Handling Ambiguous Effects in Action Learning

Boris Lesner, Bruno Zanuttini

Handling Ambiguous Effects in Action Learning

Boris Lesner and Bruno Zanuttini

GREYC, Université de Caen Basse-Normandie, CNRS UMR 6072, ENSICAEN
{boris.lesner,bruno.zanuttini}@unicaen.fr

Abstract. We study the problem of learning stochastic actions in propositional, factored environments, and precisely the problem of identifying STRIPS-like effects from transitions in which they are ambiguous. We give an unbiased, maximum likelihood approach, and show that maximally likely actions can be computed efficiently from observations. We also discuss how this study can be used to extend an RL approach for actions with independent effects to one for actions with correlated effects.

Keywords: stochastic action, maximum likelihood, factored MDP

1 Introduction

Learning how one's actions affect the environment is a central issue in Artificial Intelligence, and especially in Reinforcement Learning (RL) problems. This task is far from being trivial because of many factors. In particular, actions may affect the environment differently depending on the state where they are taken. They may also be stochastic, that is, have different possible outcomes, each of which occurs with some probability each time we take the action.

A third difficulty, which is our focus in this article, is that the effects of an action are typically *ambiguous* in some states. Precisely, if some fact is true in the state where we took the action *and* in the resulting state, then it is ambiguous whether this fact was set to true by the action, or was simply unaffected by it (and persisted true). This of course makes no difference in the starting state in question, but may matter a lot if we want to generalize the corresponding observation to other states, where the fact in question is not true.

To give a concrete example, assume that you want to learn the effects of the action "play the National Lottery" by observing your neighbours. If one of them suddenly has an expensive car parked in front of his house, then he must have won (effect "become rich" occurred). On the other hand, assume another one always has had an expensive car parked in front of his huge villa. Then you will see no change when he wins, nor when he looses. Technically, for this neighbour (state) both winning and losing (effects) provoke no change (self-transition). Then, this ambiguity must be resolved for generalizing from a few "poor" neighbours and a few "rich" neighbours to a generic player.

We address here this specific problem for environments and actions described in factored form over propositional variables, with STRIPS-like effects. We provide an in-depth study of ambiguity and of maximum likelihood approaches to

learning in the presence of ambiguous effects. Though the obvious motivation for this work is RL, we place our study in a general learning context, where some transitions between states are observed, from which the effects of the (unique) stochastic action which provoked them must be learnt together with their probabilities. In this setting, we propose a linear programming approach for computing the most likely effect distributions, and use it to propose a generalization of SLF-Rmax [12] to deal with ambiguous effects.

Ambiguous effects are discussed in several works [13, in particular], and implicitly dealt with in approaches to *relational RL* [5, 10, 11]. Still, to the best of our knowledge they have never been investigated in depth. For instance, propositional actions are often learnt as dynamic bayesian networks (DBNs) which assume independent effects on variables [6, 4, 12]. Then ambiguity is bypassed by learning the effects on x independently in states satisfying x and $\overline{x}$. But such DBNs (without synchronic arcs) cannot represent effect distributions as simple as $\{(x_1x_2, .5), (\emptyset, .5)\}$, while STRIPS-like descriptions, as we study here, are fully expressive (see the discussion by Boutilier *et al.* [1]).

Closest to ours is the work by Walsh *et al.* [13]. Our work can in fact be seen as using their linear regression approach (for known effects) with all 3^n possible effects, but introducing new techniques for reducing the exponential number of unknowns and dealing with ambiguity.

Section 2 introduces our formal setting. In Sections 3–5 we study the problem of computing most likely actions given observed transitions. We give an application to RL and conditional actions in Section 6, and conclude in Section 7.

2 Formal setting

We consider propositional environments, described through n Boolean variables $x_1, \ldots, x_n$. A *state* s is an assignment to $x_1, \ldots, x_n$. There is a hidden stochastic action a, which can be described as a probability distribution on a finite set of effects (we identify the action and the distribution to each other). We write $\{(p_i, e_i) \mid i = 1, \ldots, \ell\}$ (with $\sum_i p_i = 1$) for this distribution, meaning that each time a is taken, exactly one of the effects e_i occurs, as sampled i.i.d. according to the p_i's. Effects are STRIPS-like, that is, each effect is a consistent term on $x_1, \ldots, x_n$. When a is taken in some state s and effect e_i occurs, the resulting state s' is obtained from s by changing all affected variables to their value in e_i. We write $s' = \mathit{apply}(s, e_i)$. For convenience, states and terms are also seen as sets of literals, and 3^X denotes the set of all terms/effects.

For instance, for $n = 3$, $s = 000$ and $s' = 010$ are two states, and $e_1 = \overline{x}_1x_2$, $e_2 = \overline{x}_3$ are two effects. The distribution $\{(e_1, .7), (e_2, .2), (\emptyset, .1)\}$ defines an action a. Each time a is taken in a state, e_1 occurs (with probability .7), or e_2 occurs (.2), or nothing occurs (.1). If effect e_1 occurs in state s above, the resulting state is s', while effects e_2 and $\emptyset$ provoke a transition from s to itself.

Ambiguous Effects and Compact Representation. Ambiguity is a consequence of the following straightforward result.

Proposition 1. *Let s, s' be two states. Then the effects $e \in 3^X$ which satisfy $apply(s, e) = s'$ are exactly those for which $s' \setminus s \subseteq e \subseteq s'$ holds.*

It follows that the effects which may have occurred when a transition from s to s' is observed, can be *compactly* represented, as the set interval $[s' \setminus s, s']$. As an example, the effects which provoke a transition from $s = 000$ to $s' = 010$ are x_2, $\overline{x}_1 x_2$, $x_2\overline{x}_3$, and $\overline{x}_1 x_2 \overline{x}_3$. That is, we are sure that the (atomic) effect x_2 occurred, but any other literal in s' may have been set to true, or *left* true, by the action. These effects are compactly represented by the set interval $[x_2, \overline{x}_1 x_2 \overline{x}_3]$.

Set intervals obey the rule $[e_1, e_2] \cap [e_3, e_4] = [e_1 \cup e_3, e_2 \cap e_4]$. For instance, $[x_2, \overline{x}_1 x_2 \overline{x}_3] \cap [\emptyset, x_1 x_2 \overline{x}_3]$ is $[x_2, x_2\overline{x}_3]$, and $[x_2, \overline{x}_1 x_2 \overline{x}_3] \cap [\overline{x}_2\overline{x}_3, x_1 x_2 \overline{x}_3]$ is empty, since $x_2\overline{x}_2\overline{x}_3 \not\subseteq x_2\overline{x}_3$ holds (here, because x_2 and $\overline{x}_2\overline{x}_3$ are inconsistent together).

Sampled Effects and Induced Observations. We consider an agent which must learn the effect distribution of a unique action a, from transitions provoked by this action in the environment. Remember that (except in Section 6) we assume a to be unconditional, i.e., to have the same effect distribution D in all states s.

For generality, we assume nothing about the states s in which transitions are observed; what we only require is that the environment samples the effects fairly to D. This setting arises naturally in many contexts. For instance, a robot may train in the factory, so as to learn accurate models of its actuators, but the situations s which it can experiment may be restricted by the facilities in the factory. In an RL setting, the agent may lose control on the visited states if other agents or exogenous events can intervene at any moment in the environment.

Formally, the examples sampled by the environment form a multiset of pairs state/effect, of the form $\mathcal{T} = \{(s_i, e_i) \mid i = 1, \ldots, m\}$, and the learner sees the corresponding multiset of transitions between states $\{(s_i, s'_i) \mid i = 1, \ldots, m\}$ (with $s'_i = apply(s_i, e_i)$). For simplicity of presentation, we define the *observations* induced by $\mathcal{T}$ to be the multiset of intervals $O_\mathcal{T} = \{[s'_i \setminus s_i, s'_i] \mid i = 1, \ldots, m\}$, and assume that this is the only data available to the learner. That this information is equivalent to the multiset of transitions is a consequence of Proposition 1, together with the observation that s_i can be easily retrieved from $s'_i \setminus s_i$ and s'_i. Importantly, observe that e_i is always in the set interval $[s'_i \setminus s_i, s'_i]$.

Example 1. Let $D = \{(x_1, .5), (x_1\overline{x}_2, .3), (\emptyset, .2)\}$, and let

$$\mathcal{T} = \{(01, x_1), (00, x_1), (00, \emptyset), (00, x_1\overline{x}_2), (01, x_1)\}$$

in which, for instance, the first element corresponds to state $s = 01$ having been chosen, and effect x_1 having been sampled. The observations induced by $\mathcal{T}$ are

$$O = \{[x_1, x_1x_2], [x_1, x_1\overline{x}_2], [\emptyset, \overline{x}_1\overline{x}_2], [x_1, x_1\overline{x}_2], [x_1, x_1x_2]\}$$

The *frequency* of an observation $o = [s' \setminus s, s']$ in $O_\mathcal{T}$, written $f_{o,O_\mathcal{T}}$ (or f_o), is the proportion of indices i such that $[s'_i \setminus s_i, s'_i]$ is precisely o in $O_\mathcal{T}$. In Example 1, $f_{[x_1, x_1\overline{x}_2]}$ is $1/5 + 1/5 = .4$. Observe that this corresponds in fact to the manifestation of two different effects (but this is hidden to the learner).

3 Most Likely Actions

We now characterize the maximally likely effect distributions D (i.e., actions) given observations $O = \{[s'_i \setminus s_i, s'_i] \mid i = 1, \ldots, m\}$. First recall from Bayes rule that the likelihood of D given O satisfies $Pr(D|O) = Pr(D)\,Pr(O|D)/\,Pr(O)$. In particular, if all distributions D have the same *prior* likelihood $Pr(D)$, the most likely distribution D given O is the one which maximizes $Pr(O|D)$.

Likelihood and Fairness. We first need a few straightforward lemmas. Write O for a multiset of observations induced by states/effects in $\mathcal{T}$, and $D = \{(p_i, e_i) \mid i \in I\}$, $D' = \{(p'_j, e'_j) \mid j \in J\}$ for two effect distributions. Moreover, write $||D - D'||_1$ for the L1-distance between D, D', that is, $||D - D'||_1 = \sum_{e \in 3^X} |p_e - p'_e|$, with $p_e = p_i$ for $e = e_i$ in D and $p_e = 0$ otherwise, and similarly for p'_e. For instance, the L1-distance from the distribution in Example 1 to the distribution $\{(x_1, .6), (\emptyset, .4)\}$ is $(.6 - .5) + (.3 - 0) + (.4 - .2) = 0.6$.

Lemma 1. *If for some $o \in O$, there is no $i \in I$ with $e_i \in o$, then $Pr(O|D)$ is 0.*

This follows from the fact that when a transition from s to s' is observed, the effect which provoked it must be in the interval $[s' \setminus s, s']$.

Now write $D_{\mathcal{T}} = \{(e, p_e)\}$ for the distribution of effects in $\mathcal{T} = \{(s', e')\}$, that is, p_e is given by $p_e = |\{(s', e') \in \mathcal{T} \mid e' = e\}| \;/\; |\mathcal{T}|$.

Lemma 2. *$Pr(O|D) > Pr(O|D')$ is equivalent to $||D_{\mathcal{T}} - D||_1 < ||D_{\mathcal{T}} - D'||_1$.*

This follows, e.g., from the following bound [14], which extends Chernoff's bound to multivalued random variables:

$$Pr(||D_{\mathcal{T}} - D||_1 \geq \epsilon) \leq (2^\ell - 2)\mathrm{e}^{-m\epsilon^2/2} \tag{1}$$

where m is the number of samples observed (size of $\mathcal{T}$) and ℓ is the number of effects (with nonzero probability) in D. Note that other distances could be used. We use L1-distance because approximating the transition probabilities $T(\cdot|s, D)$ in an MDP with respect to it provides guarantees on the resulting value functions, but the infinite norm, for instance, gives similar results [7, "simulation lemma"]).

These observations lead us to introduce the following definition.

Definition 1. *Let $D = \{(p_i, e_i) \mid i \in I\}$ be an effect distribution. A multiset of observations O is said to be* ϵ-fair to D *if*

1. *for all intervals $o \in O$, there is an $i \in I$ with $p_i \neq 0$ and $e_i \in o$, and*
2. *there is a multiset of states/effects $\mathcal{T}$ such that O is induced by $\mathcal{T}$ and the distribution of effects in $\mathcal{T}$ is at L1-distance at most ϵ to D.*

Hence (Lemmas 1–2) the effect distributions D to which O is 0-fair are perfectly likely given O, and otherwise, the smaller ϵ, the more likely D given O.

Clearly enough, a multiset O is always 0-fair to at least one distribution D. Such D can be built from one effect e_o in each interval $o \in O$, with the frequency

f_o as its probability. However, there are in general many such "perfectly likely" distributions. In fact, each possible choice of effects in the construction above leads to a different one. Further, in some given observed interval it may be the case that two different effects provoked the corresponding transitions, increasing still more the number of candidate distributions. Still worse, the different most likely distributions may be arbitrarily distant from each other (hence inducing arbitrarily different planning problems, if learnt actions are used for planning).

Example 2. Let $o_1 = [x_1, x_1x_2], o_2 = [x_2, x_1x_2]$, and $O = \{o_1 \times 5, o_2 \times 5\}$ (o_1, o_2 observed 5 times each). Then O is 0-fair to $D_1 = \{(x_1, .5), (x_2, .5)\}$, to $D_2 = \{(x_1x_2, 1)\}$, and to $D_3 = \{(x_1, .4), (x_1x_2, .3), (x_2, .3)\}$. As to the last point, D_3 indeed induces O if x_1x_2 is sampled once in $s = 01$ and twice in $s = 10$.

Now despite O is 0-fair to D_1 and to D_2, their L1-distance is maximal (2), and so are the transition functions $T(\cdot|00, \cdot)$ which they induce in $s = 00$.

Summarizing, we have that (with high confidence) O is always ϵ-fair to the hidden effect distribution D which indeed generated it, for some small ϵ. Nevertheless, O can also be ϵ-fair, for some small ϵ, to some distribution which is very different from D. Another way to state this is that fairness gives a lower bound on the distance to the real, hidden D. Anyway, this lower bound is a correct, unbiased measure of likelihood (Lemma 2).

Computing fairness. If some additional constraints have to be satisfied (like accounting for other observations at the same time), in general there will not be any effect distribution to which the observations are 0-fair. Hence we investigate how to compute in general how fair a given multiset of observations is to a given effect distribution.

To that aim, we give an alternative definition of fairness. To understand the construction, observe the following informal statements for a sufficiently large multiset of observations O generated by a distribution $D = \{(p_i, e_i) \mid i \in I\}$:

- if e_i induced $o \in O$, then the observed frequency of o should be close to p_i,
- it is possible that some $e_i \in D$ with a sufficiently small p_i has never been observed, i.e., that O contains no interval o with $e_i \in o$,
- every $o \in O$ *must* be induced by some effect e_i with $p_i \neq 0$,
- if $o_1, o_2 \in O$ intersect, it may be that only one $e_i \in D$ induced both (in different states s), and then f_{o_1}, f_{o_2} should approximately sum up to p_i,
- dually, if $o \in O$ contains e_i, e_j, it may be that both participated in inducing o, and then p_i, p_j should approximately sum up to f_o.

Generalizing these observations, we can see that one effect may induce several observed intervals, and dually that one interval may be induced by several effects. We take these possibilities into account by introducing values which we call *contributions of effects to intervals*, written $c_{e,o}$, and reflecting how often, among the times when e was sampled, it induced interval o.

We are now ready to characterize ϵ-fairness. The value to be minimized in Proposition 2 is the L1-distance between D and the frequencies of the realizations

of effects. A particular combination of $c_{e,o}$'s indicates a particular hypothesis about which effect provoked each observed transition, and how often.

Proposition 2. *Let O be a multiset of observations and $D = \{(p_i, e_i) \mid i \in I\}$ be an effect distribution with $\forall o \in O, \exists i \in I, e_i \in o$. Then O is ϵ-fair to D if and only if the following inequality holds:*

$$\min \left(\sum_{e \in D} \quad |p_D(e) - \sum_{o \in O} c_{e,o}| \right) \leq \epsilon$$

where the minimum is taken over all combinations of nonnegative values for $c_{e,o}$'s (for all effects e, observed intervals o) which satisfy the three conditions:

$$\begin{aligned} \forall o \in O, e \in D \quad & \text{if } e \notin o \text{ then } c_{e,o} = 0 \\ \forall o \in O \quad & \sum_{e \in D} c_{e,o} = f_o \\ \forall o \in O \quad & \exists e \text{ such that } p_e \neq 0 \text{ and } c_{e,o} \neq 0 \end{aligned}$$

Proposition 3 below shows that the minimum is indeed well-defined (as the objective value of a linear program).

Example 3 (continued). The multiset O from Example 2 is 0-fair to $D_3 = \{(x_1, .4), (x_1x_2, .3), (x_2, .3)\}$, as witnessed by contributions $c_{x_1,o_1} = .4$, $c_{x_1x_2,o_1} = .1$, $c_{x_1x_2,o_2} = .2$, $c_{x_2,o_2} = .3$. Now for $D = \{(x_1x_2, .9), (\overline{x}_1, .1)\}$ we get 0.2-fairness with $c_{x_1x_2,o_1} = c_{x_1x_2,o_2} = .5$, and for $D = \{x_1, .45), (x_2, .55)\}$ we get 0.1-fairness. Finally, O is *not* fair to $D = \{x_1, 1)\}$, since no effect explains o_2.

Linearity. Fairness provides an effective way to compute most likely distributions. To see this, consider the following linear program[1] for given O and D:

$$\begin{array}{lll} \texttt{Variables:} & c_{e,o} & (\forall o \in O, e \in O) \\ \texttt{Minimize:} & \sum_{e \in 3^X} \ |p_e - \sum_{o \in O} c_{e,o}| & \\ \texttt{Subject to:} & c_{e,o} \geq 0 & (\forall o \in O, e \in O) \\ & \sum_{e \in o} c_{e,o} = f_o & (\forall o \in O) \end{array}$$

This program implements Proposition 2, but ignoring the last constraint $\forall o \in O, \exists e \in 3^X, p_e \neq 0$ and $c_{e,o} \neq 0$ (this would be a *strict* inequality constraint). Still, it can be shown that it correctly computes how fair O is to D.

Proposition 3. *Assume for all $o \in O$ there is an $e \in o$ with $p_e > 0$. Then the optimal value of the program above is the minimal ϵ such that O is ϵ-fair to D.*

We note that an arbitrary optimal solution (i.e., combination of $c_{e,o}$'s) of the program is not necessarily a witness of ϵ-fairness, in the sense that it may let some observation $o \in O$ be unexplained by any effect e with nonzero probability. Nevertheless, from such a solution a correct witness can always be straightforwardly retrieved (with the same, optimal value).

[1] The objective is indeed linear, since it consists in *minimizing* the absolute values.

Example 4 (continued). Let again $O = \{[x_1, x_1x_2] \times 5, [x_2, x_1x_2] \times 5\}$, and let $D = \{(x_1, 0), (x_2, .25), (x_1x_2, .25), (\overline{x}_1, .5)\}$. Observe that the fairness of O to D cannot be better than 1 (due to a .5-excess on effect $\overline{x}_1$). Hence the contributions $c_{x_1,o_1} = .5$, $c_{x_2,o_2} = .25$, $c_{x_1x_2,o_2} = .25$ are optimal. Nevertheless, they do not respect the constraint that there is an effect $e \in o_1$ with $p_e > 0$ and $c_{e,o_1} > 0$. Still, an optimal solution respecting this constraint can be easily retrieved by "redirecting" the contribution of x_1 to o_1 to the effect x_1x_2.

Clearly, this program is not practical, since it has exponentially many variables. We will see in Section 5 how to reduce this number in practice.

4 Variance of Sets of Observations

We now turn to the more general question of how likely it is that *several* multisets of observations $O_1, \ldots, O_q$ are all induced by the same effect distribution D. Naturally, we want to measure this by the radius of a ball centered at D and containing all O_i's. The smaller this radius, the more likely are all O_i's induced by D, and the smaller over all distributions D, the less the "variance" of O_i's. From the analysis in Section 3 it follows that fairness is the correct (unbiased) measure for defining the radius of such balls. The center of a ball containing all O_i's and with minimal radius is known as the *Chebyshev center* of the O_i's.

We have however taken care in our definition of fairness that any observed interval is accounted for by at least one effect with nonzero probability. As an unfortunate and rather surprising consequence, the center does not always exist.

Example 5. Let $O_1 = \{[x_1], [x_2] \times 2, [x_3] \times 2\}$, $O_2 = \{[x_2] \times 4, [x_4]\}$, and $O_3 = \{[x_3] \times 4, [x_4]\}$. It can be shown that O_2 and O_3 cannot be both 0.8-fair to any D, while for any $\epsilon > 0$ there is a D to which all of O_1, O_2, O_3 are $(0.8 + \epsilon)$-fair. Hence the center of O_1, O_2, O_3 does not exist (it is the argmin of an open set).

Still, we define $O_1, \ldots, O_q$ to be ϵ-*variant* (together) if there is an effect distribution D such that for all $i = 1, \ldots, q$, O_i is ϵ-fair to D. Clearly, variance inherits from fairness the property of being a correct, unbiased measure (of homogeneity), and the property of underestimating the real (hidden) variance.

It is important to note that when restricted to two multisets of observations, variance does not define a distance. Intuitively, this is simply because the witness D_{12} for the variance of O_1, O_2 needs not be the same as the witness D_{23} for O_2, O_3. In fact, the triangle inequality can even be made "very false".

Example 6. Consider the multisets $O_1 = \{[x_1, x_1x_2]\}$, $O_2 = \{[x_2, x_1x_2]\}$, and $O_3 = \{[x_2, \overline{x}_1x_2]\}$. Then O_1, O_2 are 0-variant (using $D = \{(x_1x_2, 1)\}$), and so are O_2, O_3 (using $D = \{(x_2, 1)\}$). Nevertheless, because the intervals in O_1 and O_3 are disjoint, for any distribution D the fairness of either O_1 or O_3 (or both) to D is at least 1, hence O_1, O_3 are not ϵ-variant for any $\epsilon < 1$.

The good news is that variance inherits the linear programming approach from fairness. Indeed, consider the following linear program (built on top of the one

for fairness) for computing the variance of $O_1, \ldots, O_q$, as well as a witness distribution $D = \{(p_e, e) \mid e \in 3^X\}$:

$$
\begin{array}{lll}
\texttt{Variables:} & p_e & (\forall e \in 3^X) \\
& c^i_{e,o} & (\forall i = 1, \ldots, q, o \in O_i, e \in o) \\
\\
\texttt{Minimize:} & \max_{i=1,\ldots,q} \sum_{e \in 3^X} \; |p_e - \sum_{o \in O_i} c^i_{e,o}| & \\
\\
\texttt{Subject to:} & p_e \geq 0 & (\forall e \in 3^X) \\
& c^i_{e,o} \geq 0 & (\forall i = 1, \ldots, q, o \in O_i, e \in o) \\
& \sum_{e \in 3^X} p_e = 1 & \\
& \sum_{e \in o} c^i_{e,o} = f_o & (\forall i = 1, \ldots, q, o \in O_i)
\end{array}
$$

As Example 5 shows, the program may compute a (nonattained) infimum value, but this is still sufficient for our application in Section 6.

Proposition 4. *The optimal value σ of the above linear program satisfies $\sigma \leq \epsilon$ if and only if $O_1, \ldots, O_q$ are $(\epsilon + \alpha)$-variant for all $\alpha > 0$.*

5 Restriction to Intersections of Intervals

The linear programming formulations of fairness and variance which we gave involve an exponential number of variables (enumerating all 3^n effects). We now show how to restrict this. The restriction will not give a polynomial bound on the number of variables in general, but it proves useful in practice.

The basic idea is to identify to only one (arbitrary) effect all the effects in some intersection of intervals $o_1 \cap \cdots \cap o_q$ (one per O_i).

Definition 2. *Let $O_1, \ldots, O_q$ be observations. A* maximal intersecting family *(MIF) for $O_1, \ldots, O_q$ is any maximal multiset of intervals $M = \{o_{i_1}, \ldots, o_{i_r}\}$ satisfying (for all j, $j' \neq j$): (1) $i_j \neq i_{j'}$, (2) $o_{i_j} \in O_{i_j}$, and (3) $o_{i_1} \cap \cdots \cap o_{i_r} \neq \emptyset$.*

A set of effects $E \subseteq 3^X$ is said to be sufficient *for $O_1, \ldots, O_q$ if for all MIFs $M = \{o_{i_1}, \ldots, o_{i_r}\}$, there is at least one effect $e \in E$ with $e \in o_{i_1} \cap \cdots \cap o_{i_r}$.*

Example 7. Let $o_1 = [x_1, x_1x_2]$, $o'_1 = [x_2, x_1x_2]$, $o_2 = [\emptyset, x_1x_2]$, $o'_2 = [\overline{x}_1\overline{x}_2]\}$, and $O_1 = \{o_1, o'_1\}$, $O_2 = \{o_2, o'_2\}$. The MIFs for O_1, O_2 are $\{o_1, o_2\}$ (with intersection $[x_1, x_1x_2]$), $\{o'_1, o_2\}$ (intersection $[x_2, x_1x_2]$), and $\{o'_2\}$ (intersection $[\overline{x}_1\overline{x}_2]$). Hence, for example, $\{x_1, x_2, \overline{x}_1\overline{x}_2\}$ and $\{x_1x_2, \overline{x}_1\overline{x}_2\}$ are sufficient sets of effects.

The proof of the following proposition (omitted) is simple though tedious.

Proposition 5. *Let $O_1, \ldots, O_q$ be multisets of observations, and E be a sufficient set of effects for $O_1, \ldots, O_q$. The program for computing variance has the same optimal value when effect variables p_e's range over $e \in E$ (instead of 3^X).*

Given multisets O_1, O_2, a sufficient set of effects can be easily computed by (1) intersecting all intervals in O_1 with those in O_2, yielding a set of intervals O_{12}, (2) retaining the $\subset$-minimal intervals in $(O_1 \cup O_2 \cup O_{12}) \setminus \{\emptyset\}$, and (3) choosing one effect in each remaining interval. For a greater number of intervals, Steps (1)–(2) are applied recursively first. Such computation is very efficient in practice, and indeed generates a small number of effects as compared to 3^n.

6 Application: Learning Conditions of Actions

We now show how the tools of Sections 3–5 can be applied to learning *conditional* actions. Precisely, we consider a hidden action a, consisting of formulas $c_1, \ldots, c_\ell$, each associated with an effect distribution D_i. Conditions c_i are assumed to partition the state space, so that when a is taken in s, an effect is sampled from D_i, where c_i is the unique condition satisfied by s. Such actions are known as *Probabilistic STRIPS Operators* (PSO [8]).

When $c_1, \ldots, c_\ell$ are unknown but a has independent effects on the variables, SLF-Rmax [12] learns a DBN representation with KWIK guarantees [9], provided a bound k on the number of variables used in $c_1 \cup \cdots \cup c_\ell$ is known *a priori*.

We also assume such a known bound k, and build on SLF-Rmax for addressing PSO's. Note that using synchronic arcs, SLF-RMAX could in principle deal with nonindependent (correlated) effects, but the bound k would then depend also on an *a priori* bound on the amount of correlation (maximum size of an effect). The approach we propose does not suffer from this dependence, and hence is more sample-efficient. The price we pay is that our approach is only *heuristic*.

Basic Principles. SLF-Rmax gathers statistics (observations) about the action behavior in specific families of states (using Rmax-like exploration). Precisely, for each set C of k variables and each term c on C, a multiset of observations O^c is gathered *in states satisfying* c. For instance, for $k = 2$ and $c = x_1\overline{x}_3$, O^c will contain transitions observed only in states s satisfying x_1 and $\overline{x}_3$. For technical reasons (see below), statistics are also gathered on terms over $2k$ variables.

The full approach works by first learning the conditions $c_1, \ldots, c_\ell$ of a, then learning the distribution D_i for each of them using maximum likelihood. As concerns conditions, define the following property, respective to a given measure of homogeneity μ for multisets of observations (think of our notion of variance).

Definition 3. *A term c over k variables is said to be an* ϵ-likely condition *for action a (wrt μ) if $\mu(O^{c\wedge c^1}, \ldots, O^{c\wedge c^K}) \leq \epsilon$ holds, where $c^1, \ldots, c^K$ enumerate all $\binom{n-k}{k}2^k$ terms over k variables disjoint from c.*

Example 8. Let $n = 3$ variables, $k = 1$ variable in the (hidden) conditions of a, and assume the following transitions (s, s') have been observed 10 times each:

$$(000, 000), (110, 111), (011, 000), (101, 101), (111, 111)$$

We have for instance $O^{x_1\overline{x}_3} = \{[x_3, x_1x_2x_3] \times 10\}$ and $O^{x_1x_3} = \{[\emptyset, x_1\overline{x}_2x_3] \times 10, [\emptyset, x_1x_2x_3] \times 10\}$. The term x_1 is an ϵ-likely condition because $O^{x_1x_2}$, $O^{x_1\overline{x}_2}$, $O^{x_1x_3}$, $O^{x_1\overline{x}_3}$ are homogeneous, e.g., they are all fair to $D = \{(x_3, 1)\}$. Contrastingly, $\overline{x}_3$ is not ϵ-likely because $O^{\overline{x}_2\overline{x}_3} = \{[\emptyset, \overline{x}_1\overline{x}_2\overline{x}_3] \times 10\}$ and $O^{x_2\overline{x}_3} = \{[x_3, x_1x_2x_3] \times 10\}$ do not have even one likely effect in common.

Given enough observations for each $2k$-term c, it turns out that with high probability, retaining ϵ-likely conditions (for small ϵ, as derived from Equation 1) is correct. The argument is essentially as follows [12]:

1. any correct condition $c = c_i$ is ϵ-likely; indeed, all $O^{c \wedge c^j}$'s are homogeneous because by definition they are all induced by D_i,
2. for any ϵ-likely c, by homogeneity O^c is close in particular to $O^{c \wedge c_i}$ for the *real* condition c_i, and since $O^{c \wedge c_i}$ is induced by D_i, by the triangle inequality the most likely explanations of O^c are close to the correct D_i.

Handling Ambiguity. With ambiguous effects, we propose to use variance as the measure μ in Definition 3. Because variance underestimates the radius of the L1-ball centered at the *real* distribution (Section 4), we get a complete approach.

Proposition 6. *With high confidence (as given by Equation 1), any real condition c_i of a is an ϵ-likely condition of a with respect to variance.*

However, because the soundness of SLF-Rmax (Observation 2 above) relies on the triangle inequality, our approach is not *sound* in general, that is, a condition c may be ϵ-likely while inducing an incorrect distribution of effects. Still, choosing the candidate condition c with the smallest ϵ (that is, with the smallest L1-ball centered at it) gives a promising heuristic. In particular, the worst cases where two most likely explanations are very distant from one another are not likely to be frequent in practice. We give some preliminary experimental results below.

Also observe that we could get a *sound but incomplete* approach by minimizing instead of maximizing the radius over all D's in the definition of variance. This however requires to solve a mixed integer instead of a linear program, and anyway did not prove interesting in our early experiments.

Importantly, our approach inherits the sample complexity of SLF-Rmax. A straightforward application of Equation 1 and an analysis parallel to the one of SLF-Rmax [12] shows that we need a similar number of samples: essentially, for each candidate $2k$-term $c \wedge c^j$ (there are $\binom{n}{2k}2^{2k}$ of them) we need the (polynomial) number of observations given by Equation 1 for $O^{c \wedge c^j}$ to be fair to the target distribution. Observe that $1/\epsilon$ can be used as a default estimate of the number of effects in the target distribution (ℓ in Equation 1). Overall, the sample complexity is polynomial for a fixed bound k, just as for SLF-Rmax.

Preliminary Experiments. We ran some proof-of-concept experiments on two 512-state problems ($n = 9$ variables). These only aim at demonstrating that though heuristic, our approach can work well in practice, in the sense that it learns a good model with the expected sample complexity. Hence we depict here the reward cumulated over time in an RL setting, averaged over 10 runs. For both problems on which we report here, learning can be detected to have converged at this point where the cumulated reward starts to increase continuously.

We first tested the Builder domain [3] (slightly modified by removing the "wait" action, forcing the agent to act indefinitely). The number of variables in the conditions is $k = 2$ for one action, 5 for another, and 1 for all others. Actions modify up to 4 variables together and have up to 3 different effects.

Figure 1 (left) shows the behaviour of our approach ("PSO-Rmax"). The behaviour of (unfactored) Rmax [2] is shown as a baseline; recall that Rmax

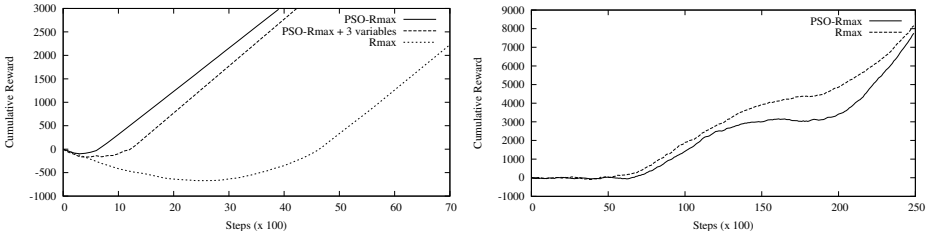

Fig. 1. Cumulative rewards for Builder (left) and Stocktrading (right).

provably converges to optimal behaviour in a number of steps polynomial in the number of states. Both algorithms were run with $m = 10$ in Equation 1 (in the corresponding equation for Rmax), as appeared to be most efficient. We have not run SLF-Rmax here because the effects of actions on variables are not independent (we leave extensions of SLF-Rmax to this setting for future work).

Each "step" on the x-axis corresponds to 100 observed transitions. As can be seen, our algorithm efficiently exploits the structure and is very fast to converge to a near-optimal policy (by exploring an average of 57 states only). In no run have we observed behaviours keeping on deviating from the optimal one, as could happen in theory since our algorithm is not guaranteed to be sound and might learn an incorrect model. Inportantly, we also tested the behaviour of our approach with 3 dummy variables added along with actions to flip them, thus artificially extending the domain to 4096 states. As Figure 1 shows, this had almost no consequence on the convergence of our approach.

We also ran some experiments on a 3×2 Stocktrading domain [12]. Here the effects are independent, so that DBNs are a compact representation ($k = 2$ variables in conditions for SLF-Rmax) while PSOs are not ($k = 6$ and $k = 7$). Experiments confirm that SLF-Rmax converges much faster than our approach. Despite this, as Figure 1 (right) shows, our algorithm behaves similarily as $R_{\max}$, which could be expected since considering all $2k$-terms over 9 variables with $k = 6$ or 7 amounts to exploiting no structure. Nevertheless, again our method always converged to optimal behaviour despite its heuristic nature.

7 Conclusion

We have studied a maximum likelihood, unbiased approach to learning actions from transitions in which effects are ambiguous. We have shown that the maximally likely actions could be computed by a linear program with exponentially many variables, but that this number could be greatly reduced in pratice. Moreover, if a known, polynomial-size set of candidate effects is learnt or known in advance, then this can be exploited directly in the program.

We have proposed an application of our study to RL, for hidden actions with correlated effects, extending SLF-Rmax [12]. Our approach is complete but not sound in general, but some proof-of-concept experiments suggest that it may

work in pratice. Nevertheless, this application is mainly illustrative, and our projects include using variance as a measure of information in techniques based on induction of decision trees [4], which prove to be more powerful in practice.

Another interesting perspective is to study information criteria able to bias the search for likely actions (e.g., MDL), and to integrate them efficiently in our linear programming approach. Finally, an important perspective in an RL context is to design exploration strategies able to help the agent disambiguate or learn the set of effects before trying to learn their relative probabilities.

Acknowledgements. This work is supported by the French National Research Agency under grant LARDONS (ANR-2010-BLAN-0215).

References

1. Boutilier, C., Dean, T., Hanks, S.: Decision-theoretic planning: Structural assumptions and computational leverage. J. Artificial Intelligence Research 11, 1–94 (1999)
2. Brafman, R.I., Tennenholtz, M.: R-max: A general polynomial time algorithm for near-optimal reinforcement learning. J. Machine Learning Research 3, 213–231 (2002)
3. Dearden, R., Boutilier, C.: Abstraction and approximate decision-theoretic planning. Artificial Intelligence 89, 219–283 (1997)
4. Degris, T., Sigaud, O., Wuillemin, P.H.: Learning the structure of factored Markov Decision Processes in reinforcement learning problems. In: Proc. International Conference on Machine Learning (ICML 2006). pp. 257–264. ACM (2006)
5. Džeroski, S., De Raedt, L., Driessens, K.: Relational reinforcement learning. Machine Learning 43, 7–52 (2001)
6. Kearns, M., Koller, D.: Efficient reinforcement learning in factored MDPs. In: Proc. 16th International Joint Conference on Artificial Intelligence (IJCAI 1999). pp. 740–474. Morgan Kaufmann (1999)
7. Kearns, M., Singh, S.: Near-optimal reinforcement learning in polynomial time. Machine Learning 49(2–3), 209–232 (2002)
8. Kushmerick, N., Hanks, S., Weld, D.S.: An algorithm for probabilistic planning. Artificial Intelligence 76, 239–286 (1995)
9. Li, L., Littman, M.L., Walsh, T.J., Strehl, A.L.: Knows what it knows: a framework for self-aware learning. Machine Learning 82, 399–443 (2011)
10. Pasula, H.M., Zettlemoyer, L.S., Kaelbling, L.P.: Learning symbolic models of stochastic domains. J. Artificial Intelligence Research 29, 309–352 (2007)
11. Rodrigues, C., Gérard, P., Rouveirol, C., Soldano, H.: Incremental learning of relational action rules. In: Proc. 9th International Conference on Machine Learning and Applications (ICMLA 2010). pp. 451–458. IEEE Computer Society (2010)
12. Strehl, A.L., Diuk, C., Littman, M.L.: Efficient structure learning in factored-state MDPs. In: Proc. 22nd AAAI Conference on Artificial Intelligence (AAAI 2007). pp. 645–650. AAAI Press (2007)
13. Walsh, T.J., Szita, I., Diuk, C., Littman, M.L.: Exploring compact reinforcement-learning representations with linear regression. In: Proc. 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009) (2009)
14. Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., Weinberger, M.J.: Inequalities for the l_1 deviation of the empirical distribution. Tech. Rep. HPL-2003-97, Hewlett-Packard Company (2003)