8 Learning to Reason with Partial Examples

The chapter discusses how the entailment problem introduced in Chapter 6 can be extended to the case of partial truth-assignments, as discussed in Chapter 7. Unlike in that initial setting, the methods and considerations of classical reasoning that we developed in Chapter 5 play a critical role here, to capture entailment under such incomplete information. We first illustrate how reasoning can fill in missing information with simple chaining in Section 8.1, and subsequently how multiple interleaved rounds of learning and reasoning can yield greater completeness in Section 8.2. Moreover, we will see that treating the learning problem as integrated with reasoning problems, where the ultimate end is the reasoning task, yields new possibilities. By focusing on the entailment task or an abduction task, we will see in Sections 8.3 and 8.4 how algorithms can learn implicitly from data while solving a reasoning task, that is, without producing an explicit representation of the learned knowledge. The benefit is this: It is intractable to explicitly learn many natural representations when they do not hold perfectly (Section 6.2), as needed for our common sense knowledge (recalling Sections 6.3 and 6.4). But, because implicit learning does not require finding explicit representations, it can nevertheless be computationally tractable.

8.1 The benefits of chaining.

In Chapter 7 we considered the problem of learning by observing attributes that we can directly sense from the environment, without requiring the presence of an explicit teacher to provide labels for learning examples. In that context, we accommodated for the partial observability of values for some of the attributes, both those that describe the features to be used for prediction and those that describe the features about which predictions are to be made. This uniform treatment of all attributes naturally leads to the consideration of attempting to predict the missing values across multiple (or even all the) attributes.

As a simple sanity check that the task at hand is not a priori impossible or trivial due to the substantial demand for structure in the possible worlds, consider a setting where through multiple partial observations, each hiding the value of at most one attribute, we have identified that it is consistent with all observations to hypothesize that the parity of all attributes is 0. Using this hypothesis we can then reliably impute each future

observation (over the exponentially many possible worlds that are consistent with our learned constraint) that has a single missing attribute.

With our example above using a single hypothesis to make predictions for all attributes notwithstanding, in the general case we would expect that we may have an individual hypothesis associated with each attribute to be predicted. We will denote this collection of hypotheses over worlds with n attributes as a vector $\vec{h}$ of size n. Given a partial observation $\rho \in \{0, 1, *\}$, each $h_i \in \vec{h}$ receives as input ρ_{-i} and makes a prediction on the value of ρ_i. We will write $\vec{h}(\rho)$ to mean the vector $\langle h_1(\rho_{-1}), \ldots, h_n(\rho_{-n})\rangle$ of the predictions of the n hypotheses. To accommodate for situations where we may wish to omit making a prediction on some attribute (e.g., because we do not have available a hypothesis for making that prediction), we will allow for some hypotheses in $\vec{h}$ to be replaced by the $\emptyset$ symbol, in which case the corresponding prediction in $\vec{h}(\rho)$ will be $*$, effectively indicating that $\vec{h}$ abstains from making that particular prediction.

As our aim here is to impute missing information in a partial observation ρ, it is important to define how ρ is updated to a partial observation ρ' following the computation of $\vec{h}(\rho) = \langle h_1(\rho_{-1}), \ldots, h_n(\rho_{-n})\rangle$. First, any attribute x_i for which $\rho_i = *$, inherits in ρ' the value the value $h_i(\rho_{-i}) \in \{0, 1, *\}$. Second, any attribute x_i for which $\rho_i \in \{0, 1\}$, retains in ρ' the value ρ_i. We will write $\rho \xrightarrow{\vec{h}} \rho'$ to denote the resulting partial observation.

Our definition above adopts the view that, in case of a conflict between the observed value and the predicted value of an attribute, we prefer to retain the observed value. This is a reasonable take, since we are assuming that observed values are always correct against the underlying ground-truth possible world. One could also have considered the alternative approach, where the predicted value takes precedence. Such an approach may be more appropriate in cases where the observations can be corrupted by noise, and where hypotheses aim not only to impute information, but also to revise noisily sensed information. (Although this is explored in the literature, we do not pursue such models here.) More important than which of the two approaches one adopts is the fact that our definition of applying hypotheses on a partial observation already goes beyond the learning-only perspective, and starts hinting at the role that reasoning has to play, with observed information being the base case of a chain of reasoning.

Indeed, it is now easy to extend our setting into one where not only are multiple attributes predicted for a given partial observation ρ, but the resulting partial observation ρ' becomes itself a new partial observation, on which predictions can be made. We will therefore consider a collection of vectors $\vec{h}^{(j)}$, each of size n, for $j = 1 \ldots d$, and we will call each such vector a *layer*. We will assume layers are applied sequentially, one after the other on a given partial observation, with the partial observation being updated after the application of each layer, as above. We will denote the collection of layers as a vector $\vec{H}$ of size d, that we will call a *policy* of *depth* d. In particular, $\vec{H}$ will be called *empty* if $d = 0$, *flat* if $d = 1$, and *chained* if $d \geq 2$. We will write $\vec{H}(\rho)$ to denote the partial observation such that $\rho = \rho^{(0)} \xrightarrow{\vec{h}^{(1)}} \rho^{(1)} \xrightarrow{\vec{h}^{(2)}} \rho^{(2)} \ldots \rho^{(d-1)} \xrightarrow{\vec{h}^{(d)}} \rho^{(d)} = \vec{H}(\rho)$.

Our treatment of how conflicts are resolved between layers is analogous to how conflicts are resolved between the first layer and the original partial observation: information observed or predicted in earlier layers is considered to prevail over information predicted in later layers. We repeat that this is just one possible way of resolving conflicts. In the general case, hypotheses in a policy can be chained to form arguments, whose priorities can be arbitrarily set, leading into policies whose reasoning semantics are considerably more nuanced. We will be discussing this further in Chapter 10. For this section, the specific reasoning semantics that we are considering suffices to establish our main result: that multi-layered policies are provably "better" than single-layered ones.

Before we can establish the result, we need to define the metric against which we will measure the performance of a policy. A first component of the performance metric, and aligned with the one adopted in Chapter 7 about individual hypotheses, is to require all imputed values in $\vec{H}(\rho)$ to be accurate against the possible world from which the partial observation ρ was drawn. Consider, therefore, a full example $\vec{x}$ over n attributes, a masking process M, a partial observation $\rho \in M(\vec{x})$, and a policy $\vec{H}$ of layers comprising n hypotheses each. We will say that $\vec{H}(\rho)$ is *sound* against $\vec{x}$ if $\vec{H}(\rho)$ is consistent with $\vec{x}$; thus, any variable whose value is missing in ρ either remains with a missing value in $\vec{H}(\rho)$, or the value is accurately recovered according to $\vec{x}$.

The soundness metric guarantees only that if we recover information, then the recovered information is appropriate. However, it does not account for whether we have actually recovered any information. We will say that $\vec{H}(\rho)$ is *complete* if $\vec{H}(\rho) \in \{0,1\}^n$; thus, every variable ends up with a value in $\vec{H}(\rho)$.

Example 8.1: Consider a story where it is not noon time yet, but Bob is hungry, and assume that we have are given the constraints that lunch takes place only at noon time, and that Bob has lunch only if he is hungry at lunch time. It follows that there is a single possible world $\vec{x}$ consistent with the story and the constraints, where it is not lunch time, and Bob is not having lunch. To formalize the type of reasoning that takes place in this story (and ignoring any temporal aspects, for simplicity), we consider possible worlds with four attributes $x_1 = NT, x_2 = LT, x_3 = BH, x_4 = BL$, corresponding to the properties "it is noon time", "it is lunch time", "Bob is hungry", "Bob is having lunch". The premise of the story corresponds, then, to a partial observation $\rho = 0 * 1*$, and our knowledge corresponds to the hypotheses $h_2 : LT \Rightarrow NT$ and $h_4 : BL \Rightarrow (BH \wedge LT)$, used respectively to predict the values of the second and fourth attribute. Our reasoning proceeds by first applying hypothesis h_2, and then hypothesis h_4. Thus, our reasoning policy $\vec{H}$ comprises two layers: $\vec{h}^{(1)} = \langle \emptyset, h_2, \emptyset, \emptyset \rangle$ and $\vec{h}^{(2)} = \langle \emptyset, \emptyset, \emptyset, h_4 \rangle$. It follows, then, that $\rho = 0 * 1* \xrightarrow{\vec{h}^{(1)}} 001* \xrightarrow{\vec{h}^{(2)}} 0010 = \vec{H}(\rho)$, which gives rise to a complete and sound conclusion $\vec{H}(\rho)$ against the expected possible world $\vec{x}$. Note, though, that had we considered a different reasoning policy with a single layer $\vec{h}^{(1)} = \langle \emptyset, h_2, \emptyset, h_4 \rangle$, we would have been led to an incomplete conclusion $001*$; indeed, hypothesis h_4 would be forced to abstain, as at the time of its application, the value of x_2 is missing.

The example above offers a first glimpse of the main result of this section: that chaining of hypotheses can be beneficial, if not in terms of increasing soundness, then in terms of increasing completeness. We will, in fact, establish a stronger result than what is implied by the example: that chaining in the context of a multi-layered policy $\vec{H}$ can help exceed the combined soundness and completeness of *any* single-layered policy, not just a flattening of $\vec{H}$. This strengthening of the result is important, since, in principle a single-layered policy could be using different hypotheses, whose learning might ensure the drawing of the same inferences as in the chain of inferences in a multi-layered policy. For instance, the single-layered policy that applies in parallel the hypotheses $h_2 : LT \Rightarrow NT$ and $h'_4 : BL \Rightarrow (BH \wedge NT)$ would yield the same conclusions as the multi-layered policy in the example. That is, the learnable nature of hypotheses might suggest that it can obviate the need for chaining, by learning reasoning shortcuts. We will show that this is not the case, since these shortcuts cause trouble on other partial observations.

Definition 8.1 (Chaining Collapsing) *Chaining collapses* on a set of partial observations $\{\rho^{(1)}, \ldots, \rho^{(k)}\}$ consistent with a corresponding set of possible worlds $\{\vec{x}^{(1)}, \ldots, \vec{x}^{(k)}\}$ if: for every policy $\vec{H}$ such that $\vec{H}(\rho^{(j)})$ is sound against $\vec{x}^{(j)}$ and complete for m values of $j = 1, \ldots, k$, there exists a flat policy $\vec{H}'$ such that $\vec{H}'(\rho^{(j)})$ is sound against $\vec{x}^{(j)}$ and complete for at least m values of $j = 1, \ldots, k$.

To establish the benefits of chaining hypotheses in multi-layered policies, it suffices, therefore, to establish that chaining does not collapse on a particular set of partial observations consistent with a corresponding set of possible worlds. To do so, we first introduce some preliminary definitions.

Definition 8.2 (Relevance and criticality) Consider a partial observation ρ that hides the value of attribute x_i, and a hypothesis h_t (meant to predict the value of attribute x_t). Consider changing the $*$ value of x_i in ρ to a $\{0, 1\}$ value. If at least one such value change causes $h_t(\rho_{-t})$ to change, then x_i is *relevant* for h_t on ρ. If both such value changes cause $h_t(\rho_{-t})$ to change (in two different ways), then x_i is *critical* for h_t on ρ.

Example 8.2: Continuing our example above, consider the partial observation $\rho = 0 * 1*$ and the hypothesis h_4 meant to predict the value of attribute x_4. Note that ρ hides the value of attribute x_2, and if we change the $*$ value of x_2 in ρ to a 0 value, then $h_4(\rho_{-4})$ changes from a $*$ to a 0 value; however, if we change the $*$ value of x_2 in ρ to a 1 value, then $h_4(\rho_{-4})$ retains its $*$ value. Thus, x_2 is relevant for h_4 on ρ, but not critical. Had h_4 been the equivalence $BL \Leftrightarrow (BH \wedge LT)$, then in addition to the effect of the former change on x_2, the latter change would have caused $h_4(\rho_{-4})$ to change from a $*$ to a 1 value, making x_2 critical for h_4 on ρ.

Lemma 8.3 (Unique Critical Attribute) *Consider a partial observation ρ, and a hypothesis h_t meant to predict the value of attribute x_t. If attribute x_i is critical for h_t on ρ, then no other attribute x_j is relevant for h_t on ρ.*

Proof Assume by way of contradiction that x_i is critical and x_j is relevant for h_t on ρ, and, therefore, $\rho_i = \rho_j = *$. Consider the partial observation $\rho[u, v]$ obtained from ρ after replacing ρ_i, ρ_j respectively with $u, v \in \{0, 1, *\}$. Since x_i is critical for h_t on ρ, then $h_t(\rho[0, *]_{-t}), h_t(\rho[1, *]_{-t})$ are distinct $\{0, 1\}$ values. Since x_j is relevant for h_t on ρ, then $h_t(\rho[*, a]_{-t}) \in \{0, 1\}$ for some $a \in \{0, 1\}$. Since h_t abstains monotonically, $h_t(\rho[*, a]_{-t}) = h_t(\rho[0, a]_{-t}) = h_t(\rho[0, *]_{-t}) \neq h_t(\rho[1, *]_{-t}) = h_t(\rho[1, a]_{-t}) = h_t(\rho[*, a]_{-t})$; a contradiction, as needed. □

The proof makes critical use of the property of formulas to abstain in a monotonic fashion; i.e., if they do not abstain on an observation, then completing some of the missing values in that observation will not cause the formula to start abstaining, and in fact, will not even cause a change in the prediction of the formula. Because of this property, no formula can have its predictions be unilaterally affected / determined by the values of more than one attribute. Although we defer to Chapter 10 for a fuller discussion of non-monotonic reasoning policies, non-monotonicity already underlies our next result, which effectively demonstrates that multi-layered policies can reason non-monotonically even with monotonic formulas, whereas this is not the case for flat policies.

Theorem 8.4 (Chaining Does Not Collapse) *There exist a set of partial observations consistent with a corresponding set of possible worlds on which chaining does not collapse.*

Proof Consider a set of possible worlds $\{\vec{x}^{(000)}, \vec{x}^{(000)}, \vec{x}^{(111)}, \vec{x}^{(111)}\}$ that assign the corresponding values in their superscript to the first three attributes and share the values in their remaining attributes. Consider the set of partial observations $\{\rho^{(0**)}, \rho^{(*0*)}, \rho^{(1**)}, \rho^{(*1*)}\}$ that assign the corresponding values in their superscript to the first three attributes, and do not hide the value of any of the remaining attributes. Observe that the set of partial observations is consistent with the corresponding set of possible worlds.

First consider a chained policy $\vec{H}$ that includes in its first layer two hypotheses h_1, h_2, with each predicting the value of the corresponding attribute x_1, x_2 from the value of the other attribute, and in its second layer a hypothesis h_3 that predicts the value of x_3 from the value of x_1. Then, $\vec{H}(\rho)$ is sound against the corresponding possible world $\vec{x}$ and complete for each of the four partial observations above. Assume, by way of contradiction, that chaining collapses, and therefore, there exists a flat policy $\vec{H}'$ such that $\vec{H}'(\rho)$ is sound against the corresponding possible world $\vec{x}$ and complete for each of the four partial observations above. Then, $\vec{H}'$ includes in its only layer a hypothesis h'_3 such that $h'_3(\rho^{(0**)}) = 0$, $h'_3(\rho^{(*0*)}) = 0$, $h'_3(\rho^{(1**)}) = 1$, $h'_3(\rho^{(*1*)}) = 1$.

Consider the partial observation $\rho^{(***)}$ that is obtained from either $\rho^{(0**)}$ or $\rho^{(1**)}$ after hiding the value of x_1. Since $h'_3(\rho^{(0**)}) = 0$ and $h'_3(\rho^{(1**)}) = 1$, it follows that $h'_3(\rho^{(***)}) = *$, and, therefore, that x_1 is critical for h'_3 on $\rho^{(***)}$. Noting that the partial observation $\rho^{(***)}$ can also be obtained from $\rho^{(*0*)}$ (or $\rho^{(*1*)}$) after hiding the value of x_2, and since $h'_3(\rho^{(*0*)}) = 0$, it follows that that x_2 is relevant for h'_3 on $\rho^{(***)}$. A contradiction by Lemma 8.3, as needed. □

8.2 Simultaneous learning and prediction.

We have thus established that it is beneficial to chain rules, in that in certain cases it can lead to performance that is beyond what could be achieved by flat policies. Rule chaining, after all, is at the heart of reasoning with knowledge bases. In cases where the rules are given (or even learned) and are individually guaranteed to be accurate (even if approximately) *when applied on the underlying possible worlds*, then these rules can be chained in an arbitrary order over partial observations, and their chaining can improve completeness without sacrificing soundness (at least significantly).

This is not the case, however, when the accuracy guarantees of rules are with respect to *their application on partial observations*, which is the case when the rules are learned in the context of this chapter. Recall that accuracy in this case allows rules to abstain without being penalized for the abstention. It is possible, then, that chaining can complete information, the lack of which was causing the abstention, leading rules to be forced to make predictions, which might be inaccurate. The next result demonstrates this situation, where the arbitrary chaining of rules can be detrimental to the soundness of the policy.

Theorem 8.5 (Loss of Soundness with Arbitrary Chaining) *There exist a partial observation $\rho \in \{0,1,*\}^n$ consistent with a possible world $\vec{x} \in \{0,1\}^n$, a policy $\vec{H}$ and a policy $\vec{H}'$ that is a flattening of $\vec{H}$, such that $\vec{H}(\rho)$ is not sound against $\vec{x}$, whereas $\vec{H}'(\rho)$ is sound against $\vec{x}$.*

Proof Order all possible worlds $\{0,1\}^n$ so that each pair of cyclically consecutive possible worlds differ in exactly on attribute (i.e., as a *Gray Code*), and consider a non-constant hypothesis h_t and two possible worlds $\vec{x}^{(0)}, \vec{x}^{(1)}$ that are cyclically consecutive in the Gray Code such that $h_t(\vec{x}^{(0)}_{-t}) = 0$ and $h_t(\vec{x}^{(1)}_{-t}) = 1$. Let $x_i \neq x_t$ be the single attribute that changes value between $\vec{x}^{(0)}, \vec{x}^{(1)}$, and without loss of generality, let $\vec{x}^{(0)}_t = \vec{x}^{(1)}_t = 1$. Hence, $h_t(\vec{x}^{(0)}_{-t}) \neq \vec{x}^{(0)}_t$. Let $\vec{x} = \vec{x}^{(0)}$, and construct the partial observation ρ to hide the values of x_i and x_t in $\vec{x}$.

Choose any hypothesis h_i such that $h_i(\rho_{-i}) = \vec{x}_i$, and note also that $h_t(\rho_{-t}) = *$. Observe then, that the policy $\vec{H}'$ that applies h_i, h_t in parallel is such that $\vec{H}'(\rho)$ is sound against $\vec{x}$. However, the policy $\vec{H}$ that includes h_i in its first layer $\vec{h}^{(1)}$, and h_t in its second layer $\vec{h}^{(2)}$ is such that $\rho = \rho^{(0)} \xrightarrow{\vec{h}^{(1)}} \rho^{(1)} \xrightarrow{\vec{h}^{(2)}} \rho^{(2)} = \vec{H}(\rho)$, where $\rho^{(1)}_i = h_i(\rho_{-i}) = \vec{x}_i$, and $\rho^{(2)}_t = h_t(\rho^{(1)}_{-t}) = h_t(\vec{x}_{-t}) \neq \vec{x}_t$. It follows that $\vec{H}(\rho)$ is not sound against $\vec{x}$, as needed. □

In trying to determine ways to chain rules that do not fall prey to the issue identified above, it is important to note that the post-hoc chaining of learned rules effectively corresponds to a shift between the distribution on which rules were trained – where they were not chained – and the distribution on which they are being evaluated, that is, where they are chained. Naturally then, to avoid this out-of-distribution evaluation, we need to ensure that the chaining that a rule faces during its evaluation corresponds to the chaining that the rule faces during its training. This naturally leads to an iterative process of learning and reasoning, where a policy is constructed one layer

at a time, with hypotheses in subsequent layers trained on the inferences of hypotheses in preceding layers. This *simultaneous learning and prediction* is formalized next.

Algorithm 13 interleaves learning and prediction to iteratively build a chained policy $\vec{H}$. It is given access to partial observations ρ drawn from a masking process M applied on possible worlds $\vec{x}$ drawn from some underlying probability distribution D, a non-negative integer d indicating the desired depth of $\vec{H}$, and value $\delta, \epsilon \in (0, 1]$ such that with probability $1 - \delta$,

$$\Pr_{\vec{x} \in D, \rho \in M(\vec{x})}[\vec{H}(\rho) \text{ is sound against } \vec{x}] \geq 1 - \epsilon.$$

Input: Set of partial observations $\{\rho^{(1)}, \ldots, \rho^{(m)}\}$, depth requirement d, error requirements δ, ϵ, complexity parameter κ, upper bound on concealment $1 - \eta$.
Output: A policy $\vec{H}$ of depth d that meets the requirements above.
Let $S_0 = \{\rho^{(1)}, \ldots, \rho^{(m)}\}$, $m_0 \leftarrow m$.
for $k = 1, \ldots, d$ **do**
  `// Learn on the current partial observations.`
  for $i = 1, \ldots, n$ **do**
    Execute algorithm L on S_{k-1} to obtain a hypothesis h_i such that with probability $1 - \delta'$, $h_i(\rho_{-i}^{(j)})$ does not have an accuracy conflict with $\vec{x}^{(j)}$ for any $j = 1, \ldots, m_{k-1}$.
  end
  Let $\vec{h}^{(k)} \leftarrow \langle h_1, \ldots, h_n \rangle$.
  `// Predict to enhance the partial observations.`
  Let $m_k \leftarrow m_{k-1} \cdot \frac{\eta \delta'}{\kappa + \log_2 \frac{1}{\delta'}}$, $S_k \leftarrow \{\rho'^{(j)} \mid \rho^{(j)} \xrightarrow{\vec{h}^{(k)}} \rho'^{(j)}, j = 1, \ldots, m_k\}$.
end
Set $\vec{H}$ so that its k-th layer is $\vec{h}^{(k)}$, for every $k = 1, \ldots, d$.
return $\vec{H}$.

Algorithm 13: Simultaneous Learning and Prediction

The operation of Algorithm 13 critically depends on the availability of a base algorithm L that can return hypotheses with the stated guarantees. Note that the design of and the types of guarantees that can be achieved by such base algorithms were the focus of Chapter 7. What remains then is to define the parameters m and δ' of the base algorithm L, so that Algorithm 13 remains efficient. These choices are provided in the proof of the next result.

Theorem 8.6 (Soundness in SLAP Setting) *Suppose that for a constant $C \geq 1$ and complexity function $\kappa : \mathbb{N} \to \mathbb{N}$, when given*

$$\frac{C}{\eta \epsilon_k}(\kappa(n) + \log \frac{1}{\delta'})$$

examples drawn from $M(D)$ for an at most $1-\eta$-concealing masking process M on n attributes, the base algorithm L returns a predictor that is $1-\epsilon_k$-accurate with probability $1-\delta'$ for any $\epsilon_k, \delta' > 0$. For any constant d,

$$m = O\left(\frac{n^d}{\eta^d \epsilon}\left(\frac{d}{\delta}\left(\kappa(n) + \log\frac{nd}{\delta}\right)\right)^{d-1}\right)$$

partial examples suffice to ensure that with probability $1-\delta$, if Algorithm 13 is invoked with depth d we obtain $\vec{H}$ such that

$$\Pr_{\vec{x}\in D, \rho\in M(\vec{x})}[\vec{H}(\rho) \text{ is sound against } \vec{x}] \geq 1-\epsilon.$$

Furthermore, if κ is a polynomial and L runs in polynomial time, Algorithm 13 runs in time $poly(n, 1/\epsilon, 1/\delta, 1/\eta)$.

Proof Put $\delta' = \frac{\delta}{2Cdn}$, $\epsilon_d = \frac{\epsilon}{n(1+\delta)}$, and $m_d = \frac{C}{\eta\epsilon_d}(\kappa(n) + \log\frac{1}{\delta'})$; now, for $k = 0, \ldots, d-1$, we have $m_{k-1} = \frac{m_k}{\eta\delta'}(\kappa(n) + \log_2\frac{1}{\delta'})$. So, if we put $\epsilon_k = \frac{\delta}{2dn\cdot m_k}$, we find by induction on k that the sample complexity guarantee for L ensures that with probability $1-\delta'$, each of the n predictors in $\vec{h}^{(k)}$ is $\frac{\delta}{2dn\cdot m_k}$-accurate and hence by a union bound, the total probability that S_k has any inaccurate predictions is at most $n\delta' + n\cdot m_k\cdot\frac{\delta}{2dn\cdot m_k} \leq \delta/d$. Therefore, there exists a $1-\eta_k$-concealing masking process M_k such that with probability $1-\delta/d$, S_k is distributed as m_k independent draws from $M_k(D)$, where $\eta_k \geq \eta$ since M was at most $1-\eta$-concealing and M_k masks a subset of the attributes masked by M; the base case is simply $M_0 = M$ with $\eta_0 \geq \eta$ by hypothesis. By a union bound over the d layers, we find that overall with probability $1-\delta$ every predictor in every kth layer is $1-\epsilon_k$-accurate; the first $d-1$ layers are $1-\frac{\delta\epsilon}{d(1+\delta)}$-accurate on the respective $M_k(D)$, and layer d is $1-\frac{\epsilon}{1+\delta}$-accurate. Hence, for future draws from $M(D)$, a union bound over the d layers gives that $\vec{H}$ is indeed $1-\epsilon$-accurate overall. Unrolling the recurrence, we now observe that $m_0 = \frac{Cn(1+\delta)}{\eta^d\delta'^{d-1}\epsilon}(\kappa(n)+\log\frac{1}{\delta'})^{d-1}$ indeed suffices, giving the claimed asymptotic sample complexity.

Given that $\kappa(n)$ is polynomial in n, for fixed d we observe that every m_k is polynomial in n, $1/\epsilon$, $1/\delta$, and $1/\eta$. Since L runs in polynomial time, where Algorithm 13 runs L n times on m_k examples for $k = 1, \ldots, d$, we indeed find that the running time of Algorithm 13 is $poly(n, 1/\epsilon, 1/\delta, 1/\eta)$ for fixed d, as claimed. □

Obviously, Algorithm 13 can only increase the completeness compared to learning a flat policy. Thus, interleaving learning and prediction is a solid approach when seeking to improve completeness without sacrificing soundness, assuming the availability of extra training time. As long as the depth d is constant, the extra time is polynomially bounded. In the general case, however, the training time grows exponentially with d (in an effort to contain the compounding error that would otherwise occur), which is clearly infeasible.

8.3 Implicit learning of testable knowledge.

So far we have seen that it may be beneficial to use both learning and reasoning when faced with partial observations. We now consider how learning can assist with common sense reasoning. Indeed, a significant portion of common sense consists of everyday knowledge that intuitively ought to be learnable from observations of the world around us. If we try to use our existing methods for learning from partial observations to acquire these rules, as in the previous sections, we hit three of the major obstacles of common sense reasoning:

1. **Qualification.** We cannot formulate a rule for our common sense inferences that has perfect validity. Hence, the rule learning task is an *agnostic* learning problem, which as we saw in Section 6.2, appears to be intractable for all but the very simplest rule representations.
2. **Scale.** The size of a common sense knowledge base is enormous, and it is not clear that it is feasible to exhaustively enumerate all of the common sense rules even in limited domains.
3. **Ramification.** Continuing on the problem of scale, the large number of common sense rules leads to a large number of inferences that may be drawn, and it does not seem feasible to enumerate all of them.

In the full information setting, we could overcome these issues by directly learning to reason in the form of validity estimation (Theorem 6.6): in this approach, we sought to directly estimate the validity of a given query representation from examples, without directly formulating an explicit knowledge base. This approach cannot be used when we only have partial examples, as we might not be able to evaluate a query given the partial information.

A natural idea at this point is to use partial evaluation (recall Definition 5.10) to obtain a conservative estimate of the validity: rather than the frequency with which the query is actually true, we merely estimate the frequency with which the query is *witnessed* true (Definition 7.16). We'll refer to this quantity as the *testability* of the representation:

Definition 8.7 We say that a De Morgan formula φ is *p-testable* with respect to a distribution on partial examples $M(D)$ if $\Pr_{\rho \in M(D)}[\varphi|_\rho = \top] \geq p$. A De Morgan knowledge base $KB = \{\psi_1, \ldots, \psi_k\}$ is p-testable if its conjunction $\varphi = \psi_1 \wedge \cdots \wedge \psi_k$ is p-testable.

It is straightforward to correspondingly adapt Theorem 6.6 to obtain an estimate of a query's *testability* as opposed to its *validity*; noting that the validity can only be higher, this indeed directly gives us a conservative estimate of the validity:

Proposition 8.8 *Let M be a masking process and D be any distribution on possible worlds. Then for any De Morgan formula φ, if φ is p-testable, φ is also p-valid.*

Proof Consider the event that we draw a mask m and possible world $\vec{x}$ such that for $\rho = m(\vec{x})$, $\varphi|_\rho = \top$. If φ is p-testable, this event has probability at least p. Observe that for

the assignment $\vec{x}'$ to the free attributes of $\varphi|_\rho$ that is consistent with the corresponding attributes of $\vec{x}$, $(\varphi|_\rho)(\vec{x}') = \top(\vec{x}') = 1$, where $\varphi(\vec{x}) = (\varphi|_\rho)(\vec{x}')$. Hence, in all such possible worlds, φ is true, and so the probability of obtaining $\vec{x}$ such that φ is true (i.e., the validity of φ) is at least p. □

There is no general relationship in the other direction, unfortunately:

Example 8.3: If D is a distribution such that the query φ is true on every possible world that occurs in D, φ is nontrivial, and M hides all of the attributes, then φ will be perfectly valid but 0-testable. The gap here is statistical.
Similarly, observe that in general for a 3DNF tautology φ and an M that hides all of the attributes, φ is perfectly valid and 0-testable, and although φ is perfectly valid independent of D (it is a tautology, after all) and hence examples are unnecessary, it is still NP-hard and hence intractable to determine that such queries are perfectly valid. In this case, the gap is computational.

In summary, what we will obtain from this approach is a near-certain lower bound on the validity of the query, the quality of which – that is, how much less than the validity – depends on how benign the masking process is.

Before giving a complete analysis of this approach, we will also consider how we can make use of additional prior knowledge. Unlike in the complete-information setting, a knowledge base may capture relationships that cannot be observed in the data. For example, we can draw inferences on attributes that are always hidden by the masking process. Indeed, in this extreme scenario, the inference problem is precisely deductive reasoning, and so we are seeking an approach that generalizes deductive reasoning. We could, in the style of the previous sections, invoke a purely deductive method (such as chaining) in which an explicit knowledge base is augmented by query access into the set of testable representations. But, it turns out that we can obtain greater efficiency through a deeper integration as follows.

Ultimately, we would like to know if there is some collection of testable formulas that, together with the given knowledge base, allows us to deduce that the query is (approximately) valid. The challenge is that this testable knowledge base is, in general, vast. So far we have avoided producing an explicit representation of it, partially mitigating the challenge posed by its size, but we still face the ramification problem: at any point in reasoning, via a method such as resolution using conflict-directed clause learning (as developed in Section 5.3), there may be many possible clausal consequences, which we cannot afford to explicitly represent.

Recall that reasoning with examples finessed the ramification problem by observing that within a given example, all of the possible conclusions are immediately observed to be valid. Indeed, the example itself is a compact summary of all of the formulas it satisfies, and the same holds more generally for partial examples. More concretely, on a given partial example, we only need to reason about the possible values taken by the missing attributes. This turns out to be equivalent to reasoning with the entire, vast knowledge base of formulas satisfied by that partial example for proof systems that are

closed under partial evaluation, as our various tractable families of resolution proofs are (recall Exercise 5.6). To be fully explicit now, the (simple) algorithm generalizing the method of Theorem 6.6 is given in Algorithm 14.

Input: Query $q(\vec{x})$, knowledge base $h_1(\vec{x}), \ldots, h_k(\vec{x})$, partial examples $\rho^{(1)} \ldots, \rho^{(m)}$.
Output: $\hat{p} \in [0, 1]$.
begin
  Initialize $count = 0$.
  for $i = 1, \ldots, m$ **do if** $q|_{\rho^{(i)}}$ *is provable from* $h_1|_{\rho^{(i)}}, \ldots, h_k|_{\rho^{(i)}}$ **then**
    Increment $count$.
  return $count/m$.
end

Algorithm 14: Reducing validity estimation with implicit learning under partial information to deductive reasoning.

Example 8.4: Before analyzing Algorithm 14, let's see an example of it in action. Suppose we have a service to support bird owners ("`birds.com`"), and among our services, we host feeding logs. These logs represent events in which a bird was fed some meal, and what was fed: seeds, grubs, fish, etc., for example including the rows in Table 8.1. We would like to make inferences about the birds that are kept on various accounts, in order to help connect owners to other services that may interest them, but that data is not being recorded. For example, we may wish to know if a given owner has flightless birds or not. Even though neither this nor the actual species of birds present is represented directly in the data, we may know for example that except for penguins, all of the birds the owners may have can fly. (So there are no ostriches, or emus, or...) We also know that the penguins only eat fish, and that when a bird is being fed one type of food, it is not being fed another type. We can represent this knowledge with rules such as

$$\neg penguin \vee eat_fish$$
$$penguin \vee flies$$
$$\neg eat_seed \vee \neg eat_fish$$
$$\neg eat_seed \vee \neg eat_grub$$
$$\vdots$$

This knowledge alone, however, does not directly indicate whether or not $flies$ holds in general; for any particular user, it may or may not hold.
What we might do is select the portions of the log for a fixed user n, and suppose that the feeding events are i.i.d., where there are some latent, unobserved attributes $flies$, $penguin$, and so on; so that we have a masking process that only reports the value of the propositional variable corresponding to the specific food that the bird ate.

User	Event	Food
9	42	*eat_seed*
9	43	*eat_worm*
9	44	*eat_grub*

Table 8.1 Example rows from `birds.com` feeding log.

Suppose we run Algorithm 14 on this data using unit propagation as our reasoning method: if indeed a bird was fed some food other than fish, say grubs, then for the partial example $\rho^{(i)} = eat_grub \leftarrow 1$, in our knowledge base the rule $\neg eat_grub \vee \neg eat_fish$ partially evaluates to $\neg eat_fish$ under $\rho^{(i)}$. In such a case, we have a unit propagation proof of $flies$: we first infer $eat_fish \leftarrow 0$, so from the rule $\neg penguin \vee eat_fish$, $penguin \leftarrow 0$, so from the rule $penguin \vee flies$ we get $flies \leftarrow 1$. (There is also a treelike resolution proof of $flies$ using low space, etc.) So indeed, each time the bird is recorded as eating any other type of food, the algorithm verifies that $flies$ holds, and thus reports for us a lower bound on the fraction of times a flying bird was fed.

Theorem 8.9 *Let Π be a set of proofs of a sound proof system that is closed under restrictions, for which there is an algorithm running in time $T(n, |q|, |KB|)$ where n is the number of attributes, $|q|$ denotes the size of the query representation, and $|KB|$ denotes the size of the knowledge base representation (all in bits). Then for any $\delta, \gamma \in (0, 1)$, knowledge base KB and query q for the proof system, distribution D on possible worlds, and masking process M, given $m \geq \frac{1}{2\gamma^2} \ln \frac{2}{\delta}$ partial examples from $M(D)$, Algorithm 14 runs in time $O(mT(n, |q|, |KB|))$ and returns a value $\hat{p}$ such that with probability $1 - \delta$,*

- ***Soundness:*** *for $p = \Pr_{\vec{x} \in D}[KB \Rightarrow q]$, $\hat{p} \leq p + \gamma$*
- ***Completeness:*** *if there is a KB I and $p' \in [0, 1]$ such that*
  1. *There is a proof of q from $I \cup KB$ in Π and*
  2. *I is p'-testable in $M(D)$*

  then $\hat{p} \geq p' - \gamma$.

Before proving Theorem 8.9, we'll elaborate on its interpretation. We interpret I as the relevant portion of an "implicitly learned" knowledge base of (simultaneously) testable rules. Observe that Algorithm 14 does not explicitly construct or consider I – indeed, it only explicitly considers the given q and KB – and the running time is entirely independent of the size of I. Observe that since I is only required to be p'-testable, where p' may be significantly less than 1, the algorithm can use an implicit KB that suffers from occasional counterexamples, as our "common sense rules of thumb" describing natural kinds might have. So, if there are some appropriate testable rules in D that we could have discovered that entail the query, the algorithm is guaranteed to certify the same bound we would have obtained from discovering those rules. Because the algorithm's running time is independent of the size of the implicit knowledge

base, it addresses the challenges posed by the scale of the implicit knowledge and ramifications in particular circumstances. Once again, as in reasoning with examples, all of these ramifications are captured by the partial examples themselves.

Example 8.5: Continuing Example 8.4, we can identify a p-testable rule that supports the conclusion $flies$ begin p-valid, namely the clause enumerating all of the possible kinds of food that the birds may be eating *except fish*:

$$eat_seed \vee eat_grub \vee eat_mouse \vee \cdots$$

Each time we observe the feeding of a type of food that is not fish – again, e.g., $\rho^{(i)} = eat_grub \leftarrow 1$ – this clause evaluates to $\top$. In this particular case, this clause suffices for a resolution proof of $flies$, where we eliminate each of the types of food using the respective clauses $\neg eat_seed \vee \neg eat_fish$, $\neg eat_grub \vee \neg eat_fish$, $\neg eat_mouse \vee \neg eat_fish$, …, ultimately yielding the clause $\neg eat_fish$, and then proceeding as before (in Example 8.4).

Proof of Theorem 8.9 **Soundness:** First observe that following Proposition 8.8, if there is a proof of $q|_{\rho^{(i)}}$ from $KB|_{\rho^{(i)}} = \{h_1|_{\rho^{(i)}}, \ldots, h_k|_{\rho^{(i)}}\}$, then by the soundness of the proof system, $KB \Rightarrow q$ must be true on the world $x^{(i)}$ drawn from D for $\rho^{(i)}$. Let V_i be the indicator random variable that is 1 iff $KB \Rightarrow q$ is true on the world $x^{(i)}$; we have just shown that $count \leq \sum_{i=1}^{m} V_i$. If $KB \Rightarrow q$ is p valid, $\mathbb{E}[V_i] = p$ for all i, and observe that $V_1, \ldots, V_m$ are mutually independent. Therefore, Hoeffding's inequality tells us that for $m \geq \frac{1}{2\gamma^2} \ln \frac{2}{\delta}$, with probability at least $1 - \delta/2$, $\frac{1}{m}\sum_{i=1}^{m} V_i \leq p + \gamma$. Hence, $\hat{p} = count/m \leq p + \gamma$ with probability at least $1 - \delta/2$ as well.
Completeness: Let T_i be the indicator random variable that is 1 iff $I|_{\rho^{(i)}} = \top$. Consider the proof of q from $I \cup KB$ in Π. Since Π is closed under restrictions, there is also a proof of $q|_{\rho^{(i)}}$ from $I|_{\rho^{(i)}} \cup KB|_{\rho^{(i)}}$. But observe, when $T_i = 1$, $I|_{\rho^{(i)}}$ simplifies to $\top$, and hence there is a proof of $q|_{\rho^{(i)}}$ from $KB|_{\rho^{(i)}}$ in Π. Therefore, $count \geq \sum_{i=1}^{m} T_i$. Since $\rho^{(1)}, \ldots, \rho^{(m)}$ are mutually independent, so are $T_1, \ldots, T_m$, and if I is p'-testable, $\mathbb{E}[T_i] \geq p'$ for all i. Hoeffding's inequality similarly gives us that $\frac{1}{m}\sum_{i=1}^{m} T_i \geq p' - \gamma$ with probability at least $1-\delta/2$ for $m \geq \frac{1}{2\gamma^2} \ln \frac{2}{\delta}$. We therefore find that $\hat{p} = count/m \geq p' - \gamma$ with probability $1 - \delta/2$ as well.

By a union bound over the soundness and completeness conditions, we thus find that with probability $1 - \delta$ overall, $p' - \gamma \leq \hat{p} \leq p + \gamma$ simultaneously, as needed. □

Another benefit of this approach is that it does not rely on an unverifiable assumption on the masking process such as bounded concealment or revealingness: with high probability, the estimate is a lower bound on the query's validity. Bounded concealment is an assumption that we saw captures a wide variety of situations where learning from partial information is possible. It includes situations where the information-hiding is independent of the attribute values, but gains significant breadth by allowing for "benign" dependence as well. Nevertheless, there are situations in which attributes may be hidden in ways that are not so benign, recalling Examples 7.3 and 7.7. Reasoning

with implicit learning can give us conservative estimates in such cases that are robust to this complex or adversarial information hiding.

To recap, the implicit learning approach allows us to dodge the intractability of agnostic learning, and thereby utilize rules with imperfect validity, addressing the qualification problem. It also addresses the key challenges of the vast size of common sense knowledge and scale of ramifications. Because new knowledge can be incorporated by introducing additional examples to our sample or adding representations to our explicit knowledge base – which likewise does not depend on the size of that knowledge base, let alone the size of the implicit knowledge base – it is elaboration tolerant. The primary drawback of this approach is its opacity. We are given no indication of what rules we might be using to reach our conclusions, beyond being able to report some examples where the query holds, that presumably witness these rules (whatever they are). It's therefore challenging to monitor the quality of the reasoning process, or audit it for compliance with various requirements we might wish to impose.

8.4 Abduction with implicit learning.

Our formulation of abduction with complete observations in Section 6.6 fundamentally relied on conditional probability: we wanted to identify a likely hypothesis that made a given "query" property likely to hold. Conditional probability expresses the incorporation of evidence into probability, as it recalculates a probability using only those outcomes that are consistent with the evidence. When the observations are incomplete, as usual we have two broad families of options: we can consider the probability of the query with respect to

1. *"Skeptical"*: partial examples where the hypothesis known to be *true*
2. *"Credulous"*: partial examples where the hypothesis is *not* known to be *false*.

In general, if the partial examples rarely determine the truth value of a hypothesis then these two could be quite different, since the credulous criteria would include the many ambiguous partial examples. Recall that abduction is intended to identify a "cause" or "explanation" for the query; we want the query to be likely to hold when the hypothesis is true. A credulous interpretation of the hypothesis may include many examples where the hypothesis is not true, and therefore the query may not hold, even if the hypothesis strongly entails the query. That is, under a credulous interpretation, even a good hypothesis may seem to be weak. We remark that the terminology may be confusing here, as the "skeptical" criteria refers to whether or not we consider a partial example as sufficiently consistent to count against the hypothesis if the query condition fails to hold, so the "skeptical" interpretation actually is less demanding of the hypothesis. We are often interested in using the abduced hypothesis as a tentative guide to further reasoning or investigation, which we might rule out later. We therefore are often willing to tolerate abduction producing some incorrect hypotheses, so long as we can eventually arrive at good ones. We therefore *focus on skeptical interpretations*

of the hypothesis, considering only those partial examples where the hypothesis can be determined to be consistent.

Unfortunately, it is often challenging to determine whether or not the partial examples are consistent with the hypothesis. In particular, recall that we were only able to give an efficient algorithm for abduction with k-DNF hypotheses (Theorem 6.14), noting that other natural families of hypotheses, such as conjunctions, were intractable (Theorems 6.13 and 6.15). But, deciding whether or not a k-DNF is necessarily true given a partial example is equivalent to deciding whether the k-DNF we obtain by partially evaluating the original formula is a tautology. Since a DNF is falsified whenever its negation is satisfied by De Morgan's law, this problem is essentially equivalent to k-SAT, which we know is NP-hard (recall Theorem 3.29). Thus, when we only have partial observations available, our one example of a tractable family of representations for abduction no longer seems so tractable.

We faced a similar obstacle in the previous section. It motivated us only considering cases where a formula was witnessed true on a partial example, rather than whenever the formula is a tautology—again, with the key benefit that witnessed evaluation has an efficient algorithm. Witnessed evaluation alone does not make use of a knowledge base, however, so we can attempt to get slightly closer to our ideal of identifying all of the examples where the hypothesis is true by checking to see if it is provable from our knowledge base. Actually, even this is still not known to yield a tractable abduction task with k-DNF hypotheses. Although there is also no known formal barrier, we only know efficient algorithms that use the partial examples where the hypothesis contains a *provable term*, as opposed to those where the entire disjunction of terms is provable.

The question of whether or not the query holds also faces the same obstacle, though given the probability distribution, the problem here is again validity estimation, and so we will be able to use the same approach as developed in the previous section for this part. Again motivated by our desire for good hypotheses to remain solutions to the task (at the possible risk of some poor hypotheses remaining candidates), *we choose the more relaxed "credulous" criteria for the query*. That is, we only penalize a hypothesis for those partial examples where we can determine that the query *does not* hold. We thus arrive at the following task:

Definition 8.10 *Weak k-DNF abduction* with respect to a given proof system, the knowledge base KB, and probability distribution $M(D)$ on possible worlds is, given a set of *example* partially specified possible worlds $\rho^{(1)}, \ldots, \rho^{(m)}$ each independently drawn from $M(D)$ for some m, parameters $\epsilon, \mu, \gamma > 0$, and a *query* CNF q such that there exists some k-DNF c with

- *Plausibility:* $\Pr_{\rho \in M(D)}[\exists \text{ term } t \in c|_\rho : KB|_\rho \vdash t] \geq \mu$ and
- *Weak entailment:* $\Pr_{\rho \in M(D)}\left[KB|_\rho \nvdash \neg q|_\rho \;\middle|\; \exists \text{ term } t \in c|_\rho : KB|_\rho \vdash t\right] = 1,$

return a k-DNF *hypothesis* representation h such that

- $\Pr_{\rho \in M(D)}[\exists \text{ term } t \in h|_\rho : KB|_\rho \vdash t] \geq (1-\gamma)\mu$ and
- $\Pr_{\rho \in M(D)}\left[KB|_\rho \nvdash \neg q|_\rho \;\middle|\; \exists \text{ term } t \in h|_\rho : KB|_\rho \vdash t\right] \geq 1 - \epsilon.$

We can interpret the plausibility condition as an estimate of the validity of the hypotheses c and h that utilizes implicit learning, using Theorem 8.9:

Corollary 8.11 *Suppose that the proof system is sound and restriction-closed, and KB is valid on the support of D. Then*

$$\Pr_{\vec{w}\in D}[c(\vec{w})] \geq \Pr_{\rho\in M(D)}[\exists t \in c|_\rho : KB|_\rho \vdash t].$$

Furthermore, if for each term t of c there is an implicit knowledge base I_t such that $KB \wedge I_t \vdash t$ and $\Pr_{M(D)}[\exists t \in c : I_t$ is witnessed true$] \geq p$, then

$$\Pr_{\rho\in M(D)}[\exists t \in c|_\rho : KB|_\rho \vdash t] \geq p.$$

The same holds replacing c with h, for the analogous plausibility property of the output hypothesis.

Example 8.6: If c itself is p-testable (meaning, recall, it is witnessed true with probability p), this means that at least one of its terms is witnessed true with probability p. We can take each I_t to consist of just the respective term t, and this satisfies the definition, ensuring p-plausibility. (Actually, since the term simplifies to $\top$, the proof of it is trivial.)

Note that the definition is more liberal than just the p-testable c, and also allows some nontrivial reasoning with KB and a nontrivial I_t to infer t.

Likewise, for weak entailment, if q is actually entailed by c on the support of D, we know that c containing a provable term means it is true on the underlying possible world, and hence q is true, so by the soundness of the proof system, $KB|_\rho \nvdash \neg q|_\rho$. Thus,

Corollary 8.12 *Suppose that the proof system is sound and restriction-closed, and KB is valid on the support of D. Then*

$$\Pr_{\substack{\vec{x}\in D,\\ \vec{\rho}=M(\vec{x})}}\left[q(\vec{x}) \;\middle|\; \exists t \in c|_\rho : KB|_\rho \vdash t\right] \leq \Pr_{\rho\in M(D)}\left[KB|_\rho \nvdash \neg q|_\rho \;\middle|\; \exists t \in c|_\rho : KB|_\rho \vdash t\right].$$

Also, if there is an implicit knowledge base I such that $KB \wedge I \vdash \neg q$ and I is p-testable w.r.t. $M(D)$ conditioned on there existing a term $t \in c|_\rho$ such that $KB|_\rho \vdash t$, then

$$\Pr_{\rho\in M(D)}\left[KB|_\rho \nvdash \neg q|_\rho \;\middle|\; \exists t \in c|_\rho : KB|_\rho \vdash t\right] \leq 1 - p.$$

And similarly for h.

Our formulation of the weak abduction task allows Elimination for disjunctive abduction (Algorithm 11) to be easily adapted to a partial information setting, using tests for the provability of the appropriate queries in place of formula evaluation, as shown in Algorithm 15.

Theorem 8.13 *Consider any fragment of a proof system that proves DNF conclusions q for a CNF knowledge base KB and is automatable in time $T(n, |q|, |KB|)$ (for*

Input: Query CNF $q(\vec{x})$, set of CNF formulas KB, partial examples $\rho^{(1)} \ldots, \rho^{(m)}$, parameter k.
Output: A k-DNF h.
begin
  Initialize h to contain all terms of at most k literals on the n attributes.
  for $i = 1, \ldots, m$ **do** **if** $KB|_{\rho^{(i)}} \vdash \neg q|_{\rho^{(i)}}$ **then**
    forall *terms* $t \in h$ **do** **if** $KB|_{\rho^{(i)}} \vdash t|_{\rho^{(i)}}$ **then**
      Remove t from h.
  return h.
end

Algorithm 15: Elimination Algorithm for weak abduction from partial examples

$T(n, |q|, |KB|) \geq \Omega(n + |q| + |KB|)$ on n attributes. Given $m = \Omega(\frac{1}{\mu\epsilon}(n^k + \log \frac{1}{\delta})$ partial examples from $M(D)$, Algorithm 15 solves the weak abduction task for k-DNF queries with respect to the given fragment in time $O(m(T(n, |q|, |KB|) + n^k T(n, k \log n, |KB|))$ (with $\gamma = 0$) with probability $1 - \delta$.

The proof is very similar to Theorems 6.11 and 6.14:

Proof Let c be a k-DNF condition with plausibility μ satisfying the weak entailment condition of the weak abduction task for q. Observe that for any term t on at most k literals, if there exists ρ in the support of $M(D)$ such that $KB|_\rho \vdash t|_\rho$ and $KB|_\rho \vdash \neg q|_\rho$, then t cannot be contained in c, as otherwise we would have

$$\Pr_{\rho \in M(D)} \left[KB|_\rho \nvdash \neg q|_\rho \;\middle|\; \exists \text{ term } t \in c|_\rho : KB|_\rho \vdash t \right] < 1$$

violating the assumption that c weakly entails q. Hence, since the k-DNF h returned by Algorithm 15 contains all of the terms for which such ρ do not exist, in particular, h contains all of the terms of c. Thus, whenever there is a term $t \in c$ such that $KB|_\rho \vdash t|_\rho$ (with $t|_\rho$ corresponding to $t' \in c|_\rho$), $t \in h$ also and hence $t|_\rho \in h|_\rho$. Thus, h also satisfies plausibility with probability μ.

Next, consider any $\tilde{h}$ that does not weakly entail q. Then

$$\Pr_{\rho \in M(D)} \left[KB|_\rho \vdash \neg q|_\rho \;\middle|\; \exists \text{ term } t \in h|_\rho : KB|_\rho \vdash t \right] > \epsilon.$$

or, rewriting the conditional probability, we find that $KB|_\rho \vdash \neg q|_\rho$ and $\exists$ term $t \in h|_\rho : KB|_\rho \vdash t$ simultaneously with probability greater than $\epsilon \Pr[\exists \text{ term } t \in h|_\rho : KB|_\rho \vdash t]$; by the above calculation, this is at least $\epsilon \cdot \mu$. Hence, the probability that when we draw ρ from $M(D)$ that we do not have $KB|_\rho \vdash \neg q|_\rho$ and $\exists$ term $t \in h|_\rho : KB|_\rho \vdash t$ simultaneously is at most $\epsilon\mu$. Observe that if we do encounter such a ρ, a term from $\tilde{h}$ is deleted from h, and so Algorithm 15 cannot return $\tilde{h}$ as output. After iterating over

$$m \geq \frac{\ln 2}{\mu\epsilon} \left(\left(\frac{2en}{k} \right)^k + \log_2 \frac{1}{\delta} \right)$$

independent examples, the probability that we never encounter such a ρ is less than $(1-\epsilon\mu)^m \leq \frac{\delta}{2^{(2en/k)^k}}$. The number of terms on the $2n$ literals of size at most k is at most $\left(\frac{2en}{k}\right)^k$, and so there are at most $2^{(2en/k)^k}$ k-DNFs. By a union bound over these $\tilde{h}$ that do not weakly entail q, the total probability that Algorithm 15 could return any of them is at most δ. Thus, with probability at least $1-\delta$, the algorithm returns a k-DNF h that satisfies both plausibility μ and weakly entails q with probability $1-\epsilon$, as needed for weak abduction.

As for the running time, we note that for each of the m partial examples, we check provability of q once and each of the $O(n^k)$ terms at most once; partial evaluation of KB, q, and the terms is a linear time operation and hence dominated by the time to automate the fragment for each provability check. □

An undesirable property of Algorithm 15 is that it tends to create large k-DNFs with many terms when a much smaller one that only contained a few terms that are often provable may exist. A smaller k-DNF is not only easier to understand and more useful for further investigation; it in turn also leads to greater efficiency with respect to the data required. Exercise 6.6 developed a simple greedy algorithm in the full-information setting, and it may be extended to the partial information setting analogously to how Algorithm 11 is extended to Algorithm 15 (Exercise 8.6).

Example 8.7: When we introduced abduction in Section 6.6, we noted that abduction underlies our interpretation of stories. As we stressed in Example 7.6, stories (and language more generally) usually only present a partial description of a possible world, specifying a few key attributes and leaving the rest unknown or implicit. Consequently, the interpretation of language is intrinsically a problem of abduction from partial information. In turn, these inferences draw on the audience's common sense knowledge, which we have sought to capture at least in part by implicit learning, and so we indeed require a formulation of the abduction task that can utilize this implicit knowledge, as our formulation of weak abduction does (cf. Corollaries 8.11 and 8.12).

Suppose we are told, "Alice walked in to the hotel lobby dripping wet." Since it is relatively unusual to be wet in a hotel lobby, we might seek an explanation for this state of affairs. Our algorithms will find a k-DNF, which we can interpret as a list of possibilities that are described by a small number of propositional statements. Of course, what statements we consider depends on our encoding of scenes. What they share is that in all previously observed scenes where we can infer that these statements are true, we cannot conclude that the subject does not walk in to a hotel while wet; i.e., it was possible that they do. For example, one condition might say that "Alice was returning to the hotel in a rainstorm without an umbrella," another might say "Alice was returning from swimming in the hotel pool without bringing a towel," "Alice spilled a drink all over herself while returning to the hotel," or "a lawn sprinkler switched on as Alice returned to the hotel." What are ruled out are conditions that were true in some story in which we could conclude that either the subject was dry, or

where the subject did not next enter the hotel. For example, in conditions that omit forgetting the umbrella or towel, then we may have examples where it was observed that Alice was safely dry. Or, another example scene may express that Alice had an umbrella but there was no need for it because it was not raining and she remained dry, which rules out conditions that fail to mention the rain. Algorithm 15 would return a formula listing any of these conditions that can be expressed by the propositional vocabulary, whereas the greedy algorithm of Exercise 8.6 would only return a few of the most likely ones, that reach a desired degree of probability.

Chapter notes

Sections 8.1 and 8.2 were developed in a sequence of works by Michael (2009; 2014). While the algorithm and the main ideas are adapted directly from those works, the guarantee for SLAP that we present in Theorem 8.6 differs from the previous guarantee of Michael (2014) in two key respects: first, Theorem 8.6 has a more restrictive assumption that the sample complexity of the base algorithm has the optimal $1/\epsilon$ dependence on the accuracy parameter, as opposed to simply a polynomial dependence. Second, this strengthening of the assumption leads to a sample complexity and running time that has only a *single exponential* dependence on the depth d, where the previous works had a *doubly-exponential* dependence. Although clearly the depth must remain small in order for the method to be tractable, the single exponential yields reasonable bounds for slightly higher depth.

Section 8.3 is adapted from Juba (2013). The only change to the formulation is that rather than testing whether the query has a fixed validity level, the formulation of Theorem 8.9 here produces a quantitative estimate $\hat{p}$ of the validity. Section 8.4 is adapted from Juba, Li, and Miller (2018), but we present a simpler algorithm based on Elimination here rather than the covering algorithm from that work. The covering algorithm has a potentially much lower sample complexity and produces a potentially much smaller hypothesis if a small DNF exists: see Exercise 8.6.

Exercise 8.4 is from a preprint (Juba, 2012) of Juba (2013). Exercise 8.5 is from Juba (2016a), which applies implicit learning to an agnostic, offline learning problem for planning.

Exercises

8.1 *Priority of rules in earlier vs later layers in chaining.* In Section 8.1, applying a layer of hypotheses to a partial example uses what we may call the *earlier-priority convention*: values already observed in the partial example, or filled in by earlier layers, are retained even if a later hypothesis predicts a different value. The section also notes that one could instead consider a convention in which predicted values take precedence. Indeed, consider the following *later-priority convention*: When a hypothesis in a later layer predicts 0 or 1 for an attribute, that prediction overwrites the current value of the attribute. When the hypothesis predicts $*$, the current value is left unchanged.

(a) Reformulate the definition of applying a depth-d policy

$$\vec{H} = \langle \vec{h}^{(1)}, \ldots, \vec{h}^{(d)} \rangle$$

to a partial example ρ under the later-priority convention. Give the coordinate-wise update rule for each layer.
(b) Revisit Definition 8.1 and Theorem 8.4. Identify any places where the proof uses the priority convention. Would the statement and proof still go through under the later-priority convention? Justify your answer.
(c) Revisit Theorem 8.5. Identify any places where the proof uses the priority convention. Would the example showing loss of soundness with arbitrary chaining still work under the later-priority convention? Justify your answer.
(d) Revisit Algorithm 13 and Theorem 8.6. In the proof, the examples updated after each layer are treated as partial examples obtained from the same underlying complete examples, with some additional attributes filled in. Under the later-priority convention, later hypotheses may also overwrite already observed or previously inferred $0/1$ values. Which parts of the proof would need to be adjusted to account for such overwrites? Explain how the accuracy guarantees for the learned hypotheses could be used to control them.
(e) Based on your answers above, is the choice between earlier-priority and later-priority essential for the formal results above, or largely a matter of convention? Explain.

8.2 Does noise-tolerant learning help create longer chains?
TO DO

8.3 *Guaranteeing validity requires testability for clauses.* We know that determining if a DNF formula is perfectly valid is NP-complete, and so a DNF may be highly valid but not testable (Example 8.3). If we consider clauses, however, testability characterizes validity under partial information: We already know that if a clause is p-testable under a masking process M and distribution on possible worlds D, it must be p-valid for D (Proposition 8.8). Prove that if a clause c is *not* p-testable under partial examples from $M(D)$, then there is a probability distribution $\tilde{D}$ and masking process $\tilde{M}$ such that $\tilde{M}(\tilde{D})$ is distributed identically to $M(D)$, but such that c is not p-valid under $\tilde{D}$.

8.4 *Feasible witnessing is necessary for implicit learning.* One important property of our notion of witnessed evaluation is that it is tractable. In this exercise, we will show that any (possibly alternative) notion of witnessed values that supports tractable implicit learning (i.e., for which there is an analogue of Theorem 8.9) must be tractable to evaluate.
Let Π be a set of proofs of a sound proof system such that given a knowledge base KB, there is a proof in Π from KB of each of the individual formulas $h \in KB$. Consider any property W of pairs of representations used by the lines of Π and partial examples such that some algorithm takes as input KB, a query q, partial examples $\rho^{(1)}, \ldots \rho^{(m)}$ drawn from $M(D)$ for a masking process M and a distribution D on possible worlds,

with parameters ϵ, δ, and γ in $(0, 1)$, runs in time polynomial in the number of attributes n, the size of KB and q (in bits), $1/\gamma$, $1/\epsilon$, and $1/\delta$, and returns a value $\hat{p}$ such that with probability $1 - \delta$,
Soundness: $\hat{p} \leq \Pr_{\vec{x} \in D}[KB \Rightarrow q] + \gamma$
Completeness: for any KB I such that there is a proof of q from $I \cup KB$ in Π, $\Pr_{\rho \in M(D)}[\forall h \in I\ W(h, \rho)] - \gamma \leq \hat{p}$.
Prove that then there is an algorithm taking as input a representation h and partial example ρ that returns *accept* if $(h, \rho) \in W$ and returns *reject* if there is a possible world $\vec{x}$ consistent with ρ such that $h(\vec{x})$ is false. Prove that your algorithm runs in polynomial time in n and the size of h.

8.5 *Implicit learning with distribution shift using importance weighting.* A limitation of the basic learning model we consider is that its guarantees only address situations where we wish to make predictions on possible worlds drawn from exactly the same distribution as the training examples. In this exercise, we will develop a technique that can, in some situations, relax this requirement, allowing us to obtain guarantees that hold even when the distribution has "shifted" somewhat.
Suppose that each of our partial examples $\rho^{(1)}, \ldots, \rho^{(m)}$ drawn from $M(D)$ is given along with a weight $w^{(i)} = Q(x^{(i)})/D(x^{(i)})$ for some other probability distribution Q. Suppose further that for some $W > 0$, every $w^{(i)} \leq W$. Now, consider the following modification of Algorithm 14: instead of incrementing a running count of the number of times $q|_{\rho^{(i)}}$ has a refutation, we instead add $w^{(i)}$ to the count (and divide the final result by m as before) to obtain $\tilde{p}$. Prove that for this modified algorithm, given $\frac{W^2}{2\gamma^2} \ln \frac{2}{\delta}$ examples from $M(D)$, with probability $1 - \delta$,
Soundness: $\tilde{p} \leq \Pr_{x \in Q}[KB \Rightarrow q] + \gamma$ and
Completeness: if there is a p-testable KB I with respect to $M(Q)$ for which q has proof from $KB \cup I$ in Π, then $\tilde{p} \geq p - \gamma$.

8.6 *Covering algorithms for abduction using partial examples.* As mentioned in Section 8.4, the Elimination Algorithm tends to create large formulas, which are less useful and harder to understand as explanations. Analogous to Exercise 6.6, in this exercise you will develop an algorithm for abduction with partial examples that guarantees that the formula remains small if possible: if the ideal k-DNF has r terms, then we find a k-DNF with $\approx r \log 1/\gamma$ terms (with plausibility $(1 - \gamma)\mu$ compared to the ideal μ). This algorithm will, moreover, have potentially much lower sample complexity, scaling with $rk \log n$ instead of n^k.
(a) Suppose that we have a set of m partial examples $\rho^{(1)}, \ldots, \rho^{(m)}$ such that for some r-term k-DNF c, on $\mu \cdot m$ of the examples, under the example $\rho^{(i)}$ some term t of $c|_{\rho^{(i)}}$ is provable from $KB|_{\rho^{(i)}}$. Suppose furthermore that we have a k-DNF h for which on at most $\alpha\mu \cdot m$ of the examples for some $\alpha < 1$ is there a term t of h such that $t|_{\rho^{(i)}}$ is provable from $KB|_{\rho^{(i)}}$. Prove that then there is a conjunction of size k in c that is not in h that is provable under $\rho^{(i)}$ from $KB|_{\rho^{(i)}}$ on at least $\frac{1-\alpha}{r}\mu \cdot m$ of the $\rho^{(i)}$.
(b) Now, prove that every r' term k-DNF on n attributes h such that in at least $m \geq \Omega(\frac{1}{\mu\epsilon}(r'k \log n + \log \frac{1}{\delta}))$ partial examples, whenever a term t of the DNF has $t|_\rho$

provable from $KB|_\rho$, $\neg q|_\rho$ is not provable from $KB|_\rho$, we have that with probability at least $1-\delta$ over the partial examples,

$$\Pr_{\rho \in M(D)}[KB|_\rho \vdash \neg q|_\rho \wedge \exists \text{term } t \in h|_\rho : KB|_\rho \vdash t] \leq \mu\epsilon.$$

(***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.)*

(c) Finally, prove that there is a polynomial time algorithm (for fixed k) such that when there is a r-term k-DNF satisfying the plausibility property with probability μ and weak entailment, using

$$m = O\left(\frac{1}{\mu}\left(\frac{1}{\epsilon} + \frac{1}{\gamma^2}\right)\left(rk \log n \log \frac{1}{\gamma} + \log \frac{1}{\delta}\right)\right)$$

partial examples, returns a $r \log \frac{2}{\gamma}$ term k-DNF solving the weak k-DNF abduction task. The algorithm should run in time polynomial in n^k, $\frac{1}{\mu}$, $\frac{1}{\epsilon}$, $\frac{1}{\gamma}$, and $\log \frac{1}{\delta}$.