Algebraic Markov Decision Processes

Patrice Perny
LIP6 - Université Paris 6
4 Place Jussieu
75252 Paris Cedex 05, France
patrice.perny@lip6.fr

Olivier Spanjaard
LIP6 - Université Paris 6
4 Place Jussieu
75252 Paris Cedex 05, France
olivier.spanjaard@lip6.fr

Paul Weng
LIP6 - Université Paris 6
4 Place Jussieu
75252 Paris Cedex 05, France
paul.weng@lip6.fr

Abstract

In this paper, we provide an algebraic approach to Markov Decision Processes (MDPs), which allows a unified treatment of MDPs and includes many existing models (quantitative or qualitative) as particular cases. In algebraic MDPs, rewards are expressed in a semiring structure, uncertainty is represented by a decomposable plausibility measure valued on a second semiring structure, and preferences over policies are represented by Generalized Expected Utility. We recast the problem of finding an optimal policy at a finite horizon as an algebraic path problem in a decision rule graph where arcs are valued by functions, which justifies the use of the Jacobi algorithm to solve algebraic Bellman equations. In order to show the potential of this general approach, we exhibit new variations of MDPs, admitting complete or partial preference structures, as well as probabilistic or possibilistic representation of uncertainty.

1 Introduction

In the field of planning under uncertainty, the theory of Markov Decision Processes (MDPs) has received much attention as a natural framework both for modeling and solving complex structured decision problems, see e.g. [Dean *et al.*, 1993; Kaebling *et al.*, 1999]. In the standard MDP approach, the utilities of actions are given by scalar rewards supposed to be additive, uncertainty in the states of the world and in the consequences of the actions are represented with probabilities, and policies are evaluated using the expected utility (EU) model. Although these choices are natural in various practical situations, many other options are worth investigating for several reasons :

• *Rewards of actions are not necessarily scalar nor additive*. In multi-agent planning or in multicriteria MDPs, the utility of any action is given by a vector of rewards (one per agent or criterion) and actions are compared according to Pareto dominance [Wakuta, 1995]. In qualitative frameworks, rewards are valued on an ordinal scale and therefore are not additive. The sum is then replaced by the min, max or any refinement of them, depending on the context.

• *Non-probabilistic representation of uncertainty might be of interest*. In practice, it is sometimes difficult to quantify precisely the plausibility of states and consequences of actions. Assessing probabilities in such situations seems difficult. For this reason, alternative approaches based on qualitative representations of uncertainty have been proposed (see e.g. [Darwiche and Ginsberg, 1992; Dubois and Prade, 1995; Wilson, 1995]) and might be used in the context of MDPs.

• *Non-EU theories offer also interesting descriptive possibilities*. Despite the appeal of the expected utility model and its theoretical foundations [von Neumann and Morgenstern, 1947; Savage, 1954], recent developments of decision theory have shown the descriptive potential of alternative representations of preferences under uncertainty. For example, the rank dependent expected utility theory (RDEU) is a sophistication of EU theory using probability transforms to better account for actual decision making behaviors under risk [Quiggin, 1993]. Besides this extension, various alternatives to EU have been proposed for decision making in a non-probabilistic setting. Among them, let us mention the qualitative expected utility (QEU) theories proposed in [Dubois and Prade, 1995] and [Giang and Shenoy, 2001], and very recently, the generalized expected utility theory (GEU) proposed in [Chu and Halpern, 2003a; 2003b] which generalizes the notion of expectation for general plausibility measures.

Despite the diversity of models proposed for decision making under uncertainty, very few of them are used in the context of dynamic decision making. This gap can be explained by the inconsistencies entailed by the use of non-EU criteria (typically RDEU) in the dynamic context [Machina, 1989; Sarin and Wakker, 1998]. Indeed, when using a given nonlinear utility criterion at each decision stage, the Bellman principle is generally violated, so that backward induction is likely to generate a dominated policy; further, there is in general no operational way to determine an optimal policy. This simple statement largely explains the predominance of the EU model in dynamic decision making under risk.

In the last decade however, some alternative models to EU have been proved to be dynamically consistent, thus providing new possibilities for sequential decision making. This is the case of qualitative expected utility theory [Dubois and Prade, 1995] that has led to a possibilistic counterpart of MDPs [Sabbadin, 1999] with efficient algorithms adapted from backward induction and value iteration, substituting operations (+, ×) by $(\max, \min)$ in computations. In the same vein, [Littman and Szepesvári, 1996] propose a generalized version of MDPs where $\max$ and $+$ are substituted by abstract operators in Bellman equations, and [Bonet and Pearl, 2002] propose a qualitative version of MDPs. Besides these positive results, very few alternatives to standard MDPs have been investigated.

Among the diversity of possible choices for defining a reward system, a plausibility measure over events and a preference over lotteries on rewards, we need to know which combination of them can soundly be used in the context of MDPs and which algorithm should be implemented to determine an optimal policy. For this reason, we propose in this paper an algebraic generalization of the standard setting, relying on the definition of a semiring structure on rewards, a semiring structure on plausibilities of events, and a generalized expectation model as decision criterion [Chu and Halpern, 2003a]. The generalization power of semirings have been already demonstrated in AI by [Bistarelli *et al.*, 1995] in the context of constraint satisfaction problems. Within this general setting, our aim is to present a unified treatment of MDPs and to provide algorithmic solutions based on the general Jacobi algorithm (initially introduced to solve systems of linear equations).

The paper is organized as follows: in Section 2, we show by example how to recast an MDP as an optimal path problem in a decision graph. Then we introduce an algebraic framework for MDPs, relying on the definition of algebraic structures on rewards, plausibilities and expectations (Section 3). In Section 4 we justify the use of the Jacobi algorithm as a general procedure to determine an optimal policy in an Algebraic MDP (AMDP). Finally, we consider in Section 5 some particular instances of AMDPs, including new probabilistic and possibilistic MDPs using partial preferences over rewards.

2 Decision Rule Graph in MDPs

We briefly recall the main characteristics of a Markov Decision Process (MDP) [Puterman, 1994]. It can be described as a tuple (S, A, T, R) where:

- $S = \{s^1, \ldots, s^n\}$ is a finite set of states,
- $A = \{a^1, \ldots, a^m\}$ is a finite set of actions,
- $T\colon S \times A \to \mathbf{Pr}(S)$ is a transition function, giving for each state and action, a probability distribution over states (in the sequel, we write $T(s, a, s')$ for $T(s, a)(s')$),
- $R\colon S \times A \to \mathbb{R}$ is a reward function giving the immediate reward for taking a given action in a given state.

A decision rule is a function from the set of states S to the set of actions A. There are $N = m^n$ available decision rules at each step. We write $\Delta = A^S = \{\delta^1, \ldots, \delta^N\}$ the set of decision rules. A policy at step t (i.e., the t^{th}-to-last step) is a sequence of t decision rules. For a policy π and a decision rule δ, we note (δ, π) the policy which consists in applying first decision rule δ and then policy π.

	δ^1	δ^2	δ^3	δ^4
s^1	a^1	a^1	a^2	a^2
s^2	a^1	a^2	a^1	a^2

Table 1: Decision rules.

A history is a realizable sequence of successive states and actions. The accumulated reward corresponding to a history $\gamma_t = (s_t, a_t, s_{t-1}, \ldots, a_1, s_0)$ (with initial state s_t) is $R(\gamma_t) = \sum_{i=1}^{t} R(s_i, a_i)$. We denote $\Gamma_t(s)$ the set of t-step histories starting from s. For an initial state s, a t-step policy $\pi = (\delta_t, \ldots, \delta_1)$ induces a probability distribution $\Pr_t^\pi(s, \cdot)$ over histories. The t-step value of being in state s and executing policy π is given by (expected accumulated reward):

$$v_t^\pi(s) = \sum_{\gamma \in \Gamma_t(s)} \Pr_t^\pi(s, \gamma) R(\gamma)$$

Denoting v_t^π the vector whose i^{th} component is $v_t^\pi(s^i)$, any two t-step policies π, π' can be compared using the componentwise dominance relation $\geq_{\mathbb{R}^n}$ defined by:

$$v_t^\pi \geq_{\mathbb{R}^n} v_t^{\pi'} \iff \big(\forall s \in S,\ v_t^\pi(s) \geq v_t^{\pi'}(s)\big)$$

The t-step value of a policy of the form (δ^i, π) is defined recursively by $v_0^{(\delta^i, \pi)} = (0, \ldots, 0)$ and $v_t^{(\delta^i, \pi)} = f^i(v_{t-1}^\pi)$, where $f^i : \mathbb{R}^n \to \mathbb{R}^n$ is the update function which associates to any vector $x = (x_1, \ldots, x_n)$ the vector $(f_1^i(x), \ldots, f_n^i(x))$ where $f_j^i(x) = R(s^j, \delta^i(s^j)) + \sum_{k=1}^{n} T(s^j, \delta^i(s^j), s^k) x_k$.

Example 1 *Consider an MDP with* $S = \{s^1, s^2\}$, $A = \{a^1, a^2\}$, $T(s^i, a^j, s^k) = 1$ *if* $i = j = k$ *and* 0.5 *otherwise,* $R(s^1, a^1) = 8$, $R(s^1, a^2) = 7$, $R(s^2, a^1) = 12$ *and* $R(s^2, a^2) = 11$. *Thus, there are* $N = 4$ *available decision rules at each step (see Table 1). For instance, decision rule* δ^2 *consists in applying action* a^1 *in states* s^1 *and action* a^2 *in state* s^2. *In this example, the functions* f^i *are given by:*

$$\begin{aligned} f^1(x_1, x_2) &= (8 + x_1, 12 + 0.5x_1 + 0.5x_2) \\ f^2(x_1, x_2) &= (8 + x_1, 11 + x_2) \\ f^3(x_1, x_2) &= (7 + 0.5x_1 + 0.5x_2, 12 + 0.5x_1 + 0.5x_2) \\ f^4(x_1, x_2) &= (7 + 0.5x_1 + 0.5x_2, 11 + x_2) \end{aligned}$$

for all $(x_1, x_2) \in \mathbb{R}^2$.

Given a finite horizon H, the optimal policy π^* can be found thanks to the following Bellman equations:

$$\begin{aligned} &v_0^{\pi^*} = (0, \ldots, 0) \\ &v_t^{\pi^*} = \max_{i=1\ldots N} f^i(v_{t-1}^{\pi^*}) \qquad t = 1 \ldots H \end{aligned} \tag{1}$$

The solution of these equations can be reduced to a vector-weighted optimal path problem in a particular graph, with update functions (the f^i's) on the arcs allowing the propagation of policy values over nodes. Indeed, consider a graph where each node δ_t^i corresponds to decision rule δ^i at step t, and each arc of the form $(\delta_t^i, \delta_{t-1}^j)$ corresponds to a transition between decision rules. Moreover, nodes δ_1^j, $j = 1 \ldots N$ are connected to a sink denoted 0, and a source denoted H is connected to nodes δ_H^j, $j = 1 \ldots N$. Hence, any path from node δ_t^i to node 0 corresponds to a t-step policy where decision δ^i is applied first. We name that graph the *decision rule graph*. Note that the Bellman update via f^i's is nicely separable (f_j^i's can be computed independently) and therefore the vector value of a path can be obtained componentwise (state by state) as usual in classic MDP algorithms. This property will be exploited later on for Algebraic MDPs. Coming back to Example 1 and assuming that $H = 2$, there are 16 available

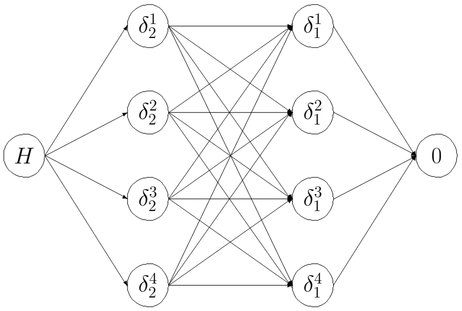

Figure 1: A decision rule graph.

policies (all possible combinations of two successive decision rules). The corresponding graph is pictured on Figure 1.

For all $i = 1 \dots n$, function f^i is associated to every arc issued from δ_t^i, $t = 1 \dots H$. Moreover, the identity function is assigned to every arc issued from H. Thus, the optimal policy can be computed thanks to backward induction on the decision rule graph, by propagating the value $(0, \dots, 0) \in \mathbb{R}^n$ from the sink 0 corresponding to the empty policy. In Example 1, backward induction leads to the labels indicated in Table 2. The optimal policy value can be recovered on node H. The optimal vector value is $(17, 23)$ and the optimal policy (recovered from bolded values in Table 2) is (δ_2^4, δ_1^1).

t	δ_t^1	δ_t^2	δ_t^3	δ_t^4
1	**(8,12)**	(8,11)	(7,12)	(7,11)
2	(16,22)	(16,23)	(17,22)	**(17,23)**

Table 2: Labels obtained during backward induction.

In the next section, we show how to generalize this approach to a wide range of MDPs. In this concern, we introduce the notion of algebraic Markov decision process.

3 Algebraic Markov Decision Process

We now define a more general setting to model rewards and uncertainty in MDPs. Our approach relies on previous works aiming at generalizing uncertainty measurement and expectation calculus. Our rewards take values in a set V and we use plausibility measures[1] [Friedman and Halpern, 1995] to model uncertainty. A plausibility measure Pl is here a function from 2^W (the set of events) to P, where W is the set of worlds, P is a set endowed with two internal operators $\oplus_P$ and $\otimes_P$ (the analogs of + and × in probability theory), a (possibly partial) order relation $\succeq_P$, and two special elements $\mathbf{0}_P$ and $\mathbf{1}_P$ such that $\mathbf{1}_P \succeq_P p \succeq_P \mathbf{0}_P$ for all $p \in P$. Furthermore, Pl verifies $\mathrm{Pl}(\emptyset) = \mathbf{0}_P$, $\mathrm{Pl}(W) = \mathbf{1}_P$ and $\mathrm{Pl}(X) \succeq_P \mathrm{Pl}(Y)$ for all X, Y such that $Y \subseteq X \subseteq W$. We assume here that Pl is decomposable, i.e. $\mathrm{Pl}(X \cup Y) = \mathrm{Pl}(X) \oplus_P \mathrm{Pl}(Y)$ for any pair of disjoint events X and Y. To combine plausibilities and rewards, we use the generalized expectation proposed by [Chu and Halpern, 2003a]. This generalized expectation is defined on an expectation domain $(V, P, \boxplus, \boxtimes)$ [2] where $\boxplus : V \times V \to V$ and $\boxtimes : P \times V \to V$ are the counterparts of $+$ and $\times$ in probabilistic expectation, and the three following requirements are satisfied: $(x \boxplus y) \boxplus z = x \boxplus (y \boxplus z)$, $x \boxplus y = y \boxplus x$, $\mathbf{1}_P \boxtimes x = x$. For any plausibility distribution Pl on V having its support in X, the generalized expectation writes: $\sum_{x \in X}^{\boxplus} \mathrm{Pl}(x) \boxtimes x$.

An Algebraic MDP (AMDP) is described as a tuple (S, A, T, R), where T and R are redefined as follows:

- $T \colon S \times A \to \mathbf{Pl}(S)$ is a transition function, where $\mathbf{Pl}(S)$ is the set of plausibility measures over S valued in P,
- $R \colon S \times A \to V$ is a reward function giving the immediate reward in V.

Consistently with the standard Markov hypothesis, the next state and the expected reward depend only on the current state and the action taken. In particular, plausibility distributions of type $T(s, a)$ are (plausibilistically) independent of the past states and actions. This plausibilistic independence refers to the notion introduced by [Friedman and Halpern, 1995] and leads to the following algebraic counterpart of the probabilistic independence property: $\mathrm{Pl}(X \cap Y) = \mathrm{Pl}(X) \otimes_P \mathrm{Pl}(Y)$ for any pair X, Y of independent events.

In this setting, rewards and plausibilities take values in two semirings. Roughly speaking, a semiring is a set endowed with two operators allowing the combination of elements (rewards or plausibilities) together. We now recall some definitions about semirings.

Definition 1 *A semiring* $(X, \oplus, \otimes, \mathbf{0}, \mathbf{1})$ *is a set* X *with two binary operations* $\oplus$ *and* $\otimes$*, such that:*
(A_1) $(X, \oplus, \mathbf{0})$ *is a commutative semigroup with* $\mathbf{0}$ *as neutral element* $(i.e., a \oplus b = b \oplus a, (a \oplus b) \oplus c = a \oplus (b \oplus c), a \oplus \mathbf{0} = a)$.
(A_2) $(X, \otimes, \mathbf{1})$ *is a semigroup with* $\mathbf{1}$ *as neutral element, and for which* $\mathbf{0}$ *is an absorbing element* $(i.e., (a \otimes b) \otimes c = a \otimes (b \otimes c), a \otimes \mathbf{1} = \mathbf{1} \otimes a = a, a \otimes \mathbf{0} = \mathbf{0} \otimes a = \mathbf{0})$.
(A_3) $\otimes$ *is distributive with respect to* $\oplus$ $(i.e., (a \oplus b) \otimes c = (a \otimes c) \oplus (b \otimes c), a \otimes (b \oplus c) = (a \otimes b) \oplus (a \otimes c))$.

A semiring is said to be *idempotent* when $(X, \oplus)$ is an idempotent commutative semigroup (i.e., a commutative semigroup such that $a \oplus a = a$). The idempotence of $\oplus$ enables to define the following canonical order relation $\succeq_X$:

$$a \succeq_X b \iff a \oplus b = a \qquad \forall a, b \in X$$

From now on, we will assume that the rewards are elements of an idempotent semiring $(V, \oplus_V, \otimes_V, \mathbf{0}_V, \mathbf{1}_V)$. Operator $\oplus_V$ is used to select the optimal values in V, whereas operator $\otimes_V$ is used to combine rewards. In classic MDPs with the total reward criterion, $\oplus_V = \max$ and $\otimes_V = +$.

Moreover the structure $(P, \oplus_P, \otimes_P, \mathbf{0}_P, \mathbf{1}_P)$ is also supposed to be a semiring. Operator $\oplus_P$ allows to combine the plausibilities of disjoint events and operator $\otimes_P$ allows to combine the plausibilities of independent events. Note that the assumption that $(P, \oplus_P, \otimes_P, \mathbf{0}_P, \mathbf{1}_P)$ is a semiring is not very restrictive since [Darwiche and Ginsberg, 1992], who use similar properties to define symbolic probability, have shown that it subsumes many representations of uncertainty, such as probability theory, possibility theory and other important calculi used in AI.

[1] This notion must not be confused with the Dempster-Shafer notion of plausibility function.

[2] In [Chu and Halpern, 2003a], an expectation domain is written $(U, P, V, \boxplus, \boxtimes)$. This structure can be simplified here since $V = U$.

Now that the general framework has been defined, we can follow the usual approach in MDPs and define a value function for policies. The accumulated reward for a history $\gamma = (s_t, a_t, s_{t-1}, \ldots, a_1, s_0)$ is $R(\gamma) = \bigotimes_{i=1}^{t} R(s_i, a_i)$. For an initial state s, a t-step policy $\pi = (\delta_t, \ldots, \delta_1)$ induces a plausibility measure $\mathrm{Pl}_t^\pi(s, \cdot)$ over histories. Such a policy will be evaluated with respect to the generalized expectation of accumulated reward, which writes:

$$v_t^\pi(s) = \sum\nolimits^{\boxplus}_{\gamma \in \Gamma_t(s)} \mathrm{Pl}_t^\pi(s, \gamma) \boxtimes R(\gamma)$$

The policies can be compared with respect to the componentwise dominance relation $\succeq_{V^n}$ between vectors in V^n:

$$x \succeq_{V^n} y \iff (\forall i = 1, \ldots, n,\ x_i \succeq_V y_i)$$

for all $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n) \in V^n$.

Most of the MDPs introduced previously in the literature are instances of our algebraic MDP:

- In standard MDPs, the underlying algebraic structure on P is $(P, \oplus_P, \otimes_P, \mathbf{0}_P, \mathbf{1}_P) = ([0,1], +, \times, 0, 1)$, and operators $\boxplus = +$ and $\boxtimes = \times$ are used to define the classic expectation operation. When rewards are defined on $(V, \oplus_V, \otimes_V, \mathbf{0}_V, \mathbf{1}_V) = (\overline{\mathbb{R}}, \max, +, -\infty, 0)$ where $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\}$, we recognize the total reward criterion. With $(\overline{\mathbb{R}}, \max, +_\gamma, -\infty, 0)$ (where $x +_\gamma y = x + \gamma y$), we recognize the weighted total reward criterion. With $(\overline{\mathbb{R}}, \max, +_H, -\infty, 0)$ where $a +_H b = \frac{1}{H} a + b$, we recognize the average reward criterion assuming that there is an initial dummy step with zero reward.

- Qualitative MDPs, introduced in [Bonet and Pearl, 2002], are AMDPs where the rewards and the plausibility measures are defined on the semiring of two-sided infinite formal series, which is a subset of the extended reals [Wilson, 1995].

In order to assign functions to the arcs in the decision rule graph, we first define $f_j^i : V^n \to V$ (the update function after applying decision rule δ^i in state s^j), for all $x \in V^n$, by:

$$f_j^i(x) = R(s^j, \delta^i(s^j)) \otimes_V \left(\sum\nolimits^{\boxplus}_{i=1..n} T(s^j, \delta^i(s^j), s^i) \boxtimes x_i \right).$$

Then, for any decision rule δ^i, we define the update function $f^i : V^n \to V^n$ which associates to any vector $x \in V^n$ the vector $(f_1^i(x), \ldots, f_n^i(x))$.

Dynamic consistency in AMDPs is guaranteed by specific properties on functions f^i. The fulfillment of these properties strongly relies on the following conditions:

(C_1) $p \boxtimes (x \oplus_V y) = (p \boxtimes x) \oplus_V (p \boxtimes y)$

(C_2) $x \boxplus (y \oplus_V z) = (x \boxplus y) \oplus_V (x \boxplus z)$

(C_3) $p \boxtimes (q \boxtimes x) = (p \otimes_P q) \boxtimes x$

(C_4) $\sum\nolimits^{\boxplus}_i p_i \boxtimes (x \otimes_V y_i) = x \otimes_V (\sum\nolimits^{\boxplus}_i p_i \boxtimes y_i)$

(C_5) $p \boxtimes (x \boxplus y) = (p \boxtimes x) \boxplus (p \boxtimes y)$

for all $p, q, p_i \in P, x, y, z, y_i \in V$.

Conditions (C_1) and (C_2) are two distributivity properties entailing a kind of additivity of $\succeq_V$ w.r.t. $\boxtimes$ and $\boxplus$ (i.e., $x \succeq_V y \Rightarrow (z * x \succeq_V z * y)$ for $* \in \{\boxtimes, \boxplus\}$). Condition (C_3) enables the reduction of lotteries. Condition (C_4) enables to isolate a sure reward in a lottery and is similar to the distributivity axiom used in [Luce, 2003]. Condition (C_5) is a distributivity condition as in classic expectation. We now establish a monotonocity result that will be used later (Prop. 3) to justify a dynamic programming approach.

Proposition 1 *If (C_1) and (C_2) hold, then f^i is non-decreasing for all $\delta^i \in \Delta$, i.e*

$$\forall x, y \in V^n,\ \big(x \succeq_{V^n} y \Rightarrow f^i(x) \succeq_{V^n} f^i(y)\big)$$

Proof. Let x, y in V^n s.t. $x \succeq_{V^n} y$. For $s_j \in S$, we have $\sum\nolimits^{\boxplus}_k T(s^j, \delta^i(s^j), s^k) \boxtimes x_k \succeq_V \sum\nolimits^{\boxplus}_k T(s^j, \delta^i(s^j), s^k) \boxtimes y_k$ by (C_1) and (C_2). Thanks to distributivity of $\otimes_V$ over $\oplus_V$, we have $f_j^i(x) \succeq_V f_j^i(y)$. Therefore, $f^i(x) \succeq_{V^n} f^i(y)$. ■

Moreover, the value of a policy π can be computed recursively thanks to the following result:

Proposition 2 *Let $\pi = (\delta^i, \pi')$ a $(t+1)$-step policy, and assume that $v_0^\pi = (\mathbf{1}_V, \ldots, \mathbf{1}_V)$. If (C_3), (C_4) and (C_5) hold, then $v_{t+1}^\pi = f^i(v_t^{\pi'})$ for all $t \geq 0$.*

Proof. Let s be a state and a denote $\delta^i(s)$. We note $\gamma_t = (s_t, a_t, \ldots, a_1, s_0)$ and $\gamma_{t+1} = (s, a, s_t, a_t, \ldots, a_1, s_0)$. We have:

$$v_{t+1}^\pi(s) = \sum^{\boxplus}_{\gamma_{t+1} \in \Gamma_{t+1}(s)} \mathrm{Pl}_{t+1}^\pi(s, \gamma_{t+1}) \boxtimes R(\gamma_{t+1})$$

$$= \sum^{\boxplus}_{s' \in S} \sum^{\boxplus}_{\gamma_t \in \Gamma_t(s')} \big(T(s,a,s') \otimes_P \mathrm{Pl}_t^{\pi'}(s', \gamma_t)\big) \boxtimes \big(R(s,a) \otimes_V R(\gamma_t)\big)$$

$$= R(s,a) \otimes_V \sum^{\boxplus}_{s' \in S} \sum^{\boxplus}_{\gamma_t \in \Gamma_t(s')} \big(T(s,a,s') \otimes_P \mathrm{Pl}_t^{\pi'}(s', \gamma_t)\big) \boxtimes R(\gamma_t)\big)$$

by (C_4)

$$= R(s,a) \otimes_V \sum^{\boxplus}_{s' \in S} T(s,a,s') \boxtimes \big(\sum^{\boxplus}_{\gamma_t \in \Gamma_t(s')} \mathrm{Pl}_t^{\pi'}(s', \gamma_t) \boxtimes R(\gamma_t)\big)$$

by (C_3) and (C_5)

$$= R(s,a) \otimes_V \sum^{\boxplus}_{s' \in S} T(s,a,s') \boxtimes v_t^{\pi'}(s')$$

Therefore, for all $s^j \in S$, $v_{t+1}^\pi(s^j) = f_j^i(v_t^{\pi'})$. ■

In order to establish the algebraic version of Bellman equations, we define the subset of maximal elements of a set with respect to an order relation $\succeq$ as:

$$\forall Y \subseteq X,\ \ M(Y, \succeq) = \{y \in Y : \forall z \in Y, \mathrm{not}(z \succ y)\}$$

Furthermore, we denote $\mathcal{P}^*(X, \succeq)$ the set $\{Y \subseteq X : Y = M(Y, \succeq)\}$. When there is no ambiguity, these sets will be denoted respectively $M(Y)$ and $\mathcal{P}^*(X)$. Besides, for any function $f : V^n \to V^n$, for all $X \in \mathcal{P}^*(V^n)$, $f(X)$ denote $\{f(x) : x \in X\}$.

We define the semiring $(\mathcal{P}^*(V^n), \oplus, \otimes, \mathbf{0}, \mathbf{1})$ where $\mathbf{0} = \{(\mathbf{0}_V, \ldots, \mathbf{0}_V)\}$, $\mathbf{1} = \{(\mathbf{1}_V, \ldots, \mathbf{1}_V)\}$, and for all $X, Y \in \mathcal{P}^*(V^n)$:

$$X \oplus Y = M(X \cup Y)$$
$$X \otimes Y = M(\{x \otimes_V y : x \in X, y \in Y\})$$

with $x \otimes_V y = (x_1 \otimes_V y_1, \ldots, x_n \otimes_V y_n)$. Hence, the algebraic generalization of Bellman equations (1) writes:

$$V_0^{\pi^*} = \mathbf{1}$$
$$V_t^{\pi^*} = \bigoplus_{i=1}^{N} f_M^i(V_{t-1}^{\pi^*}) \qquad t = 1 \ldots H$$

where $f_M^i : \mathcal{P}^*(V^n) \to \mathcal{P}^*(V^n)$ is defined by $f_M^i(X) = M(f^i(X))$ for all $X \in \mathcal{P}^*(V^n)$.

4 Generalized path algebra for AMDPs

Following the construction proposed in Section 2, a decision rule graph can be associated to an AMDP. Clearly, solving algebraic Bellman equations amounts to searching for optimal paths with respect to the canonical order associated to $\oplus$. We now show how to solve this problem.

Consider the set $\mathcal{F}$ of functions from $\mathcal{P}^*(V^n)$ to $\mathcal{P}^*(V^n)$ satisfying for all $f \in \mathcal{F}$, $X \in \mathcal{P}^*(V^n)$, $Y \in \mathcal{P}^*(V^n)$:

$$\begin{array}{lll} f(X \oplus Y) & = & f(X) \oplus f(Y) \\ f(\mathbf{0}) & = & \mathbf{0} \end{array}$$

The $\oplus$ operation on $\mathcal{P}^*(V^n)$ induces a $\oplus$ operation on $\mathcal{F}$ defined, for all $h, g \in \mathcal{F}$ and $X \in P^*(V^n)$, by:

$$(h \oplus g)(X) = h(X) \oplus g(X)$$

Let $\circ$ denote the usual composition operation between functions, *id* the identity function, and (for simplicity) $\mathbf{0}$ the constant function everywhere $\mathbf{0}$. It has been shown by [Minoux, 1977] that the algebraic structure $(\mathcal{F}, \oplus, \circ, \mathbf{0}, id)$ is a semiring. We now prove that update functions f^i_M's belong to $\mathcal{F}$.

Proposition 3 *(C_1) and (C_2) imply* $\left(\forall \delta^i \in \Delta,\, f^i_M \in \mathcal{F}\right)$.

Proof. For any δ^i, we have $f^i_M(\mathbf{0}) = \mathbf{0}$ since $\sum^{\boxplus}_i p_i \boxtimes \mathbf{0}_V = \sum^{\boxplus}_i p_i \boxtimes (\mathbf{0}_V \otimes_V \mathbf{0}_V) = \mathbf{0}_V \otimes_V (\sum^{\boxplus}_i p_i \boxtimes \mathbf{0}_V) = \mathbf{0}_V$ by (C_4) and absorption. Now, we show that $f^i_M(X \oplus Y) = f^i_M(X) \oplus f^i_M(Y), \forall X, Y \in P^*(V^n)$. Consider X, Y in $\mathcal{P}^*(V^n)$. First, we have $f^i(M(X \cup Y)) \subseteq f^i(X \cup Y)$ $(*)$. Second, we prove that $M(f^i(X \cup Y)) \subseteq f^i(M(X \cup Y))$ $(**)$. Let $z \in M(f^i(X \cup Y))$. Then it exists $w \in X \cup Y$ such that $f^i(w) = z$. If $w \in M(X \cup Y)$ then $z \in f^i(M(X \cup Y))$. If $w \notin M(X \cup Y)$ then it exists $w^* \in M(X \cup Y)$ such that $w^* \succ_V w$. As f^i is non-decreasing, we have $f^i(w^*) \succeq_V f^i(w)$. By assumption, $f^i(w) \in M(f^i(X \cup Y))$, therefore $f^i(w^*) = f^i(w)$. Finally, $z \in f^i(M(X \cup Y))$ since $z = f^i(w^*)$. By $(*)$ and $(**)$, we have $M(f^i(X \cup Y)) \subseteq f^i(M(X \cup Y)) \subseteq f^i(X \cup Y)$. Therefore $M(f^i(X \cup Y)) = M(f^i(M(X \cup Y)))$, which means, by definition, that $f^i_M(X \oplus Y) = f^i_M(X) \oplus f^i_M(Y)$. ■

Propositions 2 and 3 prove that the algebraic generalization of Jacobi algorithm solves algebraic Bellman Equations (3) and (4) [Minoux, 1977]. Thanks to the particular structure[3] of the decision rule graph, the Jacobi algorithm takes the following simple form:

```
ALGEBRAIC JACOBI ALGORITHM
1.    $V_0 \leftarrow \mathbf{1}$; $t \leftarrow 0$
2.    $Q^i_t = \mathbf{0}, \forall t = 1 \ldots H, i = 1 \ldots N$
3.    repeat
4.        $t \leftarrow t + 1$
5.        for $i = 1$ to $N$ do $Q^i_t \leftarrow f^i_M(V_{t-1})$
6.        $V_t \leftarrow \bigoplus_{i=1}^N Q^i_t$
7.    until $t = H$
```

We recognize a standard optimal path algorithm on the decision rule graph valued with functions f^i_M. When the iteration has finished, V_t is the set of generalized expected accumulated rewards at step t associated to optimal paths. The above algorithm is not efficient since lines 5 and 6 require to consider $N = m^n$ decision rules in the computation of the Bellman update. Actually, the complexity of the algorithm can be significantly improved. Indeed, remark that $\cup_{i=1}^N \{f^i_j\} = \cup_{k=1}^m \{g^k_j\}$, where $g^k_j(x) = R(s^j, a^k) \otimes_V \sum^{\boxplus}_{i=1 \ldots n} T(s^j, a^k, s^i) \boxtimes x_i$, and that $\forall v \in V^n$, $\cup_{i=1}^N \{f^i(v)\} = \cup_{k=1}^m \{g^k_1(v)\} \times .. \times \cup_{k=1}^m \{g^k_n(v)\}$. Since the maxima over a Cartesian product equals the Cartesian product of maxima over components, we have $\bigoplus_{i=1}^N f^i_M(V_{t-1}) = \bigoplus_{v \in V_{t-1}} \left(\bigoplus_{k=1}^m g^k_1(v) \times .. \times \bigoplus_{k=1}^m g^k_n(v)\right)$. Hence, lines 5 and 6 can be replaced by:

```
5.1.        $V_t \leftarrow \emptyset$
5.2.        for $v \in V_{t-1}$ do
5.3.            for $j = 1$ to $n$ do
5.4.                for $k = 1$ to $m$ do $q^k_{tj} \leftarrow g^k_j(v)$
6.1.                $V_{tj} \leftarrow \bigoplus_{k=1}^m q^k_{tj}$
6.2.            endfor
6.3.            $V_t \leftarrow V_t \oplus (V_{t1} \times .. \times V_{tn})$
6.4.        endfor
```

This algebraic counterpart of backward induction runs in polynomial time when the reward scale is completely ordered and algebraic operators can be computed in $O(1)$.

5 Examples

To illustrate the potential of the algebraic approach for MDPs, we now consider decision models that have not yet been investigated in a dynamic setting.

• *Qualitative MDPs.* In decision under possibilistic uncertainty, [Giang and Shenoy, 2001] have recently studied a new qualitative decision model (binary possibilistic utility), allowing to handle weak information about uncertainty while improving the discrimination power of qualitative utility models introduced by [Dubois and Prade, 1995]. The latter have been investigated in sequential decision problems by [Sabbadin, 1999]. We show here that the former can be exploited in sequential decision problems as well.

Possibilistic uncertainty is measured on a finite qualitative totally ordered set P, endowed with two operators $\vee$ and $\wedge$ ($\max$ and $\min$ respectively). We denote $\mathbf{0}_P$ (resp. $\mathbf{1}_P$) the least (resp. greatest) element in P. The structure $(P, \vee, \wedge, \mathbf{0}_P, \mathbf{1}_P)$ is a semiring.

In Giang and Shenoy's model, rewards are valued in an ordered scale $(U_P, \succeq)$ where $U_P = \{\langle\lambda, \mu\rangle : \lambda \in P, \mu \in P, \lambda \vee \mu = \mathbf{1}_P\}$ and $\langle\lambda, \mu\rangle \succeq \langle\lambda', \mu'\rangle \Leftrightarrow (\lambda \geq \lambda'$ and $\mu \leq \mu')$. The relevant semiring is here $(V, \oplus_V, \otimes_V, \mathbf{0}_V, \mathbf{1}_V)$ where:

$$\begin{array}{l} V = \{\langle\lambda, \mu\rangle : \lambda \in P, \mu \in P\} \\ \langle\alpha, \beta\rangle \oplus_V \langle\lambda, \mu\rangle = \langle\alpha \vee \lambda, \beta \wedge \mu\rangle \\ \langle\alpha, \beta\rangle \otimes_V \langle\lambda, \mu\rangle = \langle\alpha \wedge \lambda, \beta \vee \mu\rangle \\ \mathbf{0}_V = \langle\mathbf{0}_P, \mathbf{1}_P\rangle, \ \ \mathbf{1}_V = \langle\mathbf{1}_P, \mathbf{0}_P\rangle \end{array}$$

Note that $\oplus_V$ and $\otimes_V$ are operators $\max$ and $\min$ on U_P.

The binary possibilistic utility model is a generalized expectation with operators $\boxplus$ and $\boxtimes$ taken as componentwise $\vee$ and $\wedge$ respectively. Note that this criterion takes values in U_P. Thanks to distributivity of $\vee$ (resp. $\wedge$) over $\wedge$ (resp. $\vee$), conditions (C_1) to (C_5) hold, which proves that algebraic

[3] The graph is layered; update functions labelling arcs issued from a same node are identical and invariant from a layer to another.

backward induction yields the value of optimal policies.

• *Multicriteria MDPs.* In planning problems, different aspects (time, energy, distance...) enter into the assessment of the utility of an action, and often these aspects cannot be reduced to a single scalar reward. This shows the practical interest of investigating MDPs with multicriteria additive rewards and multicriteria comparison models. As an example, consider the following multicriteria decision model. Let Q denote the set of criteria and $\succeq_Q$ a (possibly partial) order relation over Q (reflecting the importance of criteria). Following [Grosof, 1991] and [Junker, 2002], we denote $\succ_G$ the following strict order relation between vectors:

$$x \succ_G y \Leftrightarrow \begin{cases} \exists i = 1 \ldots |Q|, x_i \neq y_i \\ \forall i : x_i \neq y_i, \big((x_i > y_i) \text{ or } (\exists j \succ_Q i,\ x_j > y_j)\big) \end{cases}$$

The interest of such a definition is to unify in a single model both the lexicographic order (when $\succeq_Q$ is a linear order) and the Pareto dominance order (when $\succeq_Q = \emptyset$).

In this case, the relevant semiring structure on rewards is defined by (neutral elements are omitted here):

$$V = \mathcal{P}^*(\overline{\mathbb{R}}^{|Q|})$$
$$X \oplus_V Y = M(X \cup Y, \succ_G)$$
$$X \otimes_V Y = M(\{x_i + y_i : x \in X, y \in Y\}, \succ_G)$$

Hence, such a reward system can easily be inserted in the classical probabilistic setting. When the generalized expectation is chosen as the componentwise (usual) expectation, conditions (C_1) to (C_5) hold, which proves that algebraic backward induction yields the value of optimal policies.

6 Conclusion

We have introduced a general approach for defining solvable MDPs in various contexts. The interest of this approach is to factorize many different positive results concerning various rewards systems, uncertainty and decision models. Once the structure on rewards, the representation of uncertainty and the decision criteria have been chosen, it is sufficient to check that we have two semirings on V and P and that conditions (C_1) through (C_5) are fulfilled to justify the use of an algorithm "à la Jacobi" to solve the problem. It is likely that this result generalizes to the infinite horizon case, provided a suitable topology is defined on the policy valuation space.

Remark that, despite its generality, our framework does not include all interesting decision theories under uncertainty. For instance, in a probabilistic setting, the RDEU model (as well as other Choquet integrals) cannot be expressed under the form of the generalized expectation used in the paper. Actually, RDEU is known as a dynamically inconsistent model [Machina, 1989; Sarin and Wakker, 1998] and it is unlikely that constructive algorithms like backward induction are appropriate. A further specific study for this class of decision models might be of interest.

References

[Bistarelli *et al.*, 1995] S. Bistarelli, U. Montanari, and F. Rossi. Constraint solving over semirings. In *Proc. of IJCAI*, 1995.

[Bonet and Pearl, 2002] B. Bonet and J. Pearl. Qualitative MDPs and POMDPs: An order-of-magnitude approximation. In *Proc. of the 18th UAI*, pages 61–68, 2002.

[Chu and Halpern, 2003a] F.C. Chu and J.Y. Halpern. Great expectations. part I: On the customizability of generalized expected utility. In *Proc. of the 18th IJCAI*, pages 291–296, 2003.

[Chu and Halpern, 2003b] F.C. Chu and J.Y. Halpern. Great expectations. part II: Generalized expected utility as a universal decision rule. In *Proc. of the 18th IJCAI*, pages 297–302, 2003.

[Darwiche and Ginsberg, 1992] A. Darwiche and M.L. Ginsberg. A symbolic generalization of probability theory. In *Proc. of the 10th AAAI*, pages 622–627, 1992.

[Dean *et al.*, 1993] T. Dean, L.P. Kaelbling, J. Kirman, and A. Nicholson. Planning with deadlines in stochastic domains. In *Proc. of the 11th AAAI*, pages 574–579, 1993.

[Dubois and Prade, 1995] D. Dubois and H. Prade. Possibility theory as a basis of qualitative decision theory. In *Proc. of the 14th IJCAI*, pages 1925–1930, 1995.

[Friedman and Halpern, 1995] N. Friedman and J. Halpern. Plausibility measures: A user's guide. In *Proc. of the 11th UAI*, pages 175–184, 1995.

[Giang and Shenoy, 2001] P.H. Giang and P.P. Shenoy. A comparison of axiomatic approaches to qualitative decision making using possibility theory. In *Proc. of the 17th UAI*, pages 162–170, 2001.

[Grosof, 1991] B. Grosof. Generalizing prioritization. In *Proc. of the 2nd KR*, pages 289–300, 1991.

[Junker, 2002] U. Junker. Preference-based search and multicriteria optimization. In *Proc. of AAAI*, pages 34–40, 2002.

[Kaebling *et al.*, 1999] L.P. Kaebling, M. Littman, and A. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1999.

[Littman and Szepesvári, 1996] M.L. Littman and C. Szepesvári. A generalized reinforcement-learning model: Convergence and applications. In *Proc. of the 13th ICML*, pages 310–318, 1996.

[Luce, 2003] R.D. Luce. Increasing increment generalizations of rank-dependent theories. *Theory and Decision*, 55:87–146, 2003.

[Machina, 1989] M.J. Machina. Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature*, 27:1622–1668, 1989.

[Minoux, 1977] M. Minoux. Generalized path algebra. In *Surveys of Mathematical Programming*, pages 359–364. Publishing House of the Hungarian Academy of Sciences, 1977.

[von Neumann and Morgenstern, 1947] J. von Neumann and O. Morgenstern. *Theory of games and economic behavior*. 2nd Ed. Princeton University Press, 1947.

[Puterman, 1994] M.L. Puterman. *Markov Decision Processes - Discrete Stochastic Dynamic Programming*. Wiley & Sons, 1994.

[Quiggin, 1993] J. Quiggin. *Generalized expected utility theory: the rank-dependant model*. Kluwer Academic Publishers, 1993.

[Sabbadin, 1999] R. Sabbadin. A possibilistic model for qualitative sequential decision problems under uncertainty in partially observable environments. In *Proc. of UAI*, pages 567–574, 1999.

[Sarin and Wakker, 1998] R. Sarin and P. Wakker. Dynamic choice and non-EU. *J. of Risk and Uncertainty*, 17:87–119, 1998.

[Savage, 1954] L.J. Savage. *The Foundations of Statistics*. J. Wiley & Sons, 1954.

[Wakuta, 1995] K. Wakuta. Vector-valued markov decision processes and the systems of linear inequalities. *Stochastic Processes and their Applications*, 56:159–169, 1995.

[Wilson, 1995] N. Wilson. An order of magnitude calculus. In *Proc. of the 11th UAI*, pages 548–555, 1995.