Learning Structured Decision Problems with Unawareness

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of the 36th International Conference on Machine Learning (ICML)

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Learning Structured Decision Problems with Unawareness

Craig Innes 1  Alex Lascarides 1

Abstract
Structured models of decision making often assume an agent is aware of all possible states and actions in advance. This assumption is sometimes untenable. In this paper, we learn influence diagrams from both domain exploration and expert assertions in a way which guarantees convergence to optimal behaviour, even when the agent starts unaware of actions or belief variables that are critical to success. Our experiments show that our agent learns optimal behaviour on small and large decision problems, and that allowing an agent to conserve information upon discovering new possibilities results in faster convergence.

1. Introduction
Probabilistic graphical models have proven useful in representing richly-structured decision tasks (Koller, 2009). Many techniques exist to learn a model’s structure and parameters from experience (Tsamardinos et al., 2006; Bartlett & Cussens, 2017) and expert advice (Masegosa & Moral, 2013). Unfortunately, all such methods assume the way the domain is conceptualized—the belief and action variables used to describe the problem—is completely known in advance. Often, this assumption does not hold.

In medicine, for example, suppose an agent prescribes a drug, but later a senior pharmacologist objects to the prescription based on a reason unforeseen by the agent—the patient carries a newly discovered genetic trait, and the drug produces harmful side effects in its carriers. Further, this discovery may occur after the agent has already learned a lot about how other (foreseen) factors impact the drug’s effectiveness. As Coenen et al. (2017) point out, such scenarios are common in human discussion—the answer to a question may not only provide information about which of the questioner’s existing hypotheses are likely, but also reveal unforeseen hypotheses not yet considered. This example also shows that while it may be infeasible for an agent to gather all relevant factors of a problem before learning, it may be easy for an expert to offer contextually relevant advice during learning. Another example is in robotic skill learning. Methods like Cakmak & Thomaz (2012) teach a robot how to perform a new action it was unaware of, but not how to integrate it into an existing decision model.

Current learning and decision making models don’t address these issues; they assume the task is to use data to refine a distribution over a fixed hypothesis space. The above examples, however, illustrate a sort of reverse bayesianism (Karni & Viero, 2013), where the hypothesis space expands over time.

Rather than overcoming unawareness of states and actions, we could instead represent unawareness as an infinite number of hidden variables (Wood et al., 2006), or abandon learning a structured model and use densely connected hidden layers to implicitly learn structure from raw sensory input (Mnih et al., 2015). Both approaches have drawbacks. First, neither currently addresses how to adapt what one has learned when a new action is discovered. More importantly, the hidden layers/variables are not tied to grounded concepts with explicit meaning, so an agent cannot easily justify its decisions to a user, or articulate the limits of its current understanding to solicit help from others.

We instead propose an agent which explicitly overcomes its unawareness, and constructs an interpretable model of the problem. This paper makes three contributions: First, an algorithm which incrementally learns all components of a decision problem as an Influence Diagram (ID). This includes the reward function and probabilistic dependencies between variables, but also the set of belief and action variables themselves (Sections 2, 3.2). Second, an agent-expert communication protocol (Section 3.1) which interleaves contextual advice with learning, and guarantees our agent converges to optimal behaviour despite starting unaware of factors critical to success. Third, experiments on decision tasks of varying sizes showing our agent successfully learns optimal behaviour in practice (Section 4). Such contributions should be most relevant in interactive task learning (Laird et al., 2017), where agents must learn the underlying concepts which define a task.
2. Learning with Full Awareness

We focus on learning observable, single-stage decision problems with discrete states and actions. In this paper, we learn optimal behavior by learning an Influence Diagram (a Bayesian network extended with actions and rewards), which defines the optimal policy. We start with an agent who is aware of all states and actions, then address the task where the agent is unaware of relevant states and actions.

2.1. Making Optimal Decisions

Our agent’s goal is to learn the optimal policy \(\pi^*\) which, given evidence \(e\), chooses the action \(a\) which maximizes its expected utility \(EU(a|e) = \sum_{s'} P(s'|a,e)R(s')\) across all states \(s'\). If \(P(s'|a,e)\) or reward function \(R\) are unknown, the agent must learn them via trial and error. Our agent balances exploiting its current estimate of \(\pi^*\) with exploring the domain by using an \(\epsilon\)-greedy policy (one which acts randomly with probability \(\epsilon > 0\), and according to the current estimate of \(\pi^*\) with probability \(1-\epsilon\)). For any policy \(\pi\), we can measure the expected loss in reward against the true optimal policy \(\pi^*\) using (1)—the policy error.

\[
PE(\pi) = \sum_e P(e) (EU(\pi^*(e)|e) - EU(\pi(e)|e))
\] (1)

If there are many states, evaluating the expected utility directly is infeasible. The next sections show how to exploit a problem's inherent structure to make computations tractable.

2.2. Bayesian Networks

Bayesian networks (BNs) decompose the state space of the problem into a set of belief variables \(X = \{X_1, \ldots, X_n\}\). Each state \(s\) is a joint assignment to all variables in \(X\) (i.e., \(s \in v(X)\), where \(v(X)\) is the cross-product of possible assignments to each variable in \(X\)). An BN is a directed acyclic graph (DAG) \(Pa\) and parameters \(\theta\) which exploit the conditional independencies between variables in \(X\). For each \(X \in X\), \(Pa_X \subseteq Pa\) is the parent set of \(X\). Given \(Pa_X\), the value of \(X\) is conditionally independent of all \(Y \in X \setminus Pa_X\) which are non-descendants of \(X\). The parameters \(\theta_X \in \theta\) then define the conditional probability distributions \(\theta_X = P(X|Pa_X)\).

Given \(X\), we can learn the most likely structure \(Pa^*\) and parameters \(\theta^*\) from observed data \(D_{0,t} = [d_0, \ldots, d_t]\) (where \(d_t \in v(X)\)) by learning \(P(Pa|D_{0,t})\) and \(P(\theta|D_{0,t}, Pa)\). A common way to calculate \(P(Pa|D_{0,t})\) is to use the Bayesian Dirichlet Score (BD-Score) (Heckerman et al., 1995):

\[
P(Pa|D_{0,t}) \propto \prod_{X \in X} BD_t(X, Pa_X)
\] (2)

\[
BD_t(X, Pa_X) = \frac{P(Pa_X) \prod_j \beta(N_{ij}^t + \alpha_{ij}, \ldots, N_{mj}^t + \alpha_{mj})}{\beta(\alpha_{ij}, \ldots, \alpha_{mj})}
\] (3)

Here, \(\beta(n_1, \ldots, n_m)\) is the multivariate beta function, and \(N_{ij}^t = \sum_{X} Pa_X = \sum_{X} Pa_X\) is the number of states in \(D_{0,t}\) where \(X\) has assignment \(i\) and its parents have joint assignment \(j\). The \(\alpha_{ij}\) parameters come from the Dirichlet priors over \(\theta\) and act as a “pseudo-count” when data is sparse. We typically choose the prior \(P(Pa)\) to penalize complex structures, as in equation (4) which assigns a cost of \(\rho < 0.5\) for each extra parent:

\[
P(Pa_X) = \rho^{|Pa_X|} (1 - \rho)^{|X\setminus Pa_X|}
\] (4)

Unfortunately, evaluating the full posterior of (2), or even finding its maximum, is infeasible for even a moderate number of variables, as the number of possible BNs is hyper-exponential in the size of \(X\) (Tian & He, 2009). To tackle this, most methods approximate \(Pa^*\) via either local search (Teyssier & Koller, 2005), sampling methods (Madigan et al., 1995), or by reducing the search space of structures considered reasonable (Buntine, 1991). For our task, we have another issue—most BN learning methods operate once on a single batch of data, and thus consider the collection of sufficient statistics (i.e., the BD-scores of each \(Pa_X\) and their associated \(N_{ij}^t\)) to be a negligible pre-computed cost. In a decision problem, agents gather data incrementally, and exploit their current beliefs during learning, so we must update the sufficient statistics each time step.

To address these issues, we follow Buntine (1991) by constructing a reasonable parent lattice \(\mathcal{P}_X\) for each \(X\). Figure 1 shows an example lattice for \(X_1\). Starting from \(\emptyset\), we construct larger parent sets by combining sets seen so far, and track the highest scoring set according to (3). Any parent-set with a score lower than some proportion \(\kappa\) of the best score, is considered “unreasonable”, and is not expanded further.

Once we have our reduced set \(\mathcal{P}_X\) for each \(X\), we can find the combination of reasonable parent sets which maximize (2) and form a valid DAG. In our task we will be learning \(X\), so cannot assume we know a total-ordering \(\prec\) over \(X\) (as Buntine (1991) does), and thus cannot choose each \(Pa_X\) independently. Instead, following Bartlett & Cussens (2017), we treat finding \(Pa^*\) as the linear program given below:

1Where the context is clear, we compress this notation to \(N_{ij}^t\)
2In contrast to Buntine (1991) we also consider all subsets of a reasonable parent set to be reasonable.
Our method has three main advantages. First, reducing the number of reasonable parent sets and using the modular BD-score means we can incrementally update the probability of each $Pa_X$ using past calculations of (3) via (7).³

\[
\text{BD}_t(X, Pa_X) = \text{BD}_{t-1}(X, Pa_X) \frac{N_{ij}^t + \alpha_{ij}}{N_{ij}^t + \alpha_{ij} - 1}
\]  

Second, expressing structure learning as a linear program makes it easy to add extra structural constraints later (as we do in section 3.1). Third, each lattice $P_X$ implicitly captures an approximate distribution over each $X$’s parents.⁴

In section 3.2, the agent uses this distribution to conserve information as its awareness of $X$ expands.

2.3. Learning to Act with Influence Diagrams

An ID (Howard & Matheson, 2005) is a BN augmented with actions and rewards. It is a tuple $\langle A, X, R, Pa, \theta \rangle$. Here,

\[A = \{A_1, \ldots, A_m\}\]  

are the action variables, the agent controls (where a full action $a$ is a member of $v(A)$). As before, $X$ are the belief variables, but they are now partitioned into $X = B \cup O$, where $B$ is the set of variables the agent observes before taking action, and $O$ are the variables which describe the outcome of that action. The reward function $R : v(scope(R)) \rightarrow R$ gives the reward received for each state, where $scope(R) \subseteq X$ are the variables which determine the reward. While $Pa$ and $\theta$ still describe the probabilistic dependencies, each $Pa_X$ can now include belief and action variables (i.e., $\forall X, Pa_X \in X \cup A$). A domain trial at time $t$ is a tuple $d_t = (s_t, a_t, r_t)$, where $s_t = (b_t, o_t)$, $b_t \in v(B)$, $o_t \in v(O)$, $a_t \in v(A)$ and $r_t = R(s_t)$.

Figure 2 shows an example ID. Here, the agent chooses assignments for action variables TOPICAL CREAM and ORAL DRUG based on observing GENE X, then receives a reward based on outcomes DISEASE, and SIDE EFFECTS. Note that IDs often also include information arcs — edges from belief variables to action variables that show which variables an agent observes before making each decision. For simplicity of explanation, we omit information arcs, and assume in this paper that all action variables are set simultaneously, with access to observations of all variables in $B$.

We can use the same techniques from section 2.2 to learn $Pa^*$ and $\theta^*$ for IDs. Similarly, we can learn reward function $R_t$ from $scope(R)$, and $D_{0.5}$ by setting $R_t(s_t[scope(R)]) = r_t$,⁵ and by defaulting to indifference for unseen states (e.g., by setting the reward of unseen states to 0). Given a learned ID, the agent then chooses a policy which maximizes the expected utility from section 2.1 where $a \in v(A)$ and $c \in v(B)$.

3. Overcoming Unawareness

So far, we’ve assumed our agent was aware of all relevant variables in $X \cup A$ and all members of $scope(R)$. We now drop this assumption. From here onward we denote the true set of belief variables, actions, and reward scope as $X^+, A^+$ and $scope^+(R)$, and the learner’s awareness of them at time $t$ as $X^t, A^t$, and $scope^t(R)$.

³We include a proof of (7) in the technical supplement.

⁴Technically, without $\prec$, each $P_X$ is a distribution over markov blankets for $X$, but we enforce acyclicity, so this is not an issue.

⁵Here, $s[Y]$ projects assignment $s$ over the variables in $Y$. 

---

Figure 1. Parent lattice for $X_1$. Grey nodes are unreasonable

Figure 2. ID for medical example. Rectangles are action variables, oval are belief variables, and diamonds are reward nodes.
Suppose \( X^+ = \{ X_1, X_2 \}, X^0 = \{ X_1 \}, A^+ = \{ A_1, A_2 \}, A^! = \{ A_1 \} \). We assume the agent can’t observe the value of variables it is unaware of. In the medical example from section 1, if \( X_2 \) is a particular gene, we assume the agent cannot detect the presence of that gene if it is unaware it exists. We also assume the agent cannot perform actions it is unaware of (i.e., if the agent is unaware of variable \( A_2 \) then it cannot set it to a value other than its default value).\(^6\) This means at time \( t = 0 \), the agent does not observe the true trial \( d_0 \), but rather \( \langle s_0[ X^0], a_0[ A^0], r_0 \rangle \). But awareness of those missing factors may be crucial to learning \( \pi_+ \). For example, the best action may depend upon \( v(X_2) \), or the optimal policy may sometimes involve setting \( A_2 \) to true. The next sections thus answer two main questions. First, how can an agent discover and overcome its own unawareness by asking for help? Second, when an agent discovers a new variable, how can they integrate it into their current model while conserving what they have learned from past experience?

3.1. Expert Guidance

Our agent can expand its awareness via advice from an expert. Teacher-apprentice learning is common in the real world, as it allows learners to receive contextually relevant advice which may inform them of unforeseen concepts.

This paper assumes the expert is cooperative and infallible. Further, we abstract away the complexity of grounding natural language statements in a formal semantics and instead assume that the agent and expert communicate via a pre-specified formal language (though see e.g., (Zettlemoyer & Collins, 2007) for work on this problem). We do not, however, assume the expert knows the agent’s beliefs about the TD. The variables \( x^e \) and \( x^a \) express the expert’s knowledge of our agent’s awareness, and include only those actions the expert has explicitly seen the agent perform, or variables the agent mentions in conversation.

As argued in the introduction, the goal is to provide a minimal set of communicative acts so that interaction between the agent and expert mirrors human teacher-apprentice interaction. Concretely, this means we want our system to have two properties. First, the expert should mostly allow the agent to learn by themselves, interjecting only when the agent performs sufficiently poorly, or when they explicitly ask a question. Second, our expert should be able to give non-exhaustive answers to queries. This is because, in practice, it is unlikely a human expert will be able to give exhaustive answers to all questions due to either cognitive bounds or a lack of information. Following the maxims of Grice (1975), a cooperative speaker should give answers which provide just enough information to resolve the agent’s current dilemma. The next sections identify four advice types, whose combination guarantee the agent eventually behaves optimally, regardless of initial awareness.

3.1.1. Better Action Advice

If the expert sees the agent perform a sub-optimal action, it can tell the agent a better action to take instead. For example: “When it is raining, take your umbrella instead of your sun hat”. Our goal is to avoid interrupting the agent every time it makes a mistake, so we specify the following conditions for when the agent performs poorly enough to warrant correction: Let \( t \) be the current time step, and \( t' \) be the time the expert last uttered advice. When (8-12) hold:

\[
\sum_{t' \leq t} \frac{t - t'}{t - t'} > \mu \quad \text{(8)}
\]

\[
EU(\pi_+(b)|b) - EU(a|b_i) > \beta \quad \text{(9)}
\]

\[
EU(a_i|b_i) \geq r_t \quad \text{(10)}
\]

\[
\exists A', a' \in v(A^c \cup A'), EU(a'|b_i) > EU(a_i|b_i) \quad \text{(11)}
\]

\[
\not\exists A'', (a'' \in v(A^c \cup A''), EU(a''|b_i) > EU(a_i|b_i) \quad \text{(12)}
\]

Then the expert will utter advice of the form (13):

\[
EU(a'|w_t^A) > EU(a_i[A^t \cup A']|w_t^A) \quad \text{(13)}
\]

Equation (8) ensures some minimum time \( \mu \) has passed since the expert last gave advice, while (9) ensures the expert won’t interrupt unless its estimate of the agent’s policy error is above some threshold \( \beta \). Together, \( \mu \) and \( \beta \) define how tolerant the expert will be to the agent’s poor performance before deciding to give advice. Equation (10) ensures that the expert’s suggested action has an expected reward higher than what the agent received at time \( t \). Finally, (11-12) ensure not only that a better action \( a' \) exists, but also that \( a' \) introduces the minimum amount of potentially unforeseen action variables \( A' \) needed to improve the agent’s behaviour. This means there isn’t another action \( a'' \) which could be advised while revealing fewer unaware variables to the agent. This follows from our desire to give non-exhaustive advice, by first suggesting improvements which use concepts the agent is already aware of, rather than introducing many new action variables all at once.

Equation (13)—the expert’s utterance—uses an ambiguous term \( w_t^A \), whose intended meaning is the true state \( b \) (i.e \( [w_t^A] \in v(B^+) \)), but whose default interpretation by the agent is \( b[B^+] \). In words, the expert says “In the last step, a’ would have been better than a_i”. By introducing ambiguity, the agent can now interpret (13) in two ways. The first is as a partial description of the true
Learning Structured Decision Problems with Unawareness

\(15\) imply that any variable the expert utters must be relevant to the problem. In other words, it must exert influence on at least one variable in scope\(_+\)(\(R\)), and that there exists some state the agent could have reached which would yield a better reward than the one it got. In fact, we can enforce all equations of the form (14) when learning \(P_{a^*}\) by adding the constraint (16) to our linear program:

\[ \forall A \in A', \, A \in A^+ \land \exists X \in \text{scope}_{+}(\mathcal{R}), \, \text{anc}(A, X) \quad (14) \]

\[ \exists s \in v(A^+), \, s[B^t] = s_t[B^t] \land \mathcal{R}_{+}(s) > r_t \quad (15) \]

Where \(\text{anc}(X, Y)\) means \(X\) is the ancestor of \(Y\) (i.e., there is a directed path from \(X\) to \(Y\) in the true \(ID\)). Equations (14-15) imply that any variable the expert utters must be relevant to the problem. In other words, it must exert influence on at least one variable in scope\(_+\)(\(R\)), and that there exists some state the agent could have reached which would yield a better reward than the one it got. In fact, we can enforce all equations of the form (14) when learning \(P_{a^*}\) by adding the constraint (16) to our linear program:

\[ \forall Y \notin \text{scope}(\mathcal{R}) : \sum_{X \in A} I(P_{aX} \to X) \geq 1 \quad (16) \]

The second way the agent could use (13) is by adding its default interpretation of the advice to its current knowledge:

\[ \forall b \in B^+ : b[B^t] = b_t[B^t] \Rightarrow EU(a' | b) > EU(a | b) \quad (17) \]

The agent can then enforce (17) by choosing \(a'\) whenever \(b[B^t] = b_t[B^t]\), regardless of what seems likely from \(D_{0,t}\). We should now see that even with a cooperative, infallible expert, even abstracting away issues of grounding natural language, misunderstandings can still happen due to differences in agent and expert awareness. As the next section shows, such misunderstandings can reveal gaps in the agent’s awareness and help to articulate queries whose answers guarantee the agent expands its awareness.

Lemma 1 guarantees the expert’s advice strategy continues to reveal unforeseen actions to the agent so long as its performance in trials exceeds the expert’s tolerance.\(^7\)

**Lemma 1.** Consider an agent with awareness \(X^t \subset X^+\), and expert following (8-13). As \(k \to \infty\), either \(PE(\pi_{t+k}) \to c \leq \beta\) or the expert utters (13) where \(A^t \neq \emptyset\).

### 3.1.2. Resolving a Misunderstanding

We noted before that the agent’s default interpretation of (13) could lead it to misunderstand the expert’s intended meaning. To illustrate, suppose the agent receives advice (18) and (19) at times \(t - k\) and \(t:\)

\[ EU(a | w_{t-k}^{b}) > EU(a' | w_{t-k}^{b}) \quad (18) \]

\[ EU(a | w_{t}^{b}) < EU(a' | w_{t}^{b}) \quad (19) \]

While the intended meaning of each statement is true, the agent’s default interpretations of \(w_{t-k}^{b}\) and \(w_{t}^{b}\) may be identical. That is, \(b_{t-k}[B^t] = b_t[B^t]\). From the agent’s perspective, (18) and (19) conflict, and thus give the agent a clue that its current awareness of \(X^+\) is deficient. To resolve this conflict, the agent asks (20) (in words, “which \(B\) has distinct values in \(b_{t-k}\) and \(b_t\?)”)

\[ ?\lambda B(B \in B^+ \land s_{t-k}[B] \neq s_t[B]) \quad (20) \]

\[ B' \in B^+ \land \exists X \in \text{scope}_{+}(\mathcal{R}), \text{anc}(B', X) \quad (21) \]

Notice there may be many variables in \(B^+ \setminus B^t\) whose assignments differ in \(b_{t-k}\) and \(b_t\). Thus, the expert’s answer can be non-exhaustive. This means the agent must abandon its previous defeasible interpretations of the form (17), but can keep (14-15), as these are true regardless of what the true \(ID\) is. Lemma 2 guarantees the expert will reveal new belief variables provided misunderstandings can still arise.

**Lemma 2.** Consider an agent with awareness \(X^t \subset X^+\), \(A^t = A^+\). If \(\exists b', \exists b \neq b'\), \(b[B^t] = b'[B^t]\) and \(\pi_+(b) \neq \pi_+(b')\), then as \(k \to \infty\), either \(PE(\pi_{t+k}) \to c \leq \beta\), or the expert utters (21) such that \(B' \notin B^t\).

### 3.1.3. Unforeseen Rewards

In typical IDs (where we assume the agent starts fully aware of \(X^+\), \(A^+\), and \(\text{scope}_{+}(\mathcal{R})\)), we tend only to think of the trials as providing counts. For an unaware agent, a trial \(d_t = (s_t, a_t, r_t)\) also encodes monotonic information:

\[ \exists s, s[X^t] = s_t[X^t] \land \mathcal{R}_{+}(s) = r_t \quad (22) \]

This, along with (15), constrain the form of \(R\) the agent can learn. Recall that \(\text{scope}_{+}(\mathcal{R})\) may be only a subset of \(\text{scope}_{+}(\mathcal{R})\), so it might be impossible to construct an \(R : v(\text{scope}_{+}(\mathcal{R})) \to R\) satisfying all descriptions (22) gathered so far. Further, those extra variables in \(\text{scope}_{+}(\mathcal{R}) \setminus \text{scope}_{+}(\mathcal{R})\) may not be in \(X^t\). To resolve this, if the agent fails to construct a valid \(R\), it asks (23) (in words, “which variable \(X\) (that I don’t already know) is in \(\text{scope}(\mathcal{R})\?)”)

\[ ?\lambda X(X \in \text{scope}_{+}(\mathcal{R}) \land X \neq X^t) \quad (23) \]

\[ X'' \in \text{scope}_{+}(\mathcal{R}) \land X'' \in X^+ \quad (24) \]
Again (24) may be non-exhaustive. Even so, we can guarantee that the agent’s reward function eventually equals $R_+$.

**Lemma 3.** Consider an $\epsilon$-greedy agent with awareness $A^t = A^+, \text{scope}_+(R) \subseteq \text{scope}_+(R)$. As $k \to \infty$, there exists a $K$ such that $\forall s, \forall k \geq K, R_{t+k}(s) = R_+(s)$.

### 3.1.4. Unknown Effect

Recall that, to keep the problem tractable, our agent searches for IIDs in the space of “reasonable” parent sets $P$. Unfortunately, there might be no valid IID within $P$ which also satisfies the constraints of the form (16). The most obvious case of this is when a variable $X \notin \text{scope}_+(R)$ has no children in any reasonable DAG (i.e., $\forall Y, \forall P_{AV} \in P_Y, X \notin P_{AV}$).

Here, the agent can ask why $X$ is relevant by asking (25) (“what $V$ does $X$ affect directly?”). The answer (e.g., $X \in P_{AV}$) informs past (26) to the agent:

$$\varnothing V(X \in P_{AV})$$

$$V' \in X^+ \land \exists Y \in \text{scope}_+(R), \text{anc}(V', Y)$$

### 3.2. Adapting to Unforeseen Factors

Section 3.1 showed four ways to expand the agent’s awareness of $A$, $A^t$, and $\text{scope}(R)$. To improve on simply restarting learning, we must now say how the agent adapts its beliefs about $P_a$ and $\theta$ when its awareness expands.

#### 3.2.1. Adding a New Belief Variable

When the agent discovers a belief variable $Z$, its old distributions over parent sets no longer cover all possible parents. That is, for $X \in A^{t-1}$, $P_X$ did not cover cases where $Z$ was a parent of $X$. Worse, the agent cannot observe $Z$’s values in $D_{0,t}$, so cannot observe $N_{t-k, Z=1|ij}$ or $N_{t-k, X=1|ij}$ when $Z \in P_{AX}$. The $\alpha$-parameters involving $Z$ are also undefined, yet we need them to calculate $P_{\alpha^*}$ and $\theta^*$ via (6-7).

Discovering a new variable makes the size of each (observed) state dynamic, in contrast to standard problems where they are static (e.g., $\{X_1 = 0, X_2 = 1\}$ becomes $\{X_1 = 0, X_2 = 1, Z = ?\}$). We could phrase this as a missing data problem: $Z$ was hidden in the past but visible in future states, so estimate missing values and structure via e.g., expectation maximization (Friedman, 1998). But such methods commit us to costly passes over the full batch of data and makes finding $P_{\alpha^*}$ for problems with more than a handful of variables intractable by eliminating the modularity of the BD-score. Alternatively, we could ignore states with missing information when we require counts involving $Z$. For example, we could use $P(P_{AX}|D_{t-n})$ to score $P_{AX}$ when $Z \in P_{AX}$ but use $P(P_{AX}|D_{0,n})$ when $Z \notin P_{AX}$.

But as Friedman & Goldszmidt (1997) point out, most structure scores, including (2), assume we evaluate models with the same data. If we compare models using different data sets (even two sets from the same distribution), the learner favours models evaluated with the smallest data set.

Instead, our method discards data $D_{0:t-1}$ gathered during the learner’s old view of the hypothesis space, but conserves the posterior probabilities learned from it to create new priors for $P_{AX}$ and $\theta$ in the expanded space. On discovering the new variable $Z$, the agent updates each distribution over the parent sets $P_{AX} (X \neq Z)$ to cover parent sets which include $Z$. Equation (27) creates a new prior $P'(P_{AX})$ from the old posterior. Here, $C$ is the normalizing constant and $0 < \gamma \leq 1$ expresses how much we trust our current $P$.

$$P'(P_{AX}) = \gamma P'(P_{AX}|P_X) + (1 - \gamma)P(P_{AX})$$

$$P'(P_{AX}|P_X) =$$

$$\begin{cases} 0 & \text{if } P_{AX} \notin P_X \\ \frac{1 - \rho}{t}BD_{t-1}(X, P_{AX}) & \text{if } Z \notin P_{AX} \\ \frac{\gamma}{t}BD_{t-1}(X, P'_{AX}) & \text{if } P_{AX} = P'_{AX} \cup \{Z\} \end{cases}$$

This update preserves the relative likelihood among the parent sets without $Z$. It also preserves our bias towards simpler structures by giving only a small portion $\rho$ of the probability mass of parent sets without $Z$ to those with $Z$. This still leaves us to define a distribution over $P_{AX}$ itself. The agent has no evidence on $Z$’s parents at the moment of $Z$’s discovery, so defaults to using the initial prior (4).

To adapt $\theta^*$, we return to the issue of the counts $N_{ij}$ and $\alpha$-parameters. As before, we wish to avoid the complexity of estimating $Z$’s past values. Instead, we throw away the past states $D_{0:t-1}$ and their counts $N_{ij}$, but keep the relative likelihoods they gave rise to by packing these into new $\alpha$-parameters, as shown in (29) ($K$ is a constant):

$$N_{ij} = 0 \text{ for all } i, j, X, \rho_{AX}$$

$$\alpha_{ij} = \begin{cases} \frac{K}{|Z|}P(j|\theta_{t-1}^*) & \text{if } X = Z \\ \frac{K}{|Z|}P(i, j|P_{AX} \setminus Z)|\theta_{t-1}^* & \text{if } Z \in P_{AX} \\ K|P(i, j|\theta_{t-1}^*) & \text{otherwise} \end{cases}$$

Equation (29) summarizes the counts $D_{0:t-1}$ based on inferences from the old IID, then encodes these inferences in the $\alpha$-parameters of the new parameter prior. The new $\alpha$-parameters ensure that the likelihoods inferred from past states bias the estimate of $P_{\alpha^*}$ and $\theta^*$ as more trials arrive. The larger $K$ is, the more what it learned before discovering $Z$ influences its reasoning after discovering $Z$.

#### 3.2.2. Adding a New Action Variable

When the agent discovers a new action variable $A$, the same issue arises—the agent did not consider $A$ as a possible
Algorithm 1 outlines the entire learning process. In most was previously invalid.

Algorithm 1

1: **Input:** \( A^0, X^0, P^0, \theta^0, \rho^0, \tau^0 \)
2: for \( t = 1 \ldots \text{maxTrials} \) do
3: \( b_t \leftarrow \text{Generate new state} \)
4: \( (\alpha_t, a_t, r_t) \leftarrow \text{\epsilon-GREEDY}(b_t[X^t]) \)
5: \( N_{ij}^t \leftarrow \text{Update with } b_t[X^t], \alpha_t[X^t] \)
6: \( BD(X, Pa_X) \leftarrow \text{Update via (7)} \) for \( Pa_X \in P^X \)
7: if \( t \equiv 0 \pmod{\tau} \) then
8: \( \text{Revise } P^t \)
9: end if
10: \{ Unknown Effect (Section 3.1.4) \}
11: \{ Unforeseen Rewards (Section 3.1.3) \}
12: \{ Resolving a Misunderstanding (Section 3.1.2) \}
13: \{ Better Action Advice (Section 3.1.1) \}
14: if \( (8-12) \) are true then
15: \( \text{actAdvice} \leftarrow \text{Expert advises (13)} \)
16: Update \( A^t, \theta^t \) according to \( \text{actAdvice} \)
17: \{ \text{Exploring a New Factor (Section 3.1.4)} \}
18: for \( \text{past utterances} \) do
19: \( \text{actAdvice} \leftarrow \text{Better Action Advice} \)
20: end for
21: end if
22: if \( X^{t-1} \neq X^t \) then
23: Update \( N_{ij}^t, \alpha_{i,j}, P, P(Pa) \) via (4, 27, 29)
24: end if
25: \( (Pa^t, \theta^t) \leftarrow \text{\textbf{Solve (5) with (16)}, and (6)}}
26: end for

parent to any \( X \in X^{t-1} \), so must revise \( P \) with \( A \) as a possibility. Unlike for belief variables, however, we don’t need to discard \( D_{0,t-1} \). Since we assume that the agent cannot influence action variables it is unaware of, we can simply fill in the default value of \( A \) in all past trials \( D_{0,t-1} \).

3.2.3. EXPANDING THE REWARD FUNCTION SCOPE

Learning that a variable \( X \) is in the scope of \( R \) may cause us to revise \( Pa^* \), even if \( X \in X^{t-1} \). This is because expanding \( \text{scope}_{e_{\text{ID}}} \) loosens the constraints of the form (16) which may allow us to construct a higher scoring \( \text{ID} \) structure that was previously invalid.

Algorithm 1 outlines the entire learning process. In most steps, the agent incrementally revises its model based on the latest trial and current reasonable parents. If enough time \( \tau \) passes, or the agent’s awareness expands, the agent makes larger changes to its model, including revising \( P \). Given algorithm 1, theorem 1 guarantees our agent converges to a near-optimal policy (where “near-optimal” is bounded by the experts tolerance \( \beta \)), regardless of initial awareness (provided \( P \) includes the true structure).

**Theorem 1.** Consider an agent with initial awareness \( \lambda^0 \subseteq \lambda^+, A^0 \subseteq A^+, \text{scope}_0(R) \subseteq \text{scope}_+(R) \) following algorithm 1 without any pruning of the parent lattice (i.e., \( \kappa = 0 \)). As \( t \to \infty \), \( PE(\pi_t) \to c \leq \beta \).

4. Experiments

Our experiments show that agents following algorithm 1 converge to near-optimal behaviour not only in theory but also in practice. We also show that conserving information improves results. We do not investigate assigning explicit costs to agent-expert communication, but do show how varying the expert’s tolerance affects the agent’s performance.

We tested agents on three randomly generated \( \text{IDs} \) of increasing size: 12, 24, and 36 variables. Our results were similar across all sizes, but the differences between agents were most pronounced on the largest case, so we present those here (Full ID specifications and results for the small and medium cases are included in the technical supplement). In each, our agent begins with minimal awareness of the true \( \text{ID} \) \( (A^0 = \{O_1\}, X^0 = \{A_1\}, \text{scope}_0(R) = \{O_1\}) \). The agent acts in 5000 trials, using an \( \epsilon \)-greedy policy \( (\epsilon = 0.1) \).

We repeat experiments 100 times and average the results.

We use the cumulative reward across trials as our evaluation metric, which acts as a proxy for the quality of the agent’s policy over time. To make the results more readable, we apply a discount of 0.99 at each step, resulting in the metric \( R_{t-1}^{\text{Disc}} = r_t + 0.99 \cdot R_{t-1}^{\text{Disc}} \) for all \( t \).

We test several variants of our agent to show our approach is effective. The default agent follows algorithm 1 as is, with parameters \( \kappa = 0.001, \tau = 100, \rho = 0.1, \gamma = 0.99, K = 5.0, \mu = 10, \beta = 0.01 \) in equations (4), (8), (9), (27), and (29). The nonConservative agent does not conserve information about \( P, Pa \) nor \( \theta \) via (27) and (29) when it discovers a new factor. Instead, it discards all trials and reverts to the original prior of (4). We include this agent to show the value of conserving information as \( X \) and \( A \) expand. The non-relevant agent is like the default, but does not include any constraints of the form (16) when searching for \( Pa^* \). This means the agent might learn IDs where variables are completely disconnected. The truePolicy and random agents start knowing the true ID, and execute an \( \epsilon \)-greedy version of \( \pi_0 \), or choose a random action respectively. These agents give an upper/lower bound on performance. The lowTolerance and highTolerance agents change the expert’s tolerance to \( \beta = 0.001 \) and \( \beta = 0.1 \).

4.1. Results

Across all \( \text{IDs} \), our default agent converges to the optimal policy, despite starting unaware of factors critical to success. Figure 3a shows the cumulative reward of the default agent compared to the non-conservative, non-relevant, true-policy and random agents. The default agent converges to the op-
Learning Structured Decision Problems with Unawareness

Figure 3. Rewards (a and b) and size of $|A|$ (c) over time on the 36 variable task. Results for other tasks included in the supplement.

timal policy, and does so faster than its non-conservative counterpart. This shows conserving one’s beliefs on discovering new factors has value. The difference in reward compared to the non-relevant agent is smaller than expected. We suspect this is because actions which exert a strong causal influence over $\text{scope}^+(R)$ will be an ancestor to $\text{scope}^+(R)$ in $P_{a^*}$ regardless of whether we enforce (16). Actions the non-relevant agent leaves disconnected are those which exert little causal influence over $\text{scope}^+(R)$ (and thus usually have little effect on the optimal policy). This means that how much abandoning connectivity affects performance depends on how noisy the causation in the domain model is, which we don’t know with certainty in advance.

Figure 3b shows how the agent’s performance differs when varying the expert’s tolerance level: As the tolerance increases, the agent takes longer to converge towards a good policy, and also learns a worse final policy. Figure 3c shows why: Early on, the agent’s policy is “good enough” for the high tolerance expert, meaning the expert does not reveal the last few extra action variables which would net only small increases in the agent’s total reward.

5. Related Literature

Models of unawareness exist in logic (Board et al., 2011; Heifetz et al., 2013) and game theory (Feinberg, 2012), but interpret (un)-awareness from an omniscient perspective. We instead model awareness from the agent’s perspective and offer methods to overcome one’s unawareness.

Contextual bandits (Langford & Zhang, 2007) also deal with single-shot decision making as covered in this paper. Unlike our work, they focus on unstructured, noisy rewards, and don’t address unawareness of actions or states. Rong (2016) present Markov Decision Processes with Unawareness, which let an agent learn optimal behaviour even when unaware of some actions. We build on this work by providing a concrete mechanism for discovering unforeseen factors and by allowing the agent to discover explicit belief variables rather than atomic states—this is necessary to tackle large, richly-structured problems. McCallum & Ballard (1996) also learn an increasingly complex representation of the state space by gradually distinguishing between states which yield different rewards. Rather than dealing with unawareness, the focus of such approaches is on refining an existing state space. In other words, they do not support introducing unforeseen states or actions that the learner was unaware of before learning.

Several works use expert interventions to correct agent behaviour (Torrey & Taylor, 2013; Stone, 2009), or via agent-driven active learning (Masegosa & Moral, 2013; Murphy, 2001). Yet all such methods assume the expert’s intended meaning can be understood without expanding the agent’s conception of the state and action space. Our work allows experts to utter advice where ambiguity arises from their greater awareness of the problem.

6. Conclusion

We have presented an expert-assisted framework for learning structured decision problems, even when the learner starts unaware of factors critical to success. Further, our experiments show that being conservative about one’s beliefs improves the effectiveness of learning. In future work, we intent to extend our model to sequential tasks, as well as single-shot decision problems (Innes & Lascarides, 2019). Additionally, we aim to lift some assumptions imposed on the expert, and expand the range of situations in which the agent can ask for advice. For instance, we could let the expert be fallible (Masegosa & Moral, 2013), or leverage clues about the location of hidden variables present in the structure of the learner’s graphical model (Elidan et al., 2000) to guide questions asked to the expert.

Acknowledgements

This work is supported by EPSRC (UK). We would like to thank Kobi Gal, Subramanian Ramamoorthy, Benji Rosman,
Learning Structured Decision Problems with Unawareness

Stefano Albrecht, Vaishak Belle, and anonymous reviewers for their helpful feedback.

References


