6 Learning to Reason with Full Examples

The chapter discusses the entailment problem under the probably approximate truth semantics when we are given complete truth-assignments as examples, introduced in Section 6.1. In particular, we will focus on the case where properties hold with less than perfect validity, introduced in Section 6.2. Such properties capture "rules of thumb" that we show, in Section 6.4, may behave "non-monotonically" when we acquire additional knowledge, as captured by the approximate validity with respect to a conditional probability distribution. These are key features of "common sense" reasoning, which as we review in Section 6.3, is a central problem in Artificial Intelligence. We will discuss some of the main issues that arise in such common sense reasoning, and see an initial glimpse of how unified learning and reasoning can address some of these challenges. Finally, in Sections 6.5 and 6.6, we will introduce problems in the framework that capture a kind of "abductive" inference, where we seek to identify properties that might explain or justify some condition, by formulating such inference as finding a suitable event to condition on under our semantics. In the following chapters, we will continue to study these problems, but without assuming the examples are complete assignments.

6.1 Reasoning with examples.

Recall that as we discussed in Section 2.4, because we cannot hope to learn knowledge with a guarantee of classical validity, we consider a weaker semantics of $1-\epsilon$-validity with respect to a probability distribution D on possible worlds, for some value ϵ (strictly) between 0 and 1. Perhaps the simplest approachable relaxation of the classical entailment task that we could consider is the following:

Definition 6.1 The *approximate reasoning with examples* task for a knowledge representation family $\mathcal{Q}$ with respect to the probability distribution D on possible worlds is, given a set of *example* possible worlds of $\vec{x}^{(1)}, \ldots, \vec{x}^{(m)}$ each independently drawn from D for some m, a parameter ϵ, and a *query* representation $q \in \mathcal{Q}$, distinguish q that are perfectly valid (such that $\Pr_{\vec{x}\in D}[q(\vec{x})] = 1$) from q that are not $1-\epsilon$-valid (such that $\Pr_{\vec{x}\in D}[q(\vec{x})] < 1-\epsilon$); that is, an algorithm should *accept* q that are perfectly valid and *reject* q that are not $1-\epsilon$ valid. (If q is $1-\epsilon$-valid but not perfectly valid, both accept and reject are acceptable outputs.)

A few remarks concerning this definition are in order: first, we will only be able to obtain algorithms that solve the approximate reasoning with examples task with high probability. That is, for a given number of examples m, there will be some confidence parameter $\delta > 0$ for which our answers are only guaranteed to solve the task with probability $1-\delta$ over the draw of m examples. Indeed, as with learning more generally, there is always a nonzero chance that, say, we observe the same possible world m times and then there is little hope of solving the task.

Second, there is a *gap* between $1-\epsilon$ and 1 ($\epsilon > 0$ strictly). It is not hard to see that if we have a distribution D_ϵ for which some q is $1-\epsilon$-valid and D_0 for which q is perfectly valid, supposing that $D_\epsilon \longrightarrow D_0$ as $\epsilon \to 0$, the joint distribution of m independent examples from D_ϵ also approaches that of m examples from D_0. Since the total variation distance (essentially $1/2$ the 1-norm) gives the optimal distinguishing probability between these distributions, we need m to be $\Omega(\frac{1}{\epsilon})$ to have any hope of solving this task. As $\epsilon \to 0$, $m \to \infty$.

Third, we can interpret the solution of the task as follows: if the algorithm accepts, then with some confidence determined by m, we can infer that the query q is *at least* $1-\epsilon$-valid with respect to the probability distribution D. On the other hand, if the algorithm rejects, then we can infer (again with some confidence given by m in general) that the query q is not perfectly valid for D; this corresponds to q not being valid for the *support* of D, i.e., the possible worlds with nonzero probability mass in D. D in this case is playing a role analogous to the knowledge base in the classical reasoning task, specifying a set of possible worlds for which the query q should be true. But, in this relaxed version of the task, we are willing to tolerate q only being true of a $1-\epsilon$ fraction of the possible worlds, as quantified by D.

Indeed, the task is set up *without* a knowledge base given as part of the input. It would seem very natural to attempt to solve the problem by first trying to learn some representations $h_1, \ldots, h_k$ from the examples in an attempt to characterize D. Such an approach is complicated by the fact that all of the examples we are given are members of the support of D; in other words, if we are thinking of the support of D as a "label" for the examples, then all of the examples have label 1 (meaning the example is in the support of D). Although it is not impossible to characterize a significant fraction of D this way, it does require a prohibitively number of examples. More importantly, this is not necessary to solve the approximate reasoning task.

Theorem 6.2 *Let $\mathcal{Q}$ be a family of efficiently evaluatable knowledge representations with B-bit representations. There is an algorithm that, for any given $\delta > 0$, solves the approximate reasoning with examples task with probability $1-\delta$ using $m \geq \Omega(\frac{1}{\epsilon}(B + \log\frac{1}{\delta}))$ examples in polynomial time (in the number of attributes n, B, and m).*

Proof The algorithm is very simple: given q, we evaluate q on all m examples, and accept iff q is true on all of them. Otherwise the algorithm rejects. Since we have assumed that $\mathcal{Q}$ is efficiently evaluatable, the algorithm is polynomial time as promised. To see that it solves the approximate reasoning with examples task, observe that for any fixed $q \in \mathcal{Q}$, if q is perfectly valid with respect to D, then the algorithm indeed accepts; by contrast, if q is not $1-\epsilon$-valid, then q is only true on each example with probability

less than $1-\epsilon$. So the probability that it is true on all m examples is at most $(1-\epsilon)^m$. For $m \geq \frac{1}{\epsilon}((B+1)\ln 2 + \ln\frac{1}{\delta})$,

$$\begin{aligned}(1-\epsilon)^m &\leq e^{-\epsilon m} \\ &\leq e^{-((B+1)\ln 2+\ln\frac{1}{\delta})} \\ &= \delta/2^{B+1}.\end{aligned}$$

As there are less than 2^{B+1} representations of length B, by a union bound over Q, we have that the probability that the algorithm does not solve the task for any $q \in Q$ is at most δ, as needed. □

Example 6.1: Let Q be the class of 3DNF queries on a set of n attributes. We know that members of q can be represented by $O(n^3)$ bit vectors, indicating for each possible term of up to three literals whether or not it is included in q; members of q can also be evaluated on a possible world in linear time in the representation size. So now, for any probability distribution D on possible worlds, approximate query evaluation for 3DNFs can be solved in polynomial time using $O(\frac{1}{\epsilon}(n^3 + \log\frac{1}{\delta}))$ examples from D: we can distinguish valid 3DNF (on the support of D) from 3DNF that are not $1-\epsilon$-valid with respect to D. Observe that by contrast, the problem of deciding classical validity of a 3DNF is NP-hard.

So, in a typical setting for learning – in which one has access to complete examples from the distribution – a natural analogue of the intractable, idealized reasoning task becomes trivial for a wide variety of representations. We did not need to limit our queries to those that could be captured by a weak proof system.

6.2 Exception-tolerance and agnostic learning.

In our model of the learning task, we had assumed that the target property could be represented by the family of representations we were using. Since, as we saw, it turns out that the kinds of properties that are learnable are rather weak, it is natural to reconsider this assumption: suppose that the representations can only approximate the target property. What can we learn then?

Indeed, this problem arises naturally when we are seeking to learn *natural concepts* rather than formal properties: most famously, the philosopher Wittgenstein proposed that there is no single attribute common of all things we call "games": there are families of games that involve pieces and a board, families that involve points and scores, families with multiple players, families that involve balls or other kinds of equipment, and so on, but we can always propose a new game that does not have one of these features or another. But even beyond the particularly elusive notion of a "game," almost every natural concept is characterized by properties that are typical but not strictly essential for counting as an instance. For example, suppose we tried to define what is an "orange." We can say that it is a citrus fruit grown from an orange tree, though

this relies on the notion of an orange tree, that is itself hard to describe, and we often recognize oranges in the absence of their trees. Rather, we may say that an orange is a round, sweet fruit with an orange colored flesh and skin. But then there are certainly exceptions to these properties: navel oranges have a distinct lumpy defect to the shape and a sliced orange is still an orange; some oranges are unpalatably tart; some oranges may (still) have green skin or red flesh; and so on. These variations are still considered to be oranges. At the same time, our stab at describing an orange is pretty accurate for typical oranges, and is a useful characterization when considering the role of oranges in human life. We would like to be able to acquire and use such descriptions, even though exceptional instances exist.

The natural objective here is to find the most accurate representation in a given class, or to at least obtain a representation that is nearly as accurate. This task is called *agnostic learning* because the model is agnostic to the form of the actual properties to be predicted:

Definition 6.3 Let C_n and $\mathcal{H}_n$ be classes of representations on the same n attributes. We say that C_n is *(improperly) agnostically PAC-learnable* using $\mathcal{H}_n$ in time $t(n, \frac{1}{\epsilon}, \frac{1}{\delta})$ and sample complexity $m(n, \frac{1}{\epsilon}, \frac{1}{\delta})$ if there is an algorithm such that for any m examples drawn independently from a distribution D on n attribute possible worlds and labeled by some $b \subseteq \mathcal{X}^n$, given these examples and the parameters ϵ and δ, with probability at least $1 - \delta$ outputs a representation $h \in \mathcal{H}_n$ such that

$$\Pr_{\vec{x} \in D}[h(\vec{x}) = b(\vec{x})] \geq \max_{c \in C_n} \Pr_{\vec{x} \in D}[c(\vec{x}) = b(\vec{x})] - \epsilon.$$

If C_n is improperly agnostically PAC-learnable in polynomial time and sample complexity by any class $\mathcal{H}_n$ for which evaluation can be computed in polynomial time, then we simply say that C_n is *efficiently (improperly) PAC-learnable*. If $C_n = \mathcal{H}_n$, we say the learning is *proper* rather than improper.

On the reasoning side, the following is a natural formulation of our more general task:

Definition 6.4 The *validity estimation* task for a knowledge representation family Q with respect to the probability distribution D on possible worlds is, given a set of *example* possible worlds of $\vec{x}^{(1)}, \ldots, \vec{x}^{(m)}$ each independently drawn from D for some m, a parameter ϵ, and a *query* representation $q \in Q$, is to return an *estimate* $\hat{p}$ such that

$$|\Pr_{\vec{x} \in D}[q(\vec{x})] - \hat{p}| \leq \epsilon.$$

In particular, for the family of representations of properties $\mathcal{H}_n$, we could consider the family Q of formulas of the form $h(\vec{x}) = b$ for $h \in \mathcal{H}_n$ on the set of attributes $X_1, \ldots, X_n, B$, i.e., treating the label as an additional attribute. The validity of these formulas corresponds precisely to the accuracy of h as a representation of b with respect to the distribution D. Note that we can trivially transform a knowledge representation for $\mathcal{H}_n$ to a knowledge representation for Q using the same encodings

(via the map $h \mapsto [h(\vec{x}) = b]$). Thus, validity estimation will give us an analogue to Occam's Razor for agnostic learning.

Naturally, we now use the empirical performance of a candidate representation on the data to estimate its degree of validity on the distribution D over possible worlds. Since we have assumed that the examples are drawn independently, we have the following standard bound that guarantees that the estimates we obtain are pretty close with exponentially high probability:

Theorem 6.5 (Hoeffding bound) *Let $X_1, \ldots, X_m$ be independent random variables taking values in the range $[0, 1]$ with $\mathbb{E}[\frac{1}{m}\sum_{i=1}^{m} X_i] = p$, and let $\varepsilon > 0$ be given. Then*

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m} X_i > p + \varepsilon\right] \leq e^{-2m\varepsilon^2}$$

Considering $X_i' = 1 - X_i$ gives another set of random variables taking values in the range $[0, 1]$, and then Hoeffding's bound gives a lower bound for the original X_i variables as well:

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m} X_i < p - \varepsilon\right] \leq e^{-2m\varepsilon^2}.$$

Thanks to the exponentially high confidence obtained by Hoeffding's bound, empirical estimates efficiently solve the validity estimation task:

Theorem 6.6 *Let Q be a family of efficiently evaluatable knowledge representations with B-bit representations. There is an algorithm that, for any given $\delta > 0$, solves the validity estimation task with probability $1 - \delta$ using $m \geq \Omega(\frac{1}{\epsilon^2}(B + \log\frac{1}{\delta}))$ examples in polynomial time (in the number of attributes n, B, and m).*

Proof For any $q \in Q$, for $m \geq \frac{1}{2\epsilon^2}((B+1)\ln 2 + \ln\frac{2}{\delta})$, noting that $\mathbb{E}_{\vec{x}\in D}[q(\vec{x})] = \Pr_{\vec{x}\in D}[q(\vec{x})]$, Hoeffding's bound gives us that

$$\Pr_{\vec{x}^{(1)},\ldots,\vec{x}^{(1)}\in D^m}\left[\frac{1}{m}\sum_{i=1}^{m} q(\vec{x}^{(i)}) > \Pr_{\vec{x}\in D}[q(\vec{x})] + \varepsilon\right] \leq e^{-2m\varepsilon^2} \leq \delta/2^{B+2}$$

$$\text{and } \Pr_{\vec{x}^{(1)},\ldots,\vec{x}^{(1)}\in D^m}\left[\frac{1}{m}\sum_{i=1}^{m} q(\vec{x}^{(i)}) < \Pr_{\vec{x}\in D}[q(\vec{x})] - \varepsilon\right] \leq e^{-2m\varepsilon^2} \leq \delta/2^{B+2}.$$

So, by a union bound over these two events and the (fewer than) 2^{B+1} members of Q, taking $\hat{p} = \frac{1}{m}\sum_{i=1}^{m} q(\vec{x}^{(i)})$ (the *empirical validity* of q) indeed gives

$$|\hat{p} - \Pr_{\vec{x}\in D}[q(\vec{x})]| \leq \epsilon$$

with probability at least $1 - \delta$. So long as q is efficiently evaluatable, $\hat{p}$ can indeed be computed in polynomial time as promised. □

As with approximate reasoning with examples, we see we have a polynomial-time algorithm for the more general task of validity estimation for 3CNF representations with respect to a probability distribution D. This again stands in contrast to the intractability of deciding the classical validity of 3CNFs.

Back to agnostic learning, it immediately follows that any representation $\hat{h}$ for which $[\hat{h}(\vec{x}) = b]$ has maximum empirical validity is a solution to the agnostic learning task, and all such solutions have near-maximum empirical validity. Thus, (approximately) maximizing the empirical validity is analogous to the problem $\text{Consis}^B_{\mathcal{H}}$ for the standard learning task:

Corollary 6.7 *Let $\mathcal{H}$ be a family of efficiently evaluatable knowledge representations with B-bit representations. Any representation $\hat{h}$ for which $[\hat{h}(\vec{x}) = b]$ has maximum empirical validity on a set of $m \geq \Omega(\frac{1}{\epsilon^2}(B+\log\frac{1}{\delta}))$ examples is a solution to the proper agnostic learning task for $\mathcal{H}$ with probability $1-\delta$; and conversely, with probability $1-\delta$, all of the solutions to the proper agnostic learning task have empirical validity within 2ϵ of optimal.*

Proof Suppose $m \geq \frac{2}{\epsilon^2}((B+1)\ln 2 + \ln\frac{2}{\delta})$. First consider any representation h^* that maximizes $\Pr_{\vec{x}\in D}[h^*(\vec{x}) = b(\vec{x})]$. With probability $1-\delta$, the empirical validity of $[h^*(\vec{x}) = b]$ is at least $\Pr_{\vec{x}\in D}[h^*(\vec{x}) = b(\vec{x})] - \epsilon/2$ and the empirical validity of $[\hat{h}(\vec{x}) = b]$ is at most $\Pr_{\vec{x}\in D}[\hat{h}(\vec{x}) = b(\vec{x})] + \epsilon/2$. Since $\hat{h}$ maximizes the empirical validity, we therefore have

$$\Pr_{\vec{x}\in D}[\hat{h}(\vec{x}) = b(\vec{x})] + \epsilon/2 \geq \Pr_{\vec{x}\in D}[h^*(\vec{x}) = b(\vec{x})] - \epsilon/2$$

and thus $\hat{h}$ is indeed sufficient. As for the moreover part, consider any h' that is a solution to the agnostic learning task, i.e., attaining

$$\Pr_{\vec{x}\in D}[h'(\vec{x}) = b(\vec{x})] \geq \Pr_{\vec{x}\in D}[h^*(\vec{x}) = b(\vec{x})] - \epsilon.$$

We also have that the empirical validity of h' is at least $\Pr_{\vec{x}\in D}[h^*(\vec{x}) = b(\vec{x})] - \frac{3}{2}\epsilon$, where the empirical validity of $\hat{h}$ is at most $\Pr_{\vec{x}\in D}[\hat{h}(\vec{x}) = b(\vec{x})] + \epsilon/2$. Since h^* maximizes the actual probability, $\Pr_{\vec{x}\in D}[\hat{h}(\vec{x}) = b(\vec{x})] \leq \Pr_{\vec{x}\in D}[h^*(\vec{x}) = b(\vec{x})]$. Hence,

$$\sum_{i=1}^{m}[\hat{h}(\vec{x}^{(i)}) = b(\vec{x}^{(i)})] - \frac{1}{2}\epsilon \leq \sum_{i=1}^{m}[h'(\vec{x}^{(i)}) = b(\vec{x}^{(i)})] + \frac{3}{2}\epsilon$$

establishing the second part. □

Unfortunately, this is the end of the good news. The computational problem of maximizing agreement with the examples is much harder than finding a representation that is correct on all of the examples (when they exist). Even our most basic class of conjunctions poses an intractable agnostic learning problem—see Exercise 6.3 and the chapter notes. Indeed, even *improper* agnostic learning of conjunctions would yield a surprising algorithm for learning DNF representations, and improper agnostic learning of halfspaces would violate the approximate shortest vector assumption in cryptography. (The proof is based on the "boosting" technique – see Section A.6 and Exercises A.9 and A.10.) Even for conjunctions, the only tractable algorithms we possess only provide a representation that achieves an error rate of at best $O(n^{1/3}\min_{h\in\mathcal{H}}\Pr[h(\vec{x}) \neq b(\vec{x})])$. (Exercise 6.4 asks for a rather worse $O(n\cdot\min_{h\in\mathcal{H}}\Pr[h(\vec{x}) \neq b(\vec{x})])$ error

guarantee.) Although such guarantees may be reasonable when a very accurate representation exists, the dependence on the number of attributes is generally prohibitive for other kinds of rough representations that we might nevertheless hope to discover.

6.3 Common sense.

Arguably, the central problem of Artificial Intelligence is to capture *common sense*. Common sense here refers to a body of tacitly shared knowledge that humans possess about their world. Humans rely on this knowledge to get around the world, communicate with one another, and more broadly engage in society. We do not know how humans come to possess this knowledge: that is, how much and which of it is learned from experience with the world, and which of it is simply innate. Indeed, a key challenge in studying common sense is its tacit nature: humans generally take this knowledge for granted, and it is challenging to even give a complete account of its contents.

Nevertheless, Cognitive Science has identified several domains of "core knowledge" that seem to comprise a large portion of such common sense knowledge. Humans are not proficient in these domains at birth, but proficiency develops at a predictable schedule during the early months and years of life. Concretely, these domains include the following:

1. **Time.** At different *times*, the world may be in different states. Time has an *ordering*: between any two points in time, one comes first. Time progresses forward. Events may have *causes*, that occur at an earlier point in time.
2. **Objects.** Objects *persist*, meaning they continue to exist even when they are not observed, and when we observe them again they are the same object. More generally, there is a notion of object *identity*, where if we see two objects at the same time, they are not the same object.
3. **Space.** Two objects cannot occupy the same space at the same time, but rather they will *collide*. But, one object can be *contained* inside another object (and might be removed from it, etc.).
4. **Naïve Physics.** Objects remain in the same place unless they are moved. Objects fall down unless they are supported, etc.
5. **Naïve Psychology.** There are other *agents* that may spontaneously act to achieve *goals*, and act in accordance with their *beliefs*. Their disposition to act in certain ways may change with their *emotional state*, which in turn may change in response to events in the world.
6. **Naïve Quantities.** Explicit representations of *small numbers*, e.g., 1–5. *Comparisons* such as "more" and "less."

We stress again, this list of domains and examples is not exhaustive. One of the core challenges of common sense is to determine exactly the knowledge it encompasses.

Much of this common sense knowledge is captured by the ability to draw certain kinds of inferences about the world: in particular many of these inferences follow from

the *frame axiom* that asserts (informally) that the state of objects remains the same unless they are changed somehow, where the change corresponds to some explicit event. Inferences drawn via the frame axiom, in contrast to classical reasoning, may be *nonmontonic* in the following sense: we might need to withdraw a conclusion in light of additional knowledge. For example, if we place a book in a bag, the frame axiom permits us to infer that the book is still in the bag. But, if we later learn that someone took the book out of the bag earlier, then wish to conclude instead that the book was *not* in the bag at the time we were asked. This is in contrast to classical knowledge as introduced in Section 2.1, where additional knowledge only reduces the set of possible worlds, and hence any property of the set of possible worlds that was true with one state of knowledge must remain true when additional knowledge is acquired. In nonmonotonic reasoning, new knowledge can *defeat* another inference. The frame axiom is just one of many such rules of nonmonotonic inferences. We will see other examples of properties we presume to hold of "typical" objects in the next section, and we will discuss general mechanisms for nonomonotonic reasoning in Chapter 10.

In addition to our habitual obliviousness to the appearance and application of our common sense knowledge, approaches to common sense reasoning must contend with four significant problems that arose in attempts to adapt classical logic to common sense reasoning:

1. **Qualification.** In our classical logical rules, as used in chaining for example, the body of the rule contains a precise list of the facts that must hold in order to apply the rule. But, it is all but impossible to formulate such a precise list of all of the conditions that must hold for some inference to be valid in the real world. For example, we might want to say that when we flip a light switch, the lights turn on. But this is only true if the light bulbs are good—or indeed, that there are light bulbs installed at all. And there must not be a power outage. And the wiring inside the walls must be connected. And the lights must not be covered up. And so on... a creative reader can doubtless come up with additional conditions that must hold.
2. **Ramification.** One new fact may yield many consequences. When a train departs from a station, it has the consequence that none of the passengers on board are at the station any longer. We cannot afford a method for which the process of incorporating a new fact has time complexity that scales with the number of such consequences.
3. **Elaboration tolerance.** Similar to ramification, we should be able to incorporate new rules efficiently, with time complexity that is independent of the size of our knowledge base. To see why this is an issue at all, consider the encoding of the frame axiom we used in our clausal encoding of classical planning in Section 5.6: it explicitly enumerates all of the actions that change the state of a fluent. So, if we learn about a new effect or a new action, the frame axioms would need to be rewritten. Such a solution would, in general, not be "elaboration tolerant."
4. **Scale.** Common sense comprises a vast body of knowledge. Rough estimates range from 5×10^8–3.4×10^9 bits or 10^7–10^8 facts/rules/etc. The CYC project famously set out to hand-engineer such a knowledge base, but still misses a

significant (perhaps majority) fraction of common sense despite several decades of effort. At the other extreme, methods that mine the Internet, Wikipedia, etc. are capable of acquiring a sufficient volume of data, but this data is surely biased away from common sense knowledge—we generally see no need to report common sense facts, write Wikipedia articles about them, and so on. Beyond the difficulty of acquiring such a large knowledge base, effectively using it also poses a daunting challenge. We cannot afford methods to answer queries that scale with the size of this knowledge base.

6.4 Nonmonotonicity of conditional reasoning with examples.

A significant portion of our common sense knowledge concerns typical properties of objects. Such "typical properties" have a nonmonotonic character. For example, we know that birds generally can fly. If we are asked to consider a bird with no further qualification, we generally presume that it can fly. But, if we are informed that the bird in question is a penguin, we will wish to retract our position on its ability to fly. A natural extension of reasoning with examples elegantly captures such inferences.

In classical logic we represent a claim such as "birds fly" as the implication "b is a bird" $\Rightarrow$ "b flies", which is equivalent to the clause $\neg$"b is a bird" $\vee$ "b flies". In our formulation of PAC-semantics, this leaves something to be desired: if we are relatively unlikely to encounter birds in the distribution D on possible worlds under consideration, the clause $\neg$"b is a bird" $\vee$ "b flies" is true with high probability for vacuous reasons. We would more appropriately consider *only* the *subset* of possible worlds that feature birds, and then consider their propensity to fly—see, for example, Table 6.1. Since the subset that remain still occur with probability proportional to their original probability w.r.t. D, this is a possible world drawn from the *conditional probability distribution* $D|\text{bird}$, i.e., $\Pr_{\vec{x}\in D}[\vec{x}|\text{bird}] \stackrel{\text{def}}{=} \frac{\Pr_{\vec{x}\in D}[\vec{x}\wedge\text{bird}(\vec{x})]}{\Pr_{\vec{x}\in D}[\text{bird}(\vec{x})]}$.

Definition 6.8 The *contextual validity estimation* task for a query knowledge representation family Q and context knowledge representation family C with respect to the probability distribution D on possible worlds is, given a set of *example* possible worlds of $\vec{x}^{(1)}, \ldots, \vec{x}^{(m)}$ each independently drawn from D for some m, a parameter ϵ, a *context* $c \in C$, and a *query* representation $q \in Q$, is to return an *estimate* $\hat{p}$ such that

$$|\Pr_{\vec{x}\in D}[q(\vec{x})|c(\vec{x})] - \hat{p}| \leq \epsilon.$$

Since $\Pr_{\vec{x}\in D}[q(\vec{x})|c(\vec{x})]$ is simply another probability distribution on possible worlds, provided that a sufficient number of our example possible worlds satisfy the context c, Theorem 6.6 may still be applied, and simple counting suffices.

Example 6.2: Consider the set of examples shown in Table 6.1, and suppose that we consider queries with estimated validity at least $2/3$ to be "typically true." (In this small example, we set aside considerations of how the conclusion generalizes for the moment. We would need many more examples to have high confidence, of course.)

bird	fly	has_beak	red	purple	penguin
1	1	1	0	0	0
1	1	1	1	0	0
1	0	1	0	0	1
1	0	1	0	0	1
1	1	1	0	0	0
1	1	1	1	0	0
1	1	1	0	0	0

Table 6.1 A portion of set of example possible worlds filtered to include only those with "bird" true. Note that additionally filtering to include only those with "has_beak" also true yields the same set of examples.

We can see that since "fly" is true in $5/7 > 2/3$ of the examples, we would regard it as typically true that birds fly. Similarly, since "has_beak" is true in all 7 examples, we also would regard it as typically true that birds have beaks. But, since "red" and "penguin" are each only true in $2/7 < 2/3$ of the examples, we do not presume that typical birds are red or penguins.

Here, we incorporate knowledge by further filtering the set of possible worlds to include only those in which the additional criteria are satisfied:

bird	fly	has_beak	red	purple	penguin
1	1	1	1	0	0
1	1	1	1	0	0

Table 6.2 The set of example possible worlds of Table 6.1 filtered to include only those with "red" and "bird" true.

Example 6.3: Continuing, let's suppose that we are asked to consider red birds. We filter the set of examples from Table 6.1 to include only those in which "red" is also true, yielding the set of examples shown in Table 6.2. We see that "red" is now true in the remaining worlds, as desired, so we conclude that red birds are, indeed, typically red. They still furthermore have beaks, fly, and are not penguins.

Since validity estimation permits tolerating a nonzero fraction of counterexamples, filtering a subset of the examples may yield a smaller subset in which the formerly-exceptional cases dominate. This results in a "nonmonotonic" reversal of a previous conclusion:

bird	fly	has_beak	red	purple	penguin
1	0	1	0	0	1
1	0	1	0	0	1

Table 6.3 The example possible worlds of Table 6.1 filtered to include only those with "bird" and "penguin" true. Note that "fly" is false in these examples.

bird	fly	has_beak	red	purple	penguin

Table 6.4 The example possible worlds of Table 6.1 filtered to include only those with "bird" and "purple" true. No examples remain.

Example 6.4: Now, suppose that instead we consider penguins: we filter the set of examples of Table 6.1 to include only those in which "penguin" is true, yielding Table 6.3. In addition to "penguin" being true, only examples in which "fly" is false remain. Thus, whereas we originally had that typical birds fly, typical penguins do not fly.

This approach also allows us to sensibly handle cases that are exceptional in multiple ways: if too few examples remain (in particular, if no examples remain) then we can and should refrain from drawing vacuous conclusions about queries.

Example 6.5: Let's suppose we are asked to consider purple birds. Filtering the set of examples of Table 6.1 to include only those with "purple" true yields *no* examples (Table 6.4). Since we possess too few examples of such worlds, we don't presume any typical properties of purple birds.

Example 6.6: A famous challenging example for declarative representations is the "Nixon diamond." We suppose we are told that Quakers are typically pacifists, and Republicans are typically not pacifists. But, what about Republican Quakers? Are they pacifists or not? These typical properties are in conflict, and this needs to be resolved somehow. As with the purple birds, reasoning with examples would refrain from drawing conclusions without sufficiently many examples.

Redundant conditions are also well-behaved when reasoning with examples:

Example 6.7: Since "has_beak" was true in all of the examples, filtering additionally on the condition that "has_beak" is true does not alter any of the conclusions we draw: we obtain precisely the same estimates of validity.

We'll conclude by briefly discussing how reasoning with examples addresses the four challenges for common sense reasoning:

1. **Qualification.** Since we permit a fraction $\epsilon > 0$ of examples to be exceptional for *any* reason, we do not need to specify the kinds of exceptions in advance. Note that this indeed relies on the tolerant, validity estimation approach to reasoning with examples.
2. **Ramification.** The examples that satisfy a given condition already satisfy all of the facts that are entailed by that condition. Queries about those facts will be regarded as true, without us having needed to explicitly consider all of the possible entailed conclusions.
3. **Elaboration tolerance.** We can incorporate new knowledge simply by adding new examples, with the time complexity scaling only with the number of examples added. We don't need to modify the existing examples.
4. **Scale.** This is perhaps the most important observation: there may be an exponential number of properties that hold of a given set of examples. The typical properties of these examples are not encoded explicitly, but nevertheless can be efficiently determined by a count on a small subset of the examples. Indeed, in general we do not even need to consider all of the examples; a relatively small subset would do to answer a single query.

So, reasoning with examples allows us to reason about the properties of "typical" objects of a given kind, and demonstrates that it can correctly retract these inferences when additional knowledge undermines this presumption of typicality. It furthermore meets the four challenges outlined above. We might thus be tempted to speculate that it could capture all of common sense reasoning. It almost surely does not. Reasoning with examples here relies on examples that consist of complete descriptions of possible worlds, whereas some of our important common sense rules concern states of the world that are *not* observed. For example, persistence of objects is a rule that explicitly concerns objects that cannot be seen. Many of the inferences we would like to be able to make using frame axioms similarly concern the state of the world remaining as it was *when we last observed it*. Naïve psychology, meanwhile, presupposes unobservable mental states of other agents, and suggests rules for reasoning about them. While one might be tempted to speculate further that these mental states might be learned as a means of "explaining" the observable behavior of these other agents, this seems implausible. We currently do not know how to learn DNF or CNF conditions, and these are widely conjectured to not be learnable by efficient algorithms. Recall that we saw in Sections 4.7 and 4.8 that the families of rules we can learn have rather limited expressive power even if we allow ourselves complete freedom in how we represent the

knowledge. Furthermore, even if we allow ourselves to approximate the conditions, we noted in Section 6.2 that even conjunctive conditions seem not be be approximable unless they were quite accurate to begin with. We seem to require another, perhaps innate mechanism for drawing such inferences. We will return to considering some possible approaches in Chapter 10.

6.5 Precondition search.

Reasoning with examples can enable us to draw inferences about properties of "typical" objects of some kind. But, suppose we are simply trying to get around in the world. We will often find ourselves in situations that are not perfectly identical to any state of the world that we have not previously encountered: no matter what prior memories we have, due to the richness of the world around us, *some* features of our surroundings will differ, so if we are looking for prior examples with perfect agreement in order to use reasoning with examples to guide our expectations, we will be disappointed. We need some way of deciding *which* features of the present world to focus on to use as our filtering rule, in order to make predictions with any confidence at all.

Whereas we would like to find the precondition that leads to a correct prediction regarding the validity of a given query, this is begging the question: we seek to make a prediction because we do not know what that prediction ought to be. A simple approach, that we will pursue for the moment, is to seek the most *"credulous"* precondition. That is, the precondition that provides the best support for the query—the best evidence in favor of the query. This has some evident flaws as a means of prediction, as we'll return to shortly, but such a formulation is nevertheless within reach at the moment.

Definition 6.9 *Credulous precondition search* for a query knowledge representation family $\mathcal{Q}$ and precondition knowledge representation families $\mathcal{C}$ and $\mathcal{H}$ with respect to the probability distribution D on possible worlds is, given a set of *example* possible worlds of $\vec{x}^{(1)}, \ldots, \vec{x}^{(m)}$ each independently drawn from D for some m, a *context* possible world $\vec{x}^*$, parameters $\epsilon, \mu, \gamma > 0$, and a *query* $q \in \mathcal{Q}$ such that there exists some precondition $c \in \mathcal{C}$ with $c(\vec{x}^*)$ true, $\Pr_{\vec{x} \in D}[c(\vec{x})] \geq \mu$, and $\Pr_{\vec{x} \in D}[q(\vec{x})|c(\vec{x})] = 1$, return a precondition representation $h \in \mathcal{H}$ such that $h(x^*) = 1$, $\Pr_{\vec{x} \in D}[h(\vec{x})] \geq (1-\gamma)\mu$, and $\Pr_{\vec{x} \in D}[q(\vec{x})|h(\vec{x})] \geq 1 - \epsilon$. If $\mathcal{C} = \mathcal{H}$, this is the *proper* version of the task.

This is a *realizable* version of the task, where the ideal precondition c perfectly entails the query q, $\Pr_{\vec{x} \in D}[q(\vec{x})|c(\vec{x})] = 1$. We could also have considered an agnostic version where, rather than $\Pr_{\vec{x} \in D}[q(\vec{x})|h(\vec{x})] \geq 1 - \epsilon$, we "merely" ask $\Pr_{\vec{x} \in D}[q(\vec{x})|h(\vec{x})] \geq \Pr_{\vec{x} \in D}[q(\vec{x})|c(\vec{x})] - \epsilon$, but even the realizable version of the task is already quite challenging:

Theorem 6.10 *If there is a polynomial-time algorithm for proper credulous precondition search for conjunctive preconditions where the queries either include all 2CNFs,*

DNFs, or may use an attribute that is not used in the precondition, then there are randomized polynomial-time algorithms for NP.

Proof We give a reduction from Independent Set. Recall that here the input is an undirected graph $G = (V, E)$ and an integer k, and the objective is to determine whether there is a subset of k vertices such that no pair have an edge in G. We note that the problem remains NP-complete even if we only consider $k = c|V|$ for a fixed rational number $0 < c < 1$.

Given G, we construct the following set of $m = |V| + |E|$ examples on $n = |V|$ attributes: For each ith vertex, $i = 1, \ldots, |V|$, we create an example:

$$(1, 1, \ldots, 1, \overbrace{0}^{x_i}, 1, \ldots, 1).$$

For each edge $\{i, j\} \in E$, we create an example:

$$(1, 1, \ldots, 1, \overbrace{0}^{x_i}, 1, \ldots, 1, \overbrace{0}^{x_j}, 1, \ldots, 1).$$

Now consider the uniform distribution on this set of examples; clearly we can sample from this distribution in time $O(n)$ per example. Consider, respectively, the queries given by the DNF $\bigvee_{i=1}^{n} \bigwedge_{j \neq i} x_j$, the 2CNF $\bigwedge_{\{i,j\} \in E} x_i \vee x_j$, or by an additional $n+1$th attribute that agrees with these two formulas on all examples; indeed, these queries all evaluate to true on the first set of examples for the vertices, and false on the second set of examples for the edges. A solution to the credulous precondition search task for conjunctive preconditions for the respective query with context $\vec{x}^* = (1, \ldots, 1)$, $\epsilon = \frac{1}{2m}$, $\mu = \frac{\lceil c|V| \rceil}{m}$, and $\gamma = \frac{1}{2\lceil c|V| \rceil}$ must be a conjunction h that cannot contain negative literals, or else h would not be satisfied by $\vec{x}^*$.

We claim that the set S of vertices corresponding to indices *not* in h is an independent set: indeed, if any two vertices i and j joined by an edge were omitted from h, then h would be satisfied on the corresponding edge example where $q(\vec{x}) = 0$ and then $\Pr[q(\vec{x})|h(\vec{x})] \leq 1 - \frac{1}{m} < 1 - \frac{1}{2m}$. Furthermore, for each positive literal not in h, the corresponding vertex example must satisfy h. Hence, $\Pr[h(\vec{x})] = \frac{|S|}{m}$, and $\Pr[h(\vec{x})] \geq (1 - \frac{1}{2\lceil c|V| \rceil})\frac{\lceil c|V| \rceil}{m}$ iff

$$|S| \geq (1 - \frac{1}{2\lceil c|V| \rceil})\lceil c|V| \rceil = \lceil c|V| \rceil - \frac{1}{2}.$$

But since $|S|$ and $\lceil c|V| \rceil$ are integer, $|S| \geq \lceil c|V| \rceil - \frac{1}{2}$ iff $|S| \geq \lceil c|V| \rceil$ (and $|S| \geq c|V|$ iff $|S| \geq \lceil c|V| \rceil$). Thus the algorithm succeeds at finding such an h iff there is an independent set of size $c|V|$. Since $\frac{1}{\epsilon}$, $\frac{1}{\mu}$, and $\frac{1}{\gamma}$ are all polynomial in the size of G, we see that this would indeed give a polynomial time algorithm for Independent Set. □

Actually, we can obtain more from this reduction: since it is known that no algorithm can approximate the size of the largest independent set to within a $O(|V|^{1-\epsilon})$ factor for any $\epsilon > 0$, no algorithm can obtain h with $\Pr[h(\vec{x})] \geq \Omega(\frac{\mu}{n^{1-\epsilon}})$ for any $\epsilon > 0$.

While this is unfortunate, since conjunctions would be arguably the most natural representation, it turns out that we *can* efficiently solve the task for disjunctions.

(Clearly, unlike the usual supervised learning task, we cannot reduce the problem for conjunctions to the problem for disjunctions here.) And moreover, we have essentially already seen the algorithm: Elimination (Algorithm 1).

Input: Query $q(\vec{x})$, examples $\vec{x}^{(1)} \ldots, \vec{x}^{(m)}$, context $\vec{x}^*$.
Output: A disjunction h with $h(x^*) = 1$ or FAIL.
begin
  Initialize h to contain all $2n$ possible literals on the n attributes.
  `// Ensure negative examples are labeled correctly`
  for $i = 1, \ldots, m$ **do** **if** $q(\vec{x}^{(i)}) = 0$ **then**
    forall $\ell \in h$ **do** **if** $\ell(\vec{x}^{(i)}) = 1$ **then** Remove ℓ from h.
  if $h(x^*) = 0$ **then return** *FAIL*
  else return h.
end

Algorithm 10: Elimination Algorithm for learning a disjunctive precondition

Theorem 6.11 *There is a polynomial time algorithm for the proper credulous precondition search task for disjunctive preconditions and any efficiently evaluatable query representations with probability* $1 - \delta$ *given* $m = \Omega(\frac{1}{\mu\epsilon}(n + \log_2 \frac{1}{\delta}))$ *examples.*

Proof Suppose that there is a disjunctive precondition $c(\vec{x})$ with $c(\vec{x}^*) = 1$, $\Pr[q(\vec{x})|c(\vec{x})] = 1$ and $\Pr[c(\vec{x})] \geq \mu$. Observe that when Algorithm 10 removes a literal ℓ from h, ℓ cannot have been in c, or else we would have $\Pr[q(\vec{x})|c(\vec{x})] < 1$. Since h initially contains all possible literals, in particular it includes all of the literals in c, and since none are removed, the returned h still contains all of the literals of c. Therefore $\Pr[h(\vec{x})] \geq \Pr[c(\vec{x})] \geq \mu$. Moreover, since $c(\vec{x}^*) = 1$, $h(\vec{x}^*) = 1$ as well.

We now argue that $\Pr[q(\vec{x})|h(\vec{x})] \geq 1 - \epsilon$. Consider any disjunction $\tilde{h}$ that could be output such that $\Pr[q(\vec{x})|\tilde{h}(\vec{x})] < 1 - \epsilon$; then $\Pr[q(\vec{x}) = 0 \wedge \tilde{h}(\vec{x}) = 1] > \epsilon \Pr[\tilde{h}(\vec{x})] \geq \mu\epsilon$. But $\tilde{h}(\vec{x}) = 1$ iff some literal of $\tilde{h}(\vec{x})$ is satisfied on $\vec{x}$, and whenever we draw some $\vec{x}$ such that $q(\vec{x}) = 0$ and $\tilde{h}(\vec{x}) = 1$, such literals are removed from h, meaning that $\tilde{h}$ cannot be output. Since the m examples are independently drawn from D, the probability that $\tilde{h}$ remains a candidate is at most $(1 - \mu\epsilon)^m < \frac{\delta}{2^n}$ for $m \geq \frac{1}{\mu\epsilon}(\ln 3 \cdot n + \ln \frac{1}{\delta})$. Since there are only 3^n possible non-tautologous disjunctions (where one literal from all complementary pairs is surely eliminated on the first example with $q(\vec{x}) = 0$), by a union bound over these, the probability that any such $\tilde{h}$ is produced is at most δ. We finally note that Algorithm 10 is indeed polynomial time since q is efficiently evaluatable. □

We stress that this "credulous" formulation is too liberal to use for prediction directly. Consider for example the agnostic variant of the task with the context $\vec{x}^*$ equal to one of the examples from Table 6.1 where penguin is true with the query fly, tolerating $1/3$ probability of an exception. Then bird – or even bird ∨ has_beak

– are satisfactory (disjunctive) preconditions, since fly is true for at least $2/3$ of the examples satisfying the preconditions. However, for all of the worlds satisfying these preconditions, there is a precondition consistent with the context, given by penguin, for which fly is false on all of the remaining examples. We can think of the preconditions as proposing an *argument* for the query; here we have a counterargument that *attacks* the argument that would support fly. We'll return to these considerations in Chapter 10.

6.6 Learning abductive inference.

The credulous precondition search task may also be viewed as a formulation of *"abduction"*: a form of inference that stands in contrast to the more familiar deductive and inductive inferences we have been considering. Informally, abduction is making plausible guesses that explain or justify some observation. For example, if we see someone walking indoors who is soaked with water, we may guess that they were caught in the rain. The hallmark of an abductive inference is that it is not strictly entailed by the facts: there may have been a plumbing problem, or our poor friend may have fallen into a swimming pool. We may prefer the guess of a sudden downpour because, in general, it might be a much more common occurrence than the alternatives. Abductive inferences underlie much of our interpretation of written stories, pictures, and so on: from a few cues – black clothing and sad faces – we may infer a larger narrative of a funeral, and then guess that some figure described as "absent" is the deceased. (But again, it may be that this is a concert and a "late" friend is merely caught in traffic.)

Aside: Credulous preconditions, as "arguments" for the query, have been used to "explain" classification decisions with simple (easy to understand) representations. When $C = \mathcal{H}$ are conjunctions, this approach is known as *"anchors."* In view of Theorem 6.10, guarantees for worst-case distributions are not possible for conjunctions, and anchors may produce very specific conjunctions for a given query point. Nevertheless, these may still be useful.

In our framework where we are considering a probability distribution over possible states of the world, it is quite natural to formulate abduction as the task of finding a representation of an event that is most likely to occur under this distribution, that entails – or at least, strongly suggests – the observed evidence. If we are interested in such an explanation for a specific world x^*, then if we consider a query representation q defining the set of possible worlds in which the observed evidence is present, the credulous precondition search task is essentially to draw such an abductive inference for x^*: we seek a representation h that is true of x^*, for which q is nearly always true when h is, and of probability near the maximum over representations c that are also true of x^* and entail q. But, in many cases – for example, in text – we have only the given observation q, and not a complete possible world x^*. We are still interested in a proposing a likely, simple hypothesis that entails q. We will call this variant our abduction task:

Definition 6.12 *Abduction* for a query knowledge representation family $\mathcal{Q}$ and hypothesis knowledge representation families $\mathcal{C}$ and $\mathcal{H}$ with respect to the probability distribution D on possible worlds is, given a set of *example* possible worlds of $\vec{x}^{(1)}, \ldots, \vec{x}^{(m)}$ each independently drawn from D for some m, parameters $\epsilon, \mu, \gamma > 0$, and a *query* $q \in \mathcal{Q}$ such that there exists some $c \in \mathcal{C}$ with $\Pr_{\vec{x} \in D}[c(\vec{x})] \geq \mu$ and $\Pr_{\vec{x} \in D}[q(\vec{x}) | c(\vec{x})] = 1$, return a hypothesis representation $h \in \mathcal{H}$ such that $\Pr_{\vec{x} \in D}[h(\vec{x})] \geq (1 - \gamma)\mu$ and $\Pr_{\vec{x} \in D}[q(\vec{x}) | h(\vec{x})] \geq 1 - \epsilon$. If $\mathcal{C} = \mathcal{H}$, this is the *proper* version of the task.

We note that in many cases, $\mathcal{C}$ and $\mathcal{H}$ may be set up to disallow trivial hypotheses. For example, if the query asserts that a single propositional attribute is true, the hypothesis that simply states that the same attribute is true would be an "optimal" solution to the abduction task, but it is not useful; we can avoid this problem by simply omitting the attribute from the vocabulary of the hypothesis representations.

We note that abduction is often formulated elsewhere in Bayesian terms: that is, in which there is a probability distribution over the possible hypotheses (called a "belief distribution") in addition to the possible worlds. In these formulations, the "probability of h" is then the probability that h is the "correct" hypothesis rather than the probability that h is satisfied in a possible world drawn from our distribution. Since more than one $h \in \mathcal{H}$ may be true of a given possible world, the two generally cannot be the same. The Bayesian framework is actually more general, since it allows the belief distribution on hypotheses to be specified arbitrarily, without regard to the distribution on the attributes describing the possible world. As usual, however, generality comes at a price: inference is only tractable for special families of distributions. At the same time, abduction – in which a hypothesis is generally *not strictly identified by the observation* – is a task in which this belief distribution plays a critical role. But, this generalization to arbitrary belief distributions discards any guidance as to which belief distribution should be chosen. The formulation we have given here, by contrast, fixes a natural alternative notion of what is a "likely" hypothesis.

Nevertheless, this formulation still has its own difficulties. As with precondition search, proper abduction for conjunctions is still intractable:

Theorem 6.13 *If there is a polynomial-time algorithm for proper abduction for conjunctive hypotheses where the queries either include all 2CNFs, DNFs, or may use an attribute that is not used in the hypothesis, then there are randomized polynomial-time algorithms for NP.*

Proof The proof is essentially identical to that of Theorem 6.10, except that we include an example $(1, \ldots, 1)$ which we give weight $1/2$ in the distribution, and specify $\mu = \frac{1}{2} + \frac{\lceil c|V| \rceil}{2m}$. Then any conjunction containing a negative literal has probability at most $1/2$ and has probability too low to satisfy the lower bound given by μ. □

But, just as with precondition search, abduction for disjunctions can be solved in polynomial time – indeed, by Elimination again, and the analysis is essentially identical.

Input: Query $q(\vec{x})$, examples $\vec{x}^{(1)} \ldots, \vec{x}^{(m)}$.
Output: A disjunction h.
begin
  Initialize h to contain all $2n$ possible literals on the n attributes.
  for $i = 1, \ldots, m$ **do** **if** $q(\vec{x}^{(i)}) = 0$ **then**
    forall $\ell \in h$ **do** **if** $\ell(\vec{x}^{(i)}) = 1$ **then** Remove ℓ from h.
  return h.
end

Algorithm 11: Elimination Algorithm for disjunctive abduction

Theorem 6.14 *There is a polynomial time algorithm for abduction for disjunctive hypotheses and any efficiently evaluatable query representations with probability* $1-\delta$ *given* $m = \Omega(\frac{1}{\mu\epsilon}(n + \log_2 \frac{1}{\delta}))$ *examples.*

Using the reduction of Theorem 4.14, we can also obtain a polynomial-time algorithm for abduction for k-DNF hypotheses given $\Omega(\frac{1}{\mu\epsilon}(n^k + \log_2 \frac{1}{\delta}))$ examples (for any fixed k).

As we saw previously (Section 4.6), sometimes an improper learning task is tractable even when the proper version is not. So it is next natural to wonder whether abduction is possible for conjunctive hypotheses by using a stronger representation. Unfortunately, the answer here is also believed to be negative: it is as hard as learning DNF representations. This problem is conjectured to be intractable.

Theorem 6.15 *If there is a polynomial-time algorithm for improper abduction for conjunctive hypotheses where the queries may use an attribute that is not used in the hypothesis, then DNF representations are efficiently PAC-learnable.*

Proof The proof relies on the "boosting" technique captured in Theorem A.19: it suffices to show that for any distribution on examples, abduction enables us to produce a representation that agrees with the DNF label (corresponding to the query) with probability at least $\frac{1}{2} + \frac{1}{\text{poly}(n)}$ (i.e., for some polynomial). So, suppose that there is a DNF $c(x)$ with k terms labeling examples drawn from a distribution D; and suppose that the labels are 1 with probability p. If $p < \frac{1}{2} - \frac{1}{2k}$ or $p > \frac{1}{2} + \frac{1}{2k}$, then the constant hypotheses $\bot$ and $\top$ are sufficient, and indeed we can simply check if the empirical fraction $\hat{p}$ of 1 labels lies in the interval $(\frac{1}{2} - \frac{1}{4k}, \frac{1}{2} + \frac{1}{4k})$, using the appropriate constant hypothesis if not, for a sample of size $O(k^2 \log \frac{1}{\delta})$ (so that if it is in the interval, $\hat{p} \geq (1 - 1/8k)p$ with probability $1 - \delta/2$ by a multiplicative Chernoff bound).

Otherwise, we know that $p \geq \frac{1}{2} - \frac{1}{2k}$, and by the pigeonhole principle, at least one of the k terms of $c(x)$, which we'll denote by $h^*(x)$, is satisfied with probability at least p/k. Note that when $c(x) = 0$, $h^*(x)$ is also 0, so $h^*(x)$ entails the label attribute given by $c(x)$. Suppose we solve the abduction task where the query is given by the label attribute, both error parameters equal to $1/8k$, and $\mu = \hat{p}/k$: note that $h^*(x)$ ensures that the algorithm produces a representation $h(x)$ such that with probability at least

$1-\delta/2$, $\Pr[h(x)] \geq (1-1/8k)\mu$ and $\Pr[c(x)|h(x)] \geq 1-1/8k$, i.e., $\Pr[\neg c(x)|h(x)] \leq 1/8k$. Now we can bound the agreement rate with $c(x)$:

$$\begin{aligned}
\Pr[c(x) = h(x)] &= \Pr[c(x) \wedge h(x)] + \Pr[\neg c(x) \wedge \neg h(x)] \\
&= \Pr[c(x)|h(x)]\Pr[h(x)] + (\Pr[\neg c(x)] - \Pr[\neg c(x)|h(x)]\Pr[h(x)]) \\
&\geq (1 - 2/8k)(1 - 1/8k)\frac{\hat{p}}{k} + (1 - p) \\
&\geq (1 - 3/8k)(1 - 1/8k)\frac{p}{k} + 1 - p \\
&\geq 1 - (1 - (1 - 1/2k)\frac{1}{k})p \\
&\geq \frac{1}{2} - \frac{1}{2k} + (1 - 1/2k)\frac{1}{2k} + (1 - 1/2k)\frac{1}{2k^2} \\
&= \frac{1}{2} + (1 - 1/k)\frac{1}{4k^2}
\end{aligned}$$

so returning $h(x)$ is sufficient with probability $1 - \delta$. □

Since most natural classes of representations beyond k-DNF can express conjunctions, it follows from Theorem 6.15 that abduction is intractable for these classes. This essentially settles the question of when our abduction task is tractable: only for hypotheses that can be expressed by k-DNF.

Chapter notes

The basic query answering task and Theorem 6.2 were introduced by Khardon and Roth (1997). But, prior to this work even, Kearns, Kautz, and Selman (1995) had observed that a style of reasoning algorithm based on evaluation at appropriately chosen assignments could be much more computationally efficient than reasoning at the level of formulas, in the style of Frege proof systems. The difference is that Kearns, Kautz, and Selman generate a set of assignments to capture a fixed knowledge base, whereas Khardon and Roth sample from an unknown data distribution. We note that Khardon and Roth pushed the approach much further than Theorem 6.2, obtaining low sample complexity and running time for much larger classes of formulas ($\log n$-width CNF queries and DNF knowledge bases).

Wittgenstein's discussion of the ill-definedness of natural concepts and in particular "games" appears in (Wittgenstein, 1953, Remarks 66–77). Natural concepts are discussed more generally by Quine (1960; 1969). The agnostic learning model was introduced by Kearns, Sellie, and Schapire (1994), who also show that (improper) agnostic learning of conjunctions would yield learning of DNF; recent evidence suggests that DNF may not be efficiently learnable (see chapter notes for Chapter 4 for a discussion). Feldman et al. (2009) showed that *proper* agnostic learning of conjunctions is NP-complete. Shalev-Shwartz et al. (2011) prove that halfspaces cannot be agnostically learned if the shortest vector problem is intractable (discussed in Section 4.8); this is Exercise A.10 in Appendix A. The state-of-the-art for approximate

agnostic learning of conjunctions/disjunctions is a $\tilde{O}(n^{1/3})$-approximation algorithm due to Awasthi, Blum, and Sheffet (2010). The simple algorithm of Exercise 6.4 is essentially a result from Valiant (1985).

"Common sense" was first posed as a goal for artificial intelligence by McCarthy (1959). Spelke and Kinzler (2007) review the notion of "core knowledge" in developmental psychology. Naïve physics was first discussed by Hayes (1979), and time is discussed in more detail by Allen (1983). The problem of spatial reasoning is reviewed by Cohn (1995). Naïve psychology is discussed by Clark (1987) and Watt (1995). The first discussion of the commonsense law of inertia, referred to as the "frame problem", as well as the first discussion of the qualification problem, appears in McCarthy and Hayes (1969); the term "qualification problem" was introduced later by McCarthy (1977). The "ramification problem" was first discussed by Finger (1987). The problem of "elaboration tolerance" was first posed by McCarthy (1988) (also see McCarthy (1998) for an elaboration). The CYC project is described by Lenat (1995). Alternatives to hand-coding, for example, by crowdsourcing were pursued in the OpenMind project by Stork (1999), and by mining the web were considered by Carlson et al. in the Never-Ending Lanugage Learning (NELL) project (Carlson et al., 2010). The relationship of this latter, "websense" knowledge to "common sense" is discussed by Michael (2013). For much more on the problems posed by common sense reasoning and existing approaches, see the book by Mueller (2014).

Early approaches to non-monotonic inference with declarative (formal logical) representations were independently proposed by Doyle and McDermott (1980), McCarthy (1980), and Reiter (1980), although the problem had been recognized earlier, and the term "non-monontonic" was first used by Minsky (1975). These approaches all suffered from serious issues, ranging from computational intractability to failure to handle simple scenarios in a reasonable way, for example the "Yale shooting problem" as discussed by Hanks and McDermott (1987) or the "Nixon Diamond" from Reiter and Criscuolo (1983). We'll review a more recent approach to such declarative representations that is at least computationally tractable (and arguably well behaved) in Chapter 10. The approach based on reasoning with subsets of examples here is based on a proposal by Roth (1995), and the examples here are selected to exemplify challenges for these earlier approaches. Following similar motivations, Grosof (1986) proposed a logic for explicitly reasoning about conditional probabilities, in contrast to the approach here where the actual probabilistic aspect lies in the semantics, not the language itself.

Indeed, reference classes featured prominently in Reichenbach's approach to "logical probability" (Reichenbach, 1949). The approach as later developed by Kyburg (1974) and Pollock (1990). Pollock, in particular, was interested in capturing non-monotonic reasoning with these conditional probabilities. Valiant's neuroidal model (Valiant, 1994, 1995, 2000a) relied on such "preconditions" to draw valid inferences, but did not discuss in detail how these preconditions could be obtained. The intractability of finding conjunctive preconditions is adapted from an analogous result in the "heuristic" learning model of Pitt and Valiant (1988) (the inapproximability was strengthened later by Håstad (1999)). The analysis of Elimination for finding disjunctive preconditions (Theorem 6.11) is essentially from Juba (2016b).

Anchors as conjunctive explanations for classification decisions were proposed by Ribeiro, Singh, and Guestrin (2018), and the formal connection to our precondition task formulation was observed by Juba and Li (2020). Juba and Li also gave algorithms that provide guarantees for the error tolerance in finding disjunctive (or k-DNF) explanations.

Abduction was first raised as a style of inference on par with induction and deduction by Peirce (1931). One of the first works to consider abduction in Artificial Intelligence, by Pople (1973), applied it to diagnosis. Later, Charniak and McDermott (1985) emphasized its central role in Artificial Intelligence. Abduction was also proposed to capture the interpretation of images (Cox and Pietrzykowski, 1986; Poole, 1990) and natural language (Hobbs et al., 1990), and many other tasks can be formulated as abduction. Abduction is also understood as a kind of "explanation", but the formulation we consider here is in general terms rather than for a specific case, as we considered in the precondition search task. For example, Qi et al. (2018) showed how characterizing the "anomalous" images taken by a given webcam can be addressed by this task.

Similar to how inductive learning has been modeled in many different ways, a variety of models have also been proposed for abduction. The primary difference in these models lies in which formulas are preferred over others. Early models used syntactic criteria, preferring explanations given by fewer literals, e.g., Reiter and de Kleer (1987); later approaches based on computing maximum-entropy distributions, such as Bacchus et al. (1996), amount to essentially the same preference. On the other hand, some Bayesian approaches, as discussed by Pearl (1988), Poole (1993), and Bylander et al. (1991), assumed possession of a "prior" distribution that captured such preferences; other approaches, like Hobbs et al. (1990), simply assumed that each individual literal had some known "cost", which corresponds to the log-likelihood in a product-distribution prior, though these costs could also be interpreted in other ways. The unknown-distribution learning formulation we present here is due to Juba (2016b). Durgin and Juba (2019) later noted that this problem is computationally equivalent to heuristic learning as proposed by Pitt and Valiant (1988), as well as "positive-reliable" learning as considered by Kalai, Kakade, and Mansour (2012). (The *sample complexity* of the abduction learning task is greater: see Juba and Le (2019) for an analysis.) Given this equivalence, Theorem 6.13 is essentially from Pitt and Valiant (1988) and Theorem 6.15 is essentially from Bshouty and Burroughs (2005). (Durgin and Juba give a simpler, direct proof of the consequence for refuting random constraint satisfaction problems, which can also be obtained from learning DNF.) Meanwhile, Theorem 6.14 is essentially equivalent to an earlier result by Kanade and Thaler (2014). (Bshouty and Burroughs also noted that Elimination solves this problem.) Zhang, Mathew, and Juba (2017) and Juba, Li, and Miller (2018) gave algorithms that can find imperfect disjunctions (k-DNFs) with guarantees. In particular, Exercise 6.6 is adapted from Juba, Li, and Miller (2018).

Exercises

6.1 *Conditional validity and nonmonotonicity from examples.* Consider the following database of complete examples describing package deliveries. Each row is a possible

world, and each attribute is Boolean.

	P	I	R	E	T
1	1	0	0	0	1
2	1	0	0	0	1
3	1	0	1	0	1
4	1	0	1	1	1
5	1	1	0	0	0
6	1	1	0	1	0
7	1	1	1	0	1
8	1	1	1	1	1
9	1	1	0	0	0
10	1	0	0	1	1
11	0	0	0	0	0
12	0	0	0	0	0

The attributes have the following meanings:

P = "the item is a package", I = "the package is international",
R = "the package has priority routing", E = "the package needs extra inspection",
T = "the package arrives on time."

Suppose that a query is regarded as "typically true" in a context when its empirical conditional probability in that context is at least $2/3$.
(a) By counting the fraction of the rows in the database satisfying P that also satisfy T, compute

$$\widehat{\Pr}[T \mid P].$$

Should we conclude from these examples that packages typically arrive on time?
(b) Next, similarly compute

$$\widehat{\Pr}[T \mid P \wedge I].$$

Should we conclude from these examples that international packages typically arrive on time?
(c) Finally, compute

$$\widehat{\Pr}[T \mid P \wedge I \wedge R].$$

Should we conclude from these examples that international packages sent with priority routing typically arrive on time?
(d) Explain how your answers to (a)–(c) illustrate the nonmonotonicity of conditional reasoning with examples. In particular, identify a conclusion that is supported in one context, retracted in a stronger context, and supported again in an even stronger context.
(e) Suppose the context $P \wedge I \wedge R \wedge E$ is used instead. How many examples satisfy this context? Compute

$$\widehat{\Pr}[T \mid P \wedge I \wedge R \wedge E].$$

What concern should we have about drawing a confident conclusion from such a small number of examples?

6.2 *Agnostic learning of decision stumps.* For this problem, we will consider attributes that take integer rather than Boolean values. A *decision stump* is a decision tree that only reads a single attribute: in the case of numerical values, thus, there is a single integer threshold τ and a fixed attribute X_i such that either the decision stump is true if and only if $X_i \geq \tau$ or if and only if $X_i < \tau$. Give a proper agnostic learning algorithm for decision stumps that runs in time polynomial in n, $1/\epsilon$, and $1/\delta$. (You may assume that integer arithmetic operations take constant time.) Prove that your algorithm meets these criteria.

6.3 *Proper agnostic learning of conjunctions is intractable.* Prove that if there is a proper agnostic PAC-learning algorithm for conjunctions running in time polynomial in n, $1/\epsilon$, and $1/\delta$, then every NP search problem has a randomized polynomial-time algorithm.

6.4 *Approximate agnostic learning of conjunctions.* In this problem, we will develop and analyze an algorithm for learning conjunctions in the agnostic learning model with some limited tolerance to adversarial noise. Specifically, we will develop a polynomial-time algorithm such that if there is some conjunction c^* such that for the target concept c, $\Pr_{x \in D}[c(x) = c^*(x)] \geq 1 - \epsilon$, then our algorithm produces a conjunction h such that $\Pr_{x \in D}[c(x) = h(x)] \geq 1 - O(n\epsilon)$ with confidence $1 - \delta$ where there are n attributes.
(a) Describe a family of counterexamples showing that the original Elimination Algorithm (Algorithm 1) *does not* provide such a guarantee: that is, prove that for any $C > 0$, there is a distribution D, property c, and conjunction c^* such that for some constants $\epsilon \in (0, 1/2)$ and $\delta \in (0, 1/2)$, $\Pr[c(x) = c^*(x)] \geq 1 - \epsilon$ and, with probability at least δ, the algorithm outputs h such that $\Pr[c(x) \neq h(x)] > C n\epsilon$, even if an arbitrarily large number of examples from D are provided.
(b) Consider the following "tolerant" version of the Elimination Algorithm: it only deletes a literal ℓ from h if $\ell(x) = 0$ and $c(x) = 1$ on more than νm out of the m examples. Use the multiplicative form of the Chernoff bound (see Exercise 2.2) to calculate a number of examples (as a function of ν and δ) that suffices to ensure that with probability $1 - \delta$, ℓ is deleted unless $\Pr_{x \in D}[\ell(x) = 0 \text{ and } c(x) = 1] \leq 2\nu$.
(c) Specify values of ν and m (as functions of n, ϵ, and δ) and prove that for these choices (when given m examples), the tolerant Elimination Algorithm produces a final hypothesis with error at most $O(n\epsilon)$ with confidence $1 - \delta$ for these parameters, given only that some c^* achieves an error rate of at most ϵ. (***Hint****: To choose μ, consider the following. Suppose that we now say that a literal is "bad" if* $\Pr_{x \in D}[\ell(x) = 0 \text{ and } c(x) = 1] > 2\nu$*. What is the error if all of the "bad" literals are removed? On the other hand, even the ideal conjunction c^* may have examples x for which some literal ℓ in c^* has $\ell(x) = 0$ and $c(x) = 1$. In particular, consider your solution to part (a)—can we afford to remove such literals, in general? When is it safe to eliminate a literal?*)

6.5 *Approximate exception-tolerant precondition search.* In Theorem 6.11, we saw that a variant of Elimination (Algorithm 10) solves credulous precondition search

for disjunctions in polynomial time. Continuing Exercise 6.4, describe a "tolerant" variant of Algorithm 10 that on input q and $\vec{x}^*$, with probability $1-\delta$ returns a disjunction h with $h(\vec{x}^*)$ true, $\Pr_{\vec{x}\in D}[h(\vec{x})] \geq (1-\gamma)\mu$ and such that for any disjunctive precondition c with $c(\vec{x}^*)$ true, $\Pr_{\vec{x}\in D}[c(\vec{x})] \geq \mu$ and $\Pr_{x\in D}[q(\vec{x})|c(\vec{x})] = 1-\epsilon(c)$, $\Pr_{\vec{x}\in D}[q(\vec{x})|h(\vec{x})] \geq 1 - O(n\epsilon(c)) - \gamma$. You should prove that your algorithm satisfies these criteria and runs in polynomial time in n, $\frac{1}{\mu}$, $\frac{1}{\gamma}$, and $\frac{1}{\delta}$. (***Hint**: suppose you have a solution to Exercise 6.4(c). How does the* conditional *probability of a counterexample* $\Pr_{\vec{x}\in D}[q(\vec{x}) = 0|h(\vec{x})]$ *relate to the overall counterexample probability* $\Pr_{\vec{x}\in D}[q(\vec{x}) = 0$ *and* $h(\vec{x}) = 1]$*?*)

6.6 *Covering algorithms for exception-tolerant k-DNF abduction.* In the previous exercises, we saw variants of the Elimination Algorithm that obtain $O(n)$ increase in the error rate where n is the number of attributes. Using the standard reduction of k-DNF to disjunctions (from Section 4.3), we could obtain such hypotheses, but at the cost of a $O(n^k)$ increase in the counterexample probability, which would often result in a vacuous output. In this exercise, we will develop an alternative style of algorithm that, when an unknown, ideal k-DNF explanation has only r terms, suffers only a $O(r \log 1/\gamma)$ increase in the probability of a counterexample when finding an explanation with plausibility $(1-\gamma)$-close to optimal. Actually, the method will produce a k-DNF hypothesis with only $O(r \log 1/\gamma)$ terms. We'll focus on the abduction task in this problem.

(a) Suppose that we have a set of m examples such that there is some r-term k-DNF c that is satisfied on $\mu \cdot m$ of the examples, and that we have a k-DNF h that is satisfied on at most $\alpha\mu \cdot m$ of the examples for some $\alpha < 1$. Prove that then there is a conjunction of size k in c that is not in h that is satisfied on at least $\frac{1-\alpha}{r}\mu \cdot m$ of the examples.

(b) Continuing from part (a), describe an algorithm that, when given a set of m examples such that a r-term k-DNF c is satisfied on $\mu \cdot m$ of the examples, returns a k-DNF h with $O(r \log \frac{1}{\gamma})$ terms such that h is satisfied on at least $(1-\gamma)\mu m$ examples, and runs in time polynomial in n^k, m, r, and $1/\gamma$. Prove that your algorithm meets these criteria.

(c) For given n, k, r, γ, and δ, and a fixed query representation q, calculate a bound m on the number of examples sufficient to guarantee that for all r-term k-DNFs c, the empirical fraction of examples on which c is true and the query q is false is within a multiplicative $(1 \pm \gamma)$ of $\Pr_{\vec{x}\in D}[q(\vec{x}) = 0$ and $c(\vec{x}) = 1]$ with probability $1-\delta$. (***Hint**: use the bound* $\binom{2n}{k} \leq \left(\frac{2ne}{k}\right)^k$ *on the number of possible terms on k literals – see Lemma A.10 for the calculation.)*

(d) Finally, describe an algorithm that, for a given n, k, r, γ, ϵ, δ, μ, and an efficiently evaluatable query representation q such that there exists a r-term k-DNF c^* with $\Pr_{\vec{x}\in D}[c^*(\vec{x})] \geq \mu$ and $\Pr_{\vec{x}\in D}[q(\vec{x})|c^*(\vec{x})] \geq 1-\epsilon$, runs in time polynomial in n^k, r, $1/\gamma$, $1/\epsilon$, $1/\delta$, $1/\mu$, and the length of q, and returns a $r \log 2/\gamma$-term k-DNF h that with probability $1-\delta$ satisfies $\Pr_{\vec{x}\in D}[h(\vec{x})] \geq (1-\gamma)\mu$ and

$$\Pr_{\vec{x}\in D}[q(\vec{x})|h(\vec{x})] \geq 1 - O(\epsilon r \log \frac{1}{\gamma}).$$

Prove that your algorithm meets these criteria. *(**Hint**: consider, which terms could be in c^*?)*