Planning under Risk and Knightian Uncertainty

Felipe W. Trevizan
Instituto de Matemática e Estatística
Universidade de São Paulo
Rua do Matão, 1010
São Paulo, SP, Brazil
trevisan@ime.usp.br

Fábio G. Cozman
Escola Politécnica
Universidade de São Paulo
Av. Prof. Mello Moraes, 2231
São Paulo, SP, Brazil
fgcozman@usp.br

Leliane N. de Barros
Instituto de Matemática e Estatística
Universidade de São Paulo
Rua do Matão, 1010
São Paulo, SP, Brazil
leliane@ime.usp.br

Abstract

Two noteworthy models of planning in AI are probabilistic planning (based on MDPs and its generalizations) and nondeterministic planning (mainly based on model checking). In this paper we: (1) show that probabilistic and nondeterministic planning are extremes of a rich continuum of problems that deal simultaneously with risk and (Knightian) uncertainty; (2) obtain a unifying model for these problems using imprecise MDPs; (3) derive a simplified Bellman's principle of optimality for our model; and (4) show how to adapt and analyze state-of-art algorithms such as (L)RTDP and LDFS in this unifying setup. We discuss examples and connections to various proposals for planning under (general) uncertainty.

1 Introduction

Planning problems can be classified, based on the effects of actions, in deterministic, probabilistic or nondeterministic. In this paper we are concerned with action dynamics under general forms of uncertainty; indeed, we are interested in planning under both *risk* and *Knightian uncertainty*. We show how to use these concepts to express probabilistic and nondeterministic planning (and combinations thereof) as *Markov decision processes with set-valued transitions (MDPSTs)*.

Similar generalizations of Markov decision processes (MDPs) have appeared before in research on artificial intelligence. For example, Givan *et al.* [2000] use intervals to encode a set of exact MDPs, which is used to conduct space state reduction of MDPs. Their *bounded-parameter* MDPs (BMDPs) form neither a superset nor a subset of MDPSTs. Buffet and Aberdeen [2005] use BMDPs to produce robust policies in probabilistic planning. They also show that *Real-Time Dynamic Programming (RTDP)* can be used in BMDPs. Our perspective is different: we wish to unify various strands of planning that have proven practical value, using a theory that has a behavioral basis on preferences and beliefs — otherwise, we do follow a similar path to Buffet and Aberdeen's in that we exploit RTDP in our models. Another recent work that should be mentioned is Perny *et al.*'s [2005], where transitions must only satisfy a few algebraic properties. Our models are a strict subset of their Algebraic MDPs (AMDPs); we forgo some generality because we want to employ models with solid behavioral justification, and to exploit the specific structure of combined probabilistic-nondeterministic planning. Even though general algorithms such as AMDP-based value iteration are useful theoretically, we find that a more specific approach based on Real-Time Dynamic Programming leads to encouraging results on computational complexity.

This necessarily brief review of closest literature should indicate the central motivation of this work: we strive to work with decision processes that have solid behavioral foundation *and* that can smoothly mix problems of practical significance. A similar approach has been proposed by Eiter and Lukasiewicz [2003], using nonmonotonic logic and causal semantics to define set-valued transitions in *Partial Observable* MDPs (and leaving algorithms for future work). We offer a model that assumes full observability, and we obtain algorithms and complexity analysis for our model.

The remainder of this paper is organized as follows. In Section 3 we discuss how MDPSTs capture the continuum of planning problems from "pure" probabilistic to "pure" nondeterministic planning. In Section 4 we show that MDPSTs are *Markov decision processes with imprecise probabilities (MDPIPs)*, a model that has received attention in operations research and that displays a solid foundation. We also comment on the various relationships between MDPSTs and other models in the literature. In Section 5 we show that MDPSTs lead to important simplifications of their "minimax" Bellman-style equation (we note that such simplifications are mentioned without a proof by Buffet and Aberdeen [2005] for BMDPs). We obtain interesting insights concerning computational complexity of MDPSTs and related models. Section 6 investigates algorithms that produce minimax policies for MDPSTs. Although our results yield easy variants of value and policy iteration for MDPSTs, we are interested in more efficient algorithms based on RTDP. In Section 6 we derive the conditions that must be true for RTDP to be applied. Section 7 brings a few concluding remarks.

2 Background

2.1 Varieties of planning

We start reviewing a few basic models of planning problems, attempting to unify them as much as possible as suggested by recent literature [Bonet and Geffner, 2006]:

M1 a discrete and finite state space $\mathcal{S}$,

M2 a nonempty set of initial states $S_0 \subseteq \mathcal{S}$,

M3 a goal given by a set $S_G \subseteq \mathcal{S}$,

M4 a nonempty set of actions $\mathcal{A}(s) \subseteq \mathcal{A}$ representing the actions applicable in each state s,

M5 a state transition function $F(s,a) \subseteq \mathcal{S}$ mapping state s and action $a \in \mathcal{A}(s)$ into nonempty sets of states, i.e. $|F(s,a)| \geq 1$, and

M6 a positive cost $C(s,a)$ for taking $a \in \mathcal{A}(s)$ in s.

Adapting M2, M5 and M6, one can produce:

- **Deterministic models** (DET), where the state transition function is deterministic: $|F(s,a)| = 1$. In "classical" planning, the following constraints are added: (i) $|S_0| = 1$; (ii) $S_G \neq \emptyset$; and (iii) $\forall s \in \mathcal{S}, a \in \mathcal{A}(s) : C(s,a) = 1$.
- **Nondeterministic models** (NONDET), where the actions may result in more than one successor state without preferences among them.
- **Probabilistic models** (MDPs), where actions have probabilistic consequences. Not only the function $|F(s,a)| \geq 1$ is given, but also the model includes: (MDP1) a probability distribution $P_0(\cdot)$ over S_0; and (MDP2) a probability distribution $P(\cdot|s,a)$ over $F(s,a)$ for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$.

For any of these models, we expect that a solution (e.g. a *policy*) is evaluated on its long-term costs. The cost of a solution can be evaluated in a *finite-horizon*, in which the maximum number of actions to be executed is limited to $k \in \mathbb{R}_+$. An alternative is to consider *discounted infinite-horizon*, in which the number of actions is not bounded and the cost of actions is discounted geometrically using a *discount factor* $0 < \gamma < 1$. Since it is difficult to find an appropriate k for each problem, in this paper we assume the discounted infinite-horizon framework.[1]

Due to the assumption of full observability and discounted infinite-horizon cost, a valid solution is a *stationary policy*, that is, a function π mapping states $s \in \mathcal{S}$ into actions $a \in \mathcal{A}(s)$. Bellman's principle of optimality defines the optimal cost function $V^*(s) = \min_{a \in \mathcal{A}(s)} Q_{V^*}(s,a)$ [Bellman, 1957], where:

$$Q_V(s,a) = \begin{cases} C(s,a) + \gamma V(s'),\ s' \in F(s,a) \text{ for DET,} \\ C(s,a) + \gamma \max_{s' \in F(s,a)} V(s') \text{ for NONDET, and} \\ C(s,a) + \gamma \sum_{s' \in F(s,a)} P(s'|s,a) V(s') \text{ for MDPs.} \end{cases} \quad (1)$$

This principle characterizes V^* (also called optimal value function) and induces the optimal policy for each model: $\pi^*(s) = \operatorname{argmin}_{a \in \mathcal{A}(s)} Q_{V^*}(s,a)$. The definition of $Q_V(s,a)$ clarifies the guarantees of each model. In DET, guarantees given by π^* do not depend on its execution; in NONDET guarantees are on the worst-case cost; and in MDPs guarantees are on expected cost. There are algorithms that compute the optimal policies for each one of these models, and algorithms that can be specialized to all of them [Bonet and Geffner, 2006]. However, we should emphasize that previous unifying frameworks do not intend to handle smooth "mixtures" of these planning problems. In fact, one of our goals in this paper is to provide a framework where NONDET and MDPs are the extreme points of a continuum of planning problems.

[1]Results presented here are also applicable to finite-horizon, and can be easily adapted to address partial observability.

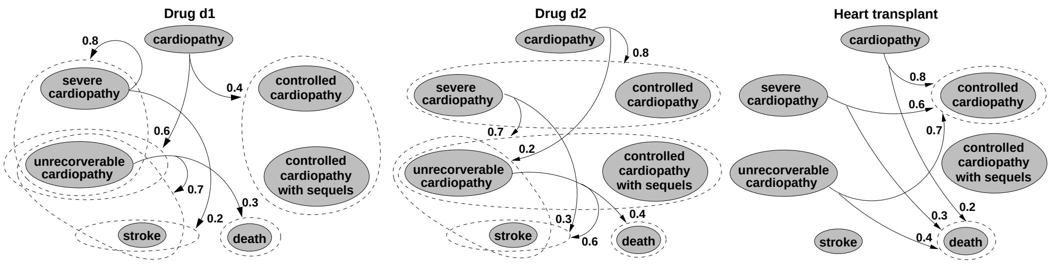
2.2 Varieties of uncertainty

Probability theory is often based on decision theory [Berger, 1985], a most appropriate scheme in the realm of planning. Thus a decision maker contemplates a set of actions, each one of which yields different rewards in different states of nature. Complete preferences over actions imply that a precise probability value is associated with each state — a situation of *risk* [Knight, 1921; Luce and Raiffa, 1957]. An obvious example of sequential decision making under "pure" risk is *probabilistic planning*. However, often preferences over actions are only partially ordered (due to incompleteness in beliefs, or lack of time/resources, or because several experts disagree), and then it is not possible to guarantee that precise probabilities represent beliefs. In those cases, a *set of probability measures* is the adequate representation for uncertainty; such sets are often referred to as *credal sets* [Levi, 1980; Kadane *et al.*, 1999; Walley, 1991]. Even though terminology is not stable, this situation is said to contain *Knightian uncertainty* (other terms are *ambiguity* or simply *uncertainty*). An extreme case is *nondeterministic planning*, where *no* probabilities are specified.[2]

Note that actual decision making is rarely restricted to either "pure" risk nor "pure" Knightian uncertainty; in fact the most realistic scenario mixes elements of both. Not surprisingly, such combinations are well studied in economics, psychology, statistics, and philosophy. We note that credal sets have raised steady interested in connection with artificial intelligence, for example in the theory of probabilistic logic [Nilsson, 1986], in Dempster-Shafer theory [Shafer, 1976], in theories of argumentation [Anrig *et al.*, 1999], and in generalizations of Bayesian networks [Cozman, 2005; Fagiuoli and Zaffalon, 1998].

The usual prescription for decision making under risk is to select an action that maximizes expected utility. In the presence of Knightian uncertainty, matters become more complex, as now a decision maker carries a set of probability measures and consequently every action is associated with an interval of expected costs [Walley, 1991]. Thus a decision maker may choose one of several criteria, such as minimaxity, maximality, E-admissibility [Troffaes, 2004]. In this paper we follow a *minimax* approach, as we are interested in actions that minimize the maximum possible expected cost; we leave other criteria for future work.

[2]The term "nondeterministic" is somewhat unfortunate as nondeterminism is often equated to probabilism; perhaps the term *planning under pure Knightian uncertainty*, although longer, would offer a better description.



Cost of actions				
State	d1	d2	HT	Noop
cardiopathy	30	20	75	---
severe cardiopathy	30	25	70	---
unrecoverable cardiopathy	40	30	70	---
controled cardiopathy	---	---	---	0
controled cardiopathy with sequels	---	---	---	2
stroke	---	---	---	85
death	---	---	---	100

Figure 1: An MDPST representing the Example 1. Dotted lines indicate each one of the reachable sets. Cost of taking actions $d1$, $d2$ and HT in the Example 1. States with "–" indicates the action is not applicable. Action `Noop` represents the persistence action for the absorbing states.

3 Markov decision processes with set-valued transitions

In this section we develop our promised synthesis of probabilistic and nondeterministic planning. We focus on the transition function; that is, on M5. Instead of taking $F(s,a) \subseteq \mathcal{S}$, we now have a set-valued $\mathbf{F}(s,a) \subseteq 2^{\mathcal{S}} \backslash \emptyset$; that is, $\mathbf{F}(s,a)$ maps each state s and action $a \in \mathcal{A}(s)$ into a *set of nonempty subsets* of $\mathcal{S}$. We refer to each set $k \in \mathbf{F}(s,a)$ as a *reachable set*. A transition from state s given action a is now associated with a probability $P(k|s,a)$; note that there is Knightian uncertainty concerning $P(s'|s,a)$ for each successor state $s' \in k$. We refer to the resulting model as a *Markov decision process with set-valued transitions (MDPSTs)*: transitions move probabilistically to reachable sets, and the probability for a particular state is not resolved by the model. In fact, there is a close connection between probabilities over $\mathbf{F}(s,a)$ and the *mass assignments* that are associated with the theory of capacities of infinite order [Shafer, 1976]; to avoid confusion between $P(k|s,a)$ and $P(s'|s,a)$, we refer to the former as mass assignments and denote them by $m(k|s,a)$.

Thus an MDPST is given by M1, M2, M3, M4, M6, MDP1,

MDPST1 a state transition function $\mathbf{F}(s,a) \subseteq 2^{\mathcal{S}} \backslash \emptyset$ mapping states s and actions $a \in \mathcal{A}(s)$ into reachable sets of $\mathcal{S}$, and

MDPST2 mass assignments $m(k|s,a)$ for all s, $a \in \mathcal{A}(s)$, and $k \in \mathbf{F}(s,a)$.

There are clearly two varieties of uncertainty in a MDPST: a probabilistic selection of a reachable set and a nondeterministic choice of a successor state from the reachable set. Another important feature of MDPSTs is that they encompass models discussed in Section 3:

- DET: There is always a single successor state: $\forall s \in \mathcal{S}, a \in \mathcal{A}(s) : |\mathbf{F}(s,a)| = 1$ and $\forall s \in \mathcal{S}, a \in \mathcal{A}(s), k \in \mathbf{F}(s,a) : |k| = 1$.
- NONDET: There is always a single reachable set, but selection within this set is left unspecified (nondeterministic): $\forall s \in \mathcal{S}, a \in \mathcal{A}(s) : |\mathbf{F}(s,a)| = 1$, and $\exists s \in \mathcal{S}, a \in \mathcal{A}(s), k \in \mathbf{F}(s,a) : |k| > 1$.
- MDPs: Selection of $k \in \mathbf{F}(s,a)$ is probabilistic and it resolves all uncertainty: $\forall s \in \mathcal{S}, a \in \mathcal{A}(s) : |\mathbf{F}(s,a)| > 1$, and $\forall s \in \mathcal{S}, a \in \mathcal{A}(s), k \in \mathbf{F}(s,a) : |k| = 1$.

Example 1 *A hospital offers three experimental treatments to cardiac patients: drug $d1$, drug $d2$ and heart transplant (HT). State $s0$ indicates patient with cardiopathy. The effects of those procedures lead to other states: severe cardiopathy ($s1$), unrecoverable cardiopathy ($s2$), cardiopathy with sequels ($s3$), controlled cardiopathy ($s4$), stroke ($s5$), and death ($s6$). There is little understanding about drugs $d1$ and $d2$, and considerable data on heart transplants. Consequently, there is "partial" nondeterminism (that is, there is Knightian uncertainty) in the way some of the actions operate. Figure 1 depicts transitions for all actions, indicating also the mass assignments and the costs. For heart transplant, we suppose that all transitions are purely probabilistic.*

4 MDPSTs, MDPIPs and BMDPs

In this section we comment on the relationship between MDPSTs and two existing models in the literature: *Markov decision processes with imprecise probabilities (MDPIPs)* [White III and Eldeib, 1994; Satia and Lave Jr, 1973] and *bounded-parameter Markov decision processes (BMDPs)* [Givan *et al.*, 2000].

An MDPIP is a Markov decision process where transitions are specified through sets of probability measures; that is, the effects of an action are modelled by a credal set $\mathcal{K}$ over the state space. An MDPIP is given by M1, M2,M3, M4, M6, MDP1 and

MDPIP1 a nonempty credal set $\mathcal{K}_s(a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$, representing probability distributions $P(s'|s,a)$ over successor states in $\mathcal{S}$.

In this paper we assume that a decision maker seeks a minimax policy (that is, she selects a policy that minimizes the maximum cost across all possible probability distributions). This adopts an implicit assumption that probabilities are selected in an adversarial manner; other interpretations for MDPIPs are possible [Troffaes, 2004]. Under the minimax interpretation, the Bellman principle of optimality is [Satia and Lave Jr, 1973]:

$$V^*(s) = \min_{a \in \mathcal{A}(s)} \max_{P(\cdot|s,a) \in \mathcal{K}_s(a)} \{C(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(\cdot|s,a) V^*(s')\}; \tag{2}$$

moreover, this equation always has a unique solution that yields the optimal stationary policy for the MDPIP. To investigate the relationship between MDPSTs and MDPIPs, the following notation is useful: when $k \in \mathbf{F}(s,a)$, we denote by

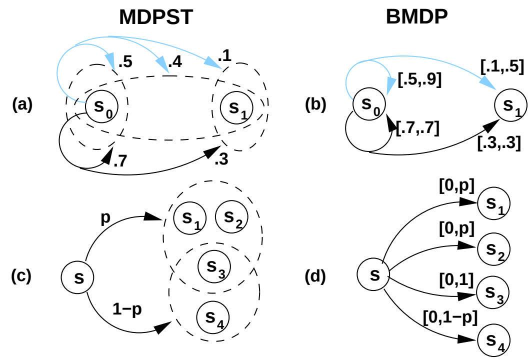


Figure 2: This figure illustrates two examples of planning under uncertainty modeled through MDPSTs and BMDPs. Example 1 is the *Heart* example from Perny *et al.* (2005). Example 2 is a simple example in which none of the models can express the problem modelled by the other one.

$\mathcal{D}(k, s, a)$ the set of states such that $k \setminus \left(\bigcup_{k' \in \mathbf{F}(s,a) \neq k} k'\right)$. Thus $\mathcal{D}(k, s, a)$ represents all nondeterministic effects of k that belong only to k. We now have:

Proposition 1 *Any MDPST $p = \langle \mathcal{S}, S_0, S_G, \mathcal{A}, \mathbf{F}, C, P_0, m\rangle$ is expressible by an MDPIP $q = \langle \mathcal{S}, S_0, S_G, \mathcal{A}, F, C, P_0, \mathcal{K}\rangle$.*

Proof (detailed in [Trevizan *et al.*, 2006]) It is enough to prove that $\forall s \in \mathcal{S}, a \in \mathcal{A}(s)$, $\mathbf{F}(s, a)$ (MDPST1) and $m(s, a)$ (MDPST2) imply $\mathcal{K}_s(a)$ (MDPIP1). First, note that MDPST2 bounds for all $s' \in \mathcal{S}$ the probability of being in state s' after applying action a in state s as follows

$$m(\{s'\}|s, a) \leq P(s'|s, a) \leq \sum_{k \in \mathbf{F}(s,a) \wedge s' \in k} m(k|s, a) \leq 1. \quad (3)$$

(To see that, use the definition of reachable sets: let $k \in \mathbf{F}(s, a)$; if $s' \notin k$, then it is not possible to select s' as a nondeterministic effect of a.)

From MDPST1 and MDPST2 it is possible to bound the sum of the probabilities of each state in a reachable set $k \in \mathbf{F}(s, a)$ and in the associated set $\mathcal{D}(k, s, a)$:

$$0 \leq \sum_{s' \in \mathcal{D}(k,s,a)} P(s'|s, a) \leq m(k|s, a) \leq \sum_{s' \in k} P(s'|s, a) \leq 1 \quad (4)$$

The set of inequalities (3) and (4) for $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$ describe a possible credal set $\mathcal{K}_s(a)$ for MDPIP1. □

Definition 1 *The MDPIP q obtained through Proposition 1 is called the **associated MDPIP** of p.*

As noted in Section 1, BMDPs are related to MDPSTs. Intuitively, BMDPs are Markov decision processes where transition probabilities and rewards are specified by intervals [Givan *et al.*, 2000]. Thus BMDPs are not comparable to MDPIPs due to possible imprecision in rewards; here we only consider those BMDPs that have real-valued rewards. Clearly these BMDPs form a strict subset of MDPIPs. The relationship between such BMDPs and MDPSTs is more complex. Figure 2.a and 2.b presents an MDPST and an BMDP that are equivalent (that is, they represent the same MDPIP). Figure 2.c and 2.d presents an MDPST that cannot be expressed

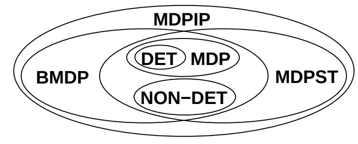


Figure 3: Relationships between models (BMDPs with precise rewards).

as an BMDP, and an BMDP that cannot be expressed as an MDPST. As a technical aside, we note that the $\mathbf{F}(s, a)$ define *Choquet capacities of infinite order*, while transitions in BMDPs define *Choquet capacities of second order* [Walley, 1991]; clearly they do not have the same representational power.

The results of this section are captured by Figure 3. In the next two sections we present our main results, where we explore properties of MDPSTs that make these models rather amenable to practical use.

5 A simplified Bellman equation for MDPSTs

We now present a substantial simplification of the Bellman principle for MDPSTs. The intuition behind the following result is this. In Equation (2), both minima and maxima are taken with respect to all combinations of actions and possible probability distributions. However, it is possible to "pull" the maximum inside the summation, so that less combinations need be considered.

Theorem 2 *For any MDPST and its associated MDPIP,* (2) *is equivalent to:*

$$V^*(s) = \min_{a \in \mathcal{A}(s)} \{C(s, a) + \gamma \sum_{k \in \mathbf{F}(s,a)} m(k|s, a) \max_{s' \in k} V^*(s')\} \quad (5)$$

Proof Define $V^*_{IP}(s)$ and $V^*_{ST}(s)$ as a shorthand for the values obtained through, respectively, (2) and (5). We want to prove that for all MDPST $p = \langle \mathcal{S}, S_0, S_G, \mathcal{A}, \mathbf{F}, C, P_0, m\rangle$, its associated MDPIP $q = \langle \mathcal{S}, S_0, S_G, \mathcal{A}, F, C, P_0, \mathcal{K}\rangle$, and $\forall s \in \mathcal{S}$, $V^*_{IP}(s) = V^*_{ST}(s)$. Due to the Proposition 1, we have that the probability measure induced by $\max_{s' \in k} V^*(s')$ for $k \in \mathbf{F}(s, a)$ in (5) is a valid choice according to $\mathcal{K}_s(a)$, therefore $\forall s \in \mathcal{S}, V^*_{ST}(s) \geq V^*_{IP}(s)$. Now, it is enough to show that $\forall s \in \mathcal{S}, V^*_{ST}(s) \leq V^*_{IP}(s)$ to conclude this proof.

For all $\hat{s} \in \mathcal{S}$, we denote by $\mathbf{F}_{\hat{s}}(s, a)$ the set of reachable sets $\{k \in \mathbf{F}(s, a) | \ \hat{s} = \mathrm{argmax}_{s' \in k} V^*_{ST}(s')\}$. The proof of $\forall s \in \mathcal{S}, V^*_{ST}(s) \leq V^*_{IP}(s)$ proceeds by contradiction as follows. For all $s \in \mathcal{S}$ and all $a \in \mathcal{A}(s)$, let $P(\cdot|s, a)$ be the probability measure chosen by the operator $\max$ in $V^*_{IP}(s)$ and suppose that $V^*_{ST}(s) > V^*_{IP}(s)$. Therefore, there is a $\overline{s} \in \mathcal{S}$ such that $\sum_{k \in \mathbf{F}_{\overline{s}}(s,a)} m(k|s, a) > P(\overline{s}|s, a)$; as $P(\cdot|s, a)$ is a probability measure, there is also a $\underline{s} \in \mathcal{S}$ s.t. $\sum_{k \in \mathbf{F}_{\underline{s}}(s,a)} m(k|s, a) < P(\underline{s}|s, a)$ and $V^*_{IP}(\overline{s}) > V^*_{IP}(\underline{s})$. Now, let $\overline{P}(\cdot|s, a)$ be a probability measure defined by: $\overline{P}(s'|s, a) = P(s'|s, a) \ \forall s' \in \mathcal{S} \setminus \{\overline{s}, \underline{s}\}$, $\overline{P}(\overline{s}|s, a) = P(\overline{s}|s, a) + \epsilon$ and $\overline{P}(\underline{s}|s, a) = P(\underline{s}|s, a) - \epsilon$, for $\epsilon > 0$. Note that $\sum_{s' \in \mathcal{S}} \overline{P}(s'|s, a) V^*_{IP} > \sum_{s' \in \mathcal{S}} P(s'|s, a) V^*_{IP}$, a contradiction by the definition of $P(\cdot|s, a)$. Thus, the rest of this proof shows that $\overline{P}(s'|s, a)$ satisfies Proposition 1.

Due to the definition of $\overline{P}(\cdot|s,a)$, we have that the left side and the right side of (3) are trivially satisfied, respectively, by $\overline{P}(\overline{s}|s,a)$ and $\overline{P}(\underline{s}|s,a)$. To treat the other case for both $\underline{s}$ and $\overline{s}$, it is sufficient to define ϵ as follows:

$$\epsilon = \min\{\sum_{k\in\mathbf{F}_{\overline{s}}(s,a)} m(k|s,a) - P(\overline{s}|s,a),\ P(\underline{s}|s,a) - \sum_{k\in\mathbf{F}_{\underline{s}}(s,a)} m(k|s,a)\}.$$

Using this definition, we have that $\epsilon > 0$ by hypothesis; since (4) gives a lower and an upper bound to the sum of $\overline{P}(\cdot|s,a)$ over, respectively, $k \in \mathbf{F}(s,a)$ and $\mathcal{D}(k,s,a) \subseteq k$. If $\{\underline{s},\overline{s}\} \in \mathcal{D}(k,s,a)$ or $\{\underline{s},\overline{s}\} \notin \mathcal{D}(k,s,a)$, then nothing changes and these bounds remain valid. There is one more case for each bound that its satisfaction is trivial too: (i) for the upper bound when $\overline{s} \notin \mathcal{D}(k,s,a)$; and (ii) for the lower bound when $\underline{s} \notin k$. A nontrivial case is for the lower bound in (4) when $\overline{s} \notin k$ and $\underline{s} \in k$. This bound still holds because $m(\{\underline{s}\}|s,a) \leq \overline{P}(\underline{s}|s,a)$ (by hypothesis) and $\sum_{s'\in k} P(s'|s,a) = \sum_{s'\in k} m(\{s'\}|s,a) + \delta \leq \sum_{s'\in k\setminus\{\underline{s}\}} m(\{s'\}|s,a) + \delta + \overline{P}(\underline{s}|s,a) \leq \sum_{s'\in k\setminus\{\underline{s}\}} P(s'|s,a) + \overline{P}(\underline{s}|s,a),\ \delta \geq 0$ (using Proposition 1 for $P(\cdot|s,a)$).

The last remaining case (to prove that Equation (4) is true for $\overline{P}(\cdot|s,a)$) happens when $\overline{s} \in \mathcal{D}(k,s,a)$ and $\underline{s} \notin \mathcal{D}(k,s,a)$ for the upper bound in (4). This case is valid because there is no $k' \in \mathbf{F}(s,a)$ such that $k' \neq k$ and $\overline{s} \in k'$ by the definition of reachable set, thus $\overline{P}(\overline{s}|s,a) \leq \sum_{k\in\mathbf{F}_{\overline{s}}(s,a)} m(k|s,a) = m(k|s,a)$ by hypothesis. If there is not a $s' \in k \setminus \{\overline{s}\}$ s.t. $P(s'|s,a) > 0$ this upper bound still holds, else, choosing s' as $\underline{s}$ will validate all the bounds. Since $\overline{P}(\cdot|s,a)$ respects Proposition 1, we get a contradiction because $\sum_{s'\in\mathcal{S}} \overline{P}(s'|s,a)V(s') > \sum_{s'\in\mathcal{S}} P(s'|s,a)V(s')$ but by hypothesis $P(\cdot|s,a) = \mathrm{argmax}_{P(\cdot|s,a)\in\mathcal{K}_s(a)} \sum_{s'\in\mathcal{S}} P(s'|s,a)V^*(s)$. Therefore, $\forall s \in \mathcal{S}, V^*_{ST}(s) \leq V^*_{IP}(s)$, what completes the proof. □

An immediate consequence of Theorem 2 is a decrease of the worst case complexity order of MDPIPs algorithms used for solving MDPSTs. Consider first one iteration of the Bellman principle of optimality for each $s \in \mathcal{S}$ (one *round*) using Equations (2). Define an upper bound of $|\mathbf{F}(s,a)|$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$ of an MDPST instance by $\overline{\mathbf{F}} = \max_{s\in\mathcal{S}}\{\max_{a\in\mathcal{A}(s)} |\mathbf{F}(s,a)|\} \leq 2^{|\mathcal{S}|}$. In the MDPIP obtained through Proposition 1, computation of $V^*(s)$ consists of solving a linear program induced by the $\max$ operator on (2). Because this linear program has $|\mathcal{S}|$ variables and its description is proportional to $\overline{\mathbf{F}}$, the worst case complexity of one round is $O(|\mathcal{A}||\mathcal{S}|^{p+1}\overline{\mathbf{F}}^q)$, for $p \geq 2$ and $q \geq 1$. The value of p and q is related to the algorithm used to solve this linear program (for instance, using the interior point algorithm [Kojima *et al.*, 1988] leads to $p = 6$ and $q = 1$, and the Karmarkar's algorithm [Karmarkar, 1984] leads to p to 3.5 and q to 3).

However, the worst case complexity for one round using Equation (5) is $O(|\mathcal{S}|^2|\mathcal{A}|\overline{\mathbf{F}})$. This is true because the probability measure that maximizes the right side of Equation (2) is represented by the choice $\max_{s'\in k} V(s')$ in Equation (5), avoiding the cost of a linear program. In the special case of an MDP modelled as an MDPST, i.e. $\forall s \in \mathcal{S}, a \in \mathcal{A}(s),\ |\mathbf{F}(s,a)| \leq |\mathcal{S}|$ and $\forall k \in \mathbf{F}(s,a),\ |k| = 1$, this worst case complexity is $O(|\mathcal{S}|^2|\mathcal{A}|)$, the same for one round using the Bellman principle for MDPs [Papadimitriou, 1994].

6 Algorithms for MDPSTs

Due to Proposition 1, every algorithm that finds the optimal policy for MDPIPs can be directly applied to MDPSTs. Instances of algorithms for MDPIP are: value iteration, policy iteration [Satia and Lave Jr, 1973], modified policy iteration [White III and Eldeib, 1994], and the algorithm to find all optimal policies presented in [Harmanec, 1999]. However, a better approach is to use Theorem 2. This proposition gives a clear path on how to adapt algorithms from the realm of MDPs — algorithms such as (L)RTDP [Bonet and Geffner, 2003] and LDFS [Bonet and Geffner, 2006]. These algorithms are defined for *Stochastic Shortest Path* problems (SSPs) [Bertsekas, 1995] (SSPs are a special case of MDPs, in which there is only one initial state (M2) and the set of goal states is nonempty (M3)). To find an optimal policy, an additional assumption is required: the goal must be reachable from every state with nonzero probability (the *reachability* assumption). For MDPSTs, this assumption can be generalized by requirement that the goal be reachable from every state with nonzero probability for all probability measures in the model. The following proposition gives a sufficient, however not necessary, condition to prove the reachability assumption for MDPSTs.

Proposition 3 *If, for all $s \in \mathcal{S}$, there exists $a \in \mathcal{A}(s)$ such that, for all $k \in \mathbf{F}(s,a)$ and all $s' \in k$, $P(s'|s,a) > 0$, then it is sufficient to prove that the reachability assumption is valid using at least one probability measure for each $s \in \mathcal{S}$ and $a \in \mathcal{A}$.*

Proof If the reachability assumption is true for a specific sequence of probability measures $\mathcal{P} = \langle P^1, P^2, \ldots . P^n\rangle$, then there exists a policy π and a *history* h, i.e. sequence of visited states and executed actions, induced by π and $\mathcal{P}$ such that $h = \langle s^0 \in S_0, \pi(s^0), s^1, \ldots, \pi(s^{n-1}), s^n \in S_G\rangle$ is minimum and $\forall s \in \mathcal{S}, P_0(s^0)\prod_{i=1}^{i\leq n} P^i(s^i|s^{i-1},\pi(s^{i-1})) > 0$. Since s^{i+1} can always be reached, because there exists an action $a \in \mathcal{A}(s^i)$ such that $P(s^{i+1}|s^i,a) > 0$, then, for any sequence of probability measures in the model, every history h' induced by π contains h, i.e., reaches $s^n \in S_G$. □

Example 2 *Consider the planning problem in the Example 1 and the cost of actions in Figure 1. We have obtained the following optimal policy for this MDPST:*

$$\pi^* = \begin{array}{c|c|c|c|c|c|c} s0 & s1 & s2 & s3 & s4 & s5 & s6 \\ \hline d1 & d2 & HT & \textit{Noop} & \textit{Noop} & \textit{Noop} & \textit{Noop} \end{array}$$

7 Conclusion

In this paper we have examined approaches to planning across many dimensions: determinism, nondeterminism, risk, uncertainty. We would like to suggest that Markov decision processes with set-valued transitions represent a remarkable entry in this space of problems. MDPSTs are quite general, as they not only capture the main existing planning models of *practical* interest, but also they can represent mixtures of these models — we have emphasized throughout the paper that MDPSTs allow one to combine nondeterminism of actions with probabilistic effects. It is particularly important to note that MDPSTs specialize rather smoothly to DET, NONDET or MDP; if an MDPST belongs to one of these cases, its solution inherits the complexity of the special case at hand. Such a "smooth" transition to special cases does not obtain if one takes the larger class of MDPIPs; the general algorithms require one to perform bilevel programming (linear programs are nested, as one linear program is needed to compute the value) and do not treat efficiently the special cases.

In fact, MDPSTs are remarkable not only because they are rather general, but because they are *not* overly general — they are sufficiently constrained that they display excellent computational properties. Consider the computation of an iteration of the Bellman equation for a state s (a round). This is an essential step both in versions of value and policy iteration and in more sophisticated algorithms such as (suitably adapted) RTDP. As discussed in Section 6, rounds in MDPSTs have much lower complexity than rounds in general MDPIPs — in essence, the simplification is the replacement of a linear program by a fractional knapsack problem.

Finally, we would like to emphasize that MDPSTs inherit the pleasant conceptual aspects of MDPIPs. They are based on solid decision theoretic principles that attempt to represent, as realistically as possible, risk *and* Knightian uncertainty. We feel that we have only scratched the surface of this space of problems; much remains to be done both on theoretical and practical fronts.

Acknowledgements

We thank FAPESP (grant 04/09568-0) and CNPq (grants 131403/05-2, 302868/04-6, 308530/03-9) for financial support and the four anonymous reviewers for the suggestions and comments.

References

[Anrig *et al.*, 1999] B. Anrig, R. Bissig, R. Haenni, J. Kohlas, and N. Lehmann. Probabilistic argumentation systems: Introduction to assumption-based modeling with ABEL. Technical Report 99-1, Institute of Informatics, University of Fribourg, 1999.

[Bellman, 1957] R. E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, New Jersey, 1957.

[Berger, 1985] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, 1985.

[Bertsekas, 1995] D.P. Bertsekas. *Dynamic programming and optimal control*. Athena Scientific Belmont, Mass, 1995.

[Bonet and Geffner, 2003] B. Bonet and H. Geffner. Labeled RTDP: Improving the convergence of real-time dynamic programming. In *Proc. of the 13th ICAPS*, pages 12–21, Trento, Italy, 2003. AAAI Press.

[Bonet and Geffner, 2006] B. Bonet and H. Geffner. Learning Depth-First Search: A unified approach to heuristic search in deterministic and non-deterministic settings, and its application to MDPs. In *Proc. of the 16th ICAPS*, 2006.

[Buffet and Aberdeen, 2005] O. Buffet and D. Aberdeen. Robust planning with (L)RTDP. In *Proc. of the 19th IJCAI*, pages 1214–1219, 2005.

[Cozman, 2005] F.G. Cozman. Graphical models for imprecise probabilities. *International Journal of Approximate Reasoning*, 39(2-3):167–184, 2005.

[Eiter and Lukasiewicz, 2003] T. Eiter and T. Lukasiewicz. Probabilistic reasoning about actions in nonmonotonic causal theories. In *Proc. of the 19th UAI*, pages 192–199, 2003.

[Fagiuoli and Zaffalon, 1998] E. Fagiuoli and M. Zaffalon. 2U: An exact interval propagation algorithm for polytrees with binary variables. *Artificial Intelligence*, 106(1):77–107, 1998.

[Givan *et al.*, 2000] R. Givan, S. M. Leach, and T. Dean. Bounded-parameter Markov decision processes. *Artificial Intelligence*, 122(1-2):71–109, 2000.

[Harmanec, 1999] D. Harmanec. A generalization of the concept of Markov decision process to imprecise probabilities. In *ISIPTA*, pages 175–182, 1999.

[Kadane *et al.*, 1999] J.B. Kadane, M.J. Schervish, and T. Seidenfeld. *Rethinking the Foundations of Statistics*. Cambridge University Press, 1999.

[Karmarkar, 1984] N. Karmarkar. A new polynomial-time algorithm for linear programming. In *Procs of the 16th annual ACM symposium on Theory of computing*, pages 302–311. ACM Press New York, NY, USA, 1984.

[Knight, 1921] F.H. Knight. *Risk, Uncertainty, and Profit.* Hart, Schaffner & Marx; Houghton Mifflin Company, Boston, 1921.

[Kojima *et al.*, 1988] M. Kojima, S. Mizuno, and A. Yoshise. A primal-dual interior point algorithm for linear programming. In *Progress in Mathematical Programming: Interior-point and related methods*, pages 29–47. Springer-Verlag, 1988.

[Levi, 1980] I. Levi. *The Enterprise of Knowledge*. MIT Press, 1980.

[Luce and Raiffa, 1957] D. Luce and H. Raiffa. *Games and Decisions*. Dover edition, Mineola, 1957.

[Nilsson, 1986] N.J. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28:71–87, 1986.

[Papadimitriou, 1994] Christos H. Papadimitriou. *Computational Complexity*. Addison-Wesley, 1994.

[Perny *et al.*, 2005] P. Perny, O. Spanjaard, and P. Weng. Algebraic Markov decision processes. In *Proc. of the 19th IJCAI*, pages 1372–1377, 2005.

[Satia and Lave Jr, 1973] J. K. Satia and R. E. Lave Jr. Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740, 1973.

[Shafer, 1976] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[Trevizan *et al.*, 2006] F. W. Trevizan, F. G. Cozman, and L. N. de Barros. Unifying nondeterministic and probabilistic planning through imprecise markov decision processes. In *Proc. of the 10th IBERAMIA/18th SBIA*, pages 502–511, 2006.

[Troffaes, 2004] M. Troffaes. Decisions making with imprecise probabilities: a short review. *The SIPTA Newsletter*, 2(1):4–7, 2004.

[Walley, 1991] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[White III and Eldeib, 1994] C. C. White III and H. K. Eldeib. Markov decision processes with imprecise transition probabilities. *Operations Research*, 42(4):739–749, 1994.