{1}{m} \sum_{j=1}^{m} w_j &\geq \mu \\
w_j &\in [0,1] && j = 1, \ldots, m \\
w(\ell) &\in [0,1] && \forall \ell
\end{aligned}
$$

We then round it as follows: each literal ℓ is included in the final disjunction with probability $w(\ell)$ independently.

The heart of the analysis, below, lies in showing that with non-negligible probability, this rounding obtains a disjunction that approximately solves an empirical version of the conditional distribution search problem:

Theorem 8 *For any $\nu \in (0,1)$, with probability at least $(1 - 1/e)(1 - \nu)\nu\mu^2$, the disjunction obtained is satisfied on at least a $(1 - 1/e)\nu\mu$ fraction of the examples, and $b^{(j)} = 0$ on at most a $O\left(\frac{n}{\nu(1-\nu)\mu}\epsilon\right)$ fraction of these examples.*

Proof: Let W_j denote the event that example j satisfies the disjunction we obtain by our randomized rounding. A standard calculation (see either Goemans and Williamson [8, Lemma 3.1] or Motwani and Raghavan [15, Lemma 5.3]) shows

$$\mathbb{E}[W_j] \geq \min\{1, (1-1/e) \sum_{\ell:\ell(x^{(j)})=1} w(\ell)\}.$$

Likewise, a union bound gives $\mathbb{E}[W_j] \leq n \max_{\ell:\ell(x^{(j)})=1} w(\ell)$. Now, since the LP constraints guarantee $w_j \leq \min\{1, \sum_{\ell:\ell(x^{(j)})=1} w(\ell)\}$ and $w_j \geq \max_{\ell:\ell(x^{(j)})=1} w(\ell)$,

$$\mathbb{E}\left[\sum_{j=1}^{m} W_j\right] \geq (1-1/e) \sum_{j=1}^{m} w_j \geq (1-1/e)\mu m.$$

Similarly,

$$\mathbb{E}\left[\sum_{j:b^{(j)}=0} W_j\right] \leq n \sum_{j:b^{(j)}=0} w_j \leq n\epsilon \sum_{j=1}^{m} w_j.$$

We now show that with moderate probability, both of the conditions are approximately satisfied. First, we note that by Markov's inequality, the probability that $\sum_{j:b^{(j)}=0} W_j$ exceeds its expectation by a $\frac{1}{(1-1/e)\nu\mu}$-factor is at most $(1-1/e)\nu\mu$. Similarly, as $\frac{1}{m}\sum_{j=1}^{m} W_j \in [0,1]$, $p = \Pr\left[\frac{1}{m}\sum_{j=1}^{m} W_j < \nu\mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m} W_j\right]\right]$ satisfies

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m} W_j\right] &\leq p\nu\mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m} W_j\right] + (1-p) \\
p\left(1 - \nu\mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m} W_j\right]\right) &\leq 1 - \mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m} W_j\right] \\
p &\leq 1 - \nu\mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m} W_j\right] \frac{1}{1-(1-\nu)\mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m} W_j\right]} \\
&\leq 1 - (1-1/e)\nu\mu\frac{1}{1-(1-1/e)(1-\nu)\mu}
\end{aligned}$$

By a union bound, therefore, both sums satisfy these approximate bounds with probability at least

$$(1-1/e)\nu\mu\frac{1}{1-(1-1/e)(1-\nu)\mu} - (1-1/e)\nu\mu \leq (1-1/e)(1-\nu)\nu\mu^2.$$

Note that in this case we have $\frac{1}{m}\sum_{j=1}^{m} W_j \geq (1-1/e)\nu\mu$ and

$$\begin{aligned}
\sum_{j:b^{(j)}=0} W_j &\leq \frac{1}{(1-1/e)\nu\mu}\mathbb{E}\left[\sum_{j:b^{(j)}=0} W_j\right] \\
&\leq \frac{n}{(1-1/e)\nu\mu}\sum_{j:b^{(j)}=0} w_j \\
&\leq \frac{n}{(1-1/e)\nu\mu}\epsilon\sum_{i=1}^{j} w_j \\
&\leq \frac{n}{(1-1/e)^2\nu\mu}\epsilon\mathbb{E}\left[\sum_{i=1}^{j} W_j\right] \\
&\leq \frac{n}{(1-1/e)^2(1-\nu)\nu\mu}\epsilon\sum_{j=1}^{m} W_j
\end{aligned}$$

as needed. ■

As promised, Theorem 8 shows that our approximation algorithm successfully solves an *empirical* version of the problem. A straightforward application of (multiplicative) Chernoff bounds now allows us to specify a number of examples that guarantees that the event we identify will solve the conditional distribution search problem, essentially completing the analysis of our algorithm. After $O(\frac{1}{(1-1/e)(1-\nu)\nu\mu^2})$ trials at rounding the LP, with constant probability one obtains a suitable disjunction:

Theorem 9 *Given* $m = \Omega\left(\left(\frac{1}{\nu\mu} + \frac{\nu(1-\nu)}{n\epsilon}\right)\left(n + \log\frac{1}{\delta}\right)\right)$ *examples, with probability* $1-\delta$*, every disjunction* c *that is satisfied on at least a* $\Omega(\nu\mu)$ *fraction of the examples, and has* $b = 0$ *on at most a* $O\left(\frac{n}{\nu(1-\nu)\mu}\epsilon\right)$ *fraction of the examples where it is satisfied has* $\Pr[c] \geq \Omega(\nu\mu)$ *and*

$$\Pr[b = 0 | c(x) = 1] \leq O\left(\frac{n}{\nu(1-\nu)\mu}\epsilon\right).$$

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

5 Applications and motivations

Our problems are "PAC" formalizations of *abductive reasoning*, a basic variety of task in AI in which one seeks *plausible* conditions or "explanations" for an observation or desired outcome. A variety of applications are mentioned in the survey by McIlraith [14] (with emphasis on logic-based formalizations of abductive reasoning); beyond the application to story understanding we sketched earlier, abductive reasoning is the main tool used in diagnosis and plan recognition. One application that our results seem particularly suitable for is in planning: given a collection of example traces and a (possibly complex) goal predicate, our algorithm 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.

The problems we consider here also arise in Valiant's Robust Logic framework [25], which was a proposal to capture certain kinds of common sense reasoning [23, 24, 26]. The problem also arises in a similar probabilistic formalization of common sense reasoning by Roth [20]. Roughly, the issue is that Valiant and Roth show that the (in)famous non-monontonic 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 [23, p.164] informally suggests that such "context" might be given by some attributes are simply 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.

The problem of choosing a precondition is related to the selection of a *"reference class"* for estimation, a problem that featured prominently in Reichenbach's theory of probability [19]. But, we do not follow Reichenbach's suggestion (or its refinements by Kyburg [12] or Pollock [17]), to find a *most specific* reference class, and relatedly, the algorithms we give propose *disjunctive* classes, which were disallowed in the refinements; see Bacchus et al. [1] for a discussion of this approach and some problems it encounters. Alas, the most natural representations for reference classes are arguably conjunctions, which Theorem 3 suggests may be out of reach.

6 Directions for future work

As noted in the introduction, one view of our work is that it presents a first, simple example of an algorithm for *learning in special cases*. That is, we explicitly identify a special case in which a trivial, constant predictor succeeds. A more general problem, of supervised learning of a "special case" is as follows:[3]

Definition 10 (Conditional supervised learning) *For a* representation class $\mathcal{C}$ *of* $c : \{0,1\}^n \to \{0,1\}$ *and* hypothesis class $\mathcal{H}$ *of* $h : \{0,1\}^n \to \{0,1\}$*, the (proper)* conditional supervised learning problem *is as follows. Given as input* m *i.i.d. examples* $(x^{(1)}, b^{(1)}), \ldots, (x^{(m)}, b^{(m)})$ *from an arbitrary distribution* D *over* $\{0,1\}^n \times \{0,1\}$ *for which there exists* $c^* \in \mathcal{C}$ *and* $h^* \in \mathcal{H}$ *such that*

*1**. $\Pr[h^*(x) = b | c^*(x) = 1] = 1$ *and*

*2**. $\Pr[c^*(x) = 1] \geq \mu$,

with probability $1 - \delta$*, find some* $c \in \mathcal{C}$ *and* $h \in \mathcal{H}$ *such that*

1. $\Pr[h(x) = b | c(x) = 1] \geq 1 - \epsilon$ *and*

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

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

[3] We thank Madhu Sudan for suggesting this generalization.

Thus, the problem is essentially to *jointly* learn a predictor h for a nontrivial special case c when some rule in $\mathcal{H}$ would succeed for some (unknown) condition in $\mathcal{C}$. For example, can we learn linear threshold functions under disjunctive conditions? ($\mathcal{C}$ is then the class of disjunctions and $\mathcal{H}$ is the class of linear threshold functions.) This definition captures cases in which the target b may *not* be described by a linear threshold function everywhere, or is described by multiple linear threshold functions in different portions of the domain. Perhaps an algorithm can be designed by generalizing the LP we presented for the agnostic setting to optimize over sets of variables for the two distinct representations.

As we noted in the introduction, these problems capture cases in which simple rules succeed only in special cases, which seems to occur frequently in nature: where multiple causes and phenomena interact, the resulting relationships between the examples and the labels are often complicated, and beyond our ability to learn efficiently. In such cases, we often do proceed by finding such special cases in which individual laws can be observed essentially in isolation. In this vein, one could also ask whether linear regression in (say, disjunctive) special cases is possible, generalizing the attribute b from Boolean to real-valued.

Of course we could also consider an improper version in which the algorithm is not constrained to produce representations in $\mathcal{C}$ or $\mathcal{H}$, as well as two "semi-proper" versions where either $\mathcal{C}$ or $\mathcal{H}$ is constrained; the version in which the algorithm *is* constrained to produce a representation in $\mathcal{C}$ but *not* $\mathcal{H}$ may be the most interesting.

Another, more technical question raised by this work concerns the quality of approximation algorithms for agnostic conditional distribution search. The randomized rounding algorithm presented in Section 4 achieves a rather poor approximation guarantee in the following sense: our intent is to focus on cases where μ is small but non-negligible. Thus, the appearance of a factor of $1/\mu$ in the approximation factor for the error is dissatisfying. Can it be removed? Although in expectation our randomized rounding achieves an approximation that has no dependence on μ, it appears on account of our use of (what is essentially) Markov's inequality to give a lower-bound on the measure of the event we find. Observe that we need *both* conditions on this disjunction to hold simultaneously, and the conditions are negatively correlated, so we can't simply invoke techniques such as the method of conditional expectations to round to a fixed disjunction. In the first place, we need to argue that some good solution *exists* that satisfies both conditions without the dependence on μ.

Finally, of course, the question of whether or not the conditional distribution search problem for conjunctive conditions has an efficient algorithm is still interesting. Given the connection to PAC-learning of DNF representations presented by Theorem 3, however, one might conjecture not (or at least that such an algorithm will be challenging to find). Still, it is natural to ask if the hardness assumption for this problem can be weakened beyond what was achieved by Daniely and Shalev-Shwartz [6] via the connection to distribution-free DNF learning. For example, can we show that conjunctive condition search is intractable under Feige's [7] original assumption directly?

Acknowledgements

We thank Leslie Valiant, Madhu Sudan, Or Sheffet, and Sam Wiseman for conversations that helped direct this work.

References

[1] Fahiem Bacchus, Adam J. Grove, Joseph Y. Halpern, and Daphne Koller. From statistical knowledge bases to degrees of belief. *Artificial Intelligence*, 87:75–143, 1996.

[2] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Occam's razor. *Information Processing Letters*, 24:377–380, 1987.

[3] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989.

[4] Tom Bylander, Dean Allemang, Michael C. Tanner, and John R. Josephson. The computational complexity of abduction. *Artificial Intelligence*, 49:25–60, 1991.

[5] Amit Daniely, Nati Linial, and Shai Shalev-Shwartz. From average case complexity to improper learning complexity. In *Proc. 46th STOC*, pages 441–448, 2014.

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

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

[8] Michel X. Goemans and David Williamson. New 3/4-approximation algorithms for the maximum satisfiability problem. *SIAM J. Discrete Math.*, 7(4):56–666, 1994.

[9] Adam Tauman Kalai, Varun Kanade, and Yishay Mansour. Reliable agnostic learning. *JCSS*, 78:1481–1495, 2012.

[10] Varun Kanade and Justin Thaler. Distribution-independent reliable learning. In *Proc. 27th COLT*, volume 35 of *JMLR Workshops and Conference Proceedings*, pages 3–24, 2014.

[11] Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Towards efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

[12] Henry E. Kyburg. *The Logical Foundations of Statistical Inference*. Reidel, Dordrecht, 1974.

[13] Diane Lambert. Statistics at google scale. In *Third Annual Symposium on the Future of Computation in Science and Engineering*, January 24, 2014. Video available from `http://computefest.seas.harvard.edu/data-storm`.

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

[15] Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

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

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

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

[19] Hans Reichenbach. *Theory of Probability*. University of California Press, Berkeley, CA, 1949.

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

[21] Robert E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

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

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

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

[25] Leslie G. Valiant. Robust logics. *Artificial Intelligence*, 117:231–253, 2000.

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

[27] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.