Extending the Bellman equation for MDPs to continuous actions and continuous time in the discounted case

Emmanuel Rachelson
ONERA-DCSD
2, avenue Edouard Belin
F-31055 Toulouse, FRANCE
emmanuel.rachelson@onera.fr

Frédérick Garcia
INRA-BIA
Chemin de Borde Rouge
F-31326 Castanet-Tolosan FRANCE
fgarcia@toulouse.inra.fr

Patrick Fabiani
ONERA-DCSD
2, avenue Edouard Belin
F-31055 Toulouse, FRANCE
patrick.fabiani@onera.fr

Abstract

Recent work on Markov Decision Processes (MDPs) covers the use of continuous variables and resources, including time. This work is usually done in a framework of bounded resources and finite temporal horizon for which a total reward criterion is often appropriate. However, most of this work considers discrete effects on continuous variables while considering continuous variables often allows for parametric (possibly continuous) quantification of actions effects. On top of that, infinite horizon MDPs often make use of discounted criterions in order to insure convergence and to account for the difference between a reward obtained now and a reward obtained later. In this paper, we build on the standard MDP framework in order to extend it to continuous time and resources and to the corresponding parametric actions. We aim at providing a framework and a sound set of hypothesis under which a classical Bellman equation holds in the discounted case, for parametric continuous actions and hybrid state spaces, including time. We illustrate our approach by applying it to the TMDP representation of (Boyan & Littman 2001).

1 Introduction

Some decision problems that deal with continuous variables imply chosing both the actions to undertake and the parameters of these actions. For example, in a robotic path planning problem, even a high level "move forward" action might need a description of its effects based on the actual length of the movement. In a medical decision problem, discrete actions may correspond to injecting different drugs but all these actions would be parameterized by the injection's volume. Lastly, in time dependent problems, one often needs a "wait" action which really is a parametric "wait for duration τ" action.

Considering continuous planning domains often leads to considering continuous effects on the state and continuous parameters for actions (namely continuous parametric actions). In this paper, we present a way of introducing continuous actions in Markov Decision Processes (MDPs) and extend the standard Bellman equation for a discounted criterion on these generalized problems. We call "XMDPs" the parametric action MDPs for notation convenience.

The MDP framework has become a popular framework for representing decision problems under uncertainty. Concerning the type of problems we presented above, consequent research has been done in the field of planning under uncertainty with continuous resources (Bresina *et al.* 2002), planning with time-dependencies (Boyan & Littman 2001; Younes & Simmons 2004) and taking into account continuous and discrete variables in the state space (Guestrin, Hauskrecht, & Kveton 2004; Hauskrecht & Kveton 2006; Feng *et al.* 2004). However, little work has been undertaken concerning continuous actions, even though the problem of dealing with parametric actions arose in the conclusion of (Feng *et al.* 2004). While considering a continuous action space may not present immediate physical meaning, using parametric actions makes sense in real world domains: our research deals with extending the MDP framework to these actions, especially in the case of random durative actions.

This work relates to the close link between control theory and decision theory (Bertsekas & Shreve 1996). As in optimal control, we deal with continuous control spaces, but the analogy doesn't go further: contrary to control problems, our framework deals with sequential decision problems with successive decision epochs defined on a continuous action space, whereas control theory deals with continuous control of a state variable. Therefore, our problem remains a discrete event control problem defined over continuous variables. On top of that we deal with random decision epochs, which are not taken into account in classical MDP models. This key feature spans the complexity of the proofs provided here but also allows one to plan in unstationary environments with a continuous observable time variable as we will see in section 3.

We recall the basics of MDPs and of Bellman's optimality equation in section 2. We then introduce the extended formalism (XMDPs) we use in order to describe hybrid state spaces and parametric actions together with the discounted reward criterion in section 3. Section 4 extends the standard Bellman equation for MDPs to this extended XMDP framework. We finally illustrate this equation on the TMDP model in section 5.

2 MDPs and Bellman equation

We assume standard MDP notations (Puterman 1994) and describe a classical MDP as a 5-tuple $< S, A, P, r, T >$

where S is a discrete, countable state space, A is a discrete, countable action space, P is a transition function mapping transitions (s', a, s) to probability values, r is a reward function mapping pairs of action and state to rewards or costs and T is a set of decision epochs. In general infinite horizon MDPs, T is isomorphic to $\mathbb{N}$.

An MDP is a sequential stochastic control problem where one tries to optimize a given criterion by choosing the actions to undertake. These actions are provided as decision rules. Decision rule d_δ maps states to the actions to be performed at decision epoch δ. (Puterman 1994) proves that for an infinite horizon problem, there is an optimal control policy that is Markovian, ie. that only relies on knowledge of the current state. Such a policy π is thus a mapping from states to actions. The optimization criterions used in the infinite horizon case are often the discounted reward criterion, the total reward criterion or the average reward criterion. We focus on the first one here. The discounted criterion for standard infinite horizon MDPs evaluates the sums of expected future rewards, each reward being discounted by a factor γ. This factor insures the convergence of the series and can be interpreted as a probability of non-failure between two decision epochs.

$$V_\gamma^\pi(s) = E\left(\sum_{\delta=0}^{\infty} \gamma^\delta r(s_\delta, \pi(s_\delta)\right) \quad (1)$$

One can evaluate a policy with regard to the discounted reward criterion. The value V^π of policy π obeys the following equation:

$$V^\pi(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} P(s'|s, \pi(s)) V^\pi(s') \quad (2)$$

The Bellman equation (or dynamic programming equation) is an implicit equation yielding the optimal value function for a given MDP and criterion. This optimal value function is the value of the optimal policy and therefore, finding the optimal value function V^* immediately yields the optimal policy π^*. The Bellman equation for discounted MDPs is:

$$V^*(s) = \max_{a \in A} \left[r(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) V^*(s') \right] \quad (3)$$

3 Hybrid state spaces and parametric actions

3.1 Model definition

In order to illustrate the following definitions on a simple example, we propose the game presented in figure 1. In this game, the goal is to bring the ball form the start box to the finish box. Unfortunately, the problem depends on a continuous time variable because the boxes' floors retract at known dates and because actions durations are uncertain and real-valued. At each decision epoch, the player has five possible actions: he can either push the ball in one of the four directions or he can wait for a certain duration in order to reach a better configuration. Finally the "push" actions are uncertain and the ball can end up in the wrong box. This problem has an hybrid state space composed of discrete variables - the ball's position - and continuous ones - the current date.

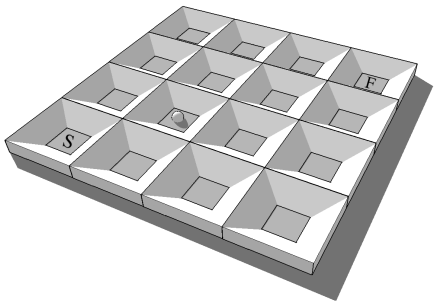

Figure 1: Illustrative example

It also has four non-parametric actions - the "push" actions - and one parametric action - the "wait" action. We are therefore trying to find a policy on a stochastic process with continuous and discrete variables and parametric actions (with real valued parameters). Keeping this example in mind, we introduce the notion of parametric MDP:

Definition (XMDP)**.** *A parametric action MDP is a tuple* $< S, A(X), p, r, T >$ *where:*
S is a Borel state space which can describe continuous or discrete state variables. A is an action space describing a finite set of actions $a_i(x)$ where x is a vector of parameters taking its values in X. Therefore, the action space of our problem is a continuous action space, factored by the different actions an agent can undertake. p is a probability density transition function $p(s'|s, a(x))$. r is a reward function $r(s, a(x))$. T is a set of timed decision epochs.

As we will see in the next sections, the time variable has a special importance regarding the discounted reward criterion. If we consider variable durations and non-integer decision epochs then we have to make the time variable observable, ie. we need to include it in the state space. In order to deal with the more general case, we will consider a real-valued time variable t and will write the state (s, t) in order to emphasize the specificity of this variable in the discounted case.

Note that for discrete variables, the $p()$ function of the XMDP is a discrete probability distribution function and that writing integrals over $p()$ is equivalent to writing a sum over the discrete variables.

On top of the definitions above, we make the following hypothesis which will prove themselves necessary in the proofs below:

- action durations are all positive and non-zero.
- the reward model is upper semi-continuous

Lastly, as previously, we will write δ the number of the current decision epoch, and, consequently, t_δ the time at which decision epoch δ occurs.

3.2 Policies and criterion

We define the *decision rule* at decision epoch δ as the mapping from states to actions: $d_\delta : \begin{cases} S \times \mathbb{R} & \rightarrow & A \times X \\ s, t & \mapsto & a, x \end{cases}$.

d_δ specifies the parametric action to undertake in state (s,t) at decision epoch δ. A *policy* is defined as a set of decision rules (one for each δ) and we consider, as in (Puterman 1994), the set $\mathcal{D}$ of stationary (with regard to δ) markovian deterministic policies.

In order to find optimal policies for our problem, we need to define a criterion. The SMDP model (Puterman 1994), proposes an extension of MDPs to continuous time, stationary models. The SMDP model is described with discrete state and action spaces S and A, a transition probability function $P(s'|s,a)$ and a duration probability function $F(t|s,a)$. The discounted criterion for SMDPs integrates the expected reward over all possible transition durations. Similarly to the discounted criterion for SMDPs, we introduce the discounted criterion for XMDPs as the expected sum of the successive discounted rewards, with regard to the application of policy π starting in state (s,t):

$$V_\gamma^\pi(s,t) = E_{(s,t)}^\pi \left\{ \sum_{\delta=0}^{\infty} \gamma^{t_\delta - t} r_\pi(s_\delta, t_\delta) \right\} \tag{4}$$

In order to make sure this series has a finite limit, our model introduces three more hypothesis:

- $|r\left((s,t),a(x)\right)|$ is bounded by M,
- $\forall \delta \in T, \quad t_{\delta+1} - t_\delta \geq \alpha > 0$, where α is the smallest possible duration of an action,
- $\gamma < 1$.

The discount factor γ^t insures the convergence of the series. Physically, it can be seen as a probability of still being functional after time t. With these hypothesis, it is easy to see that for all $(s,t) \in S \times \mathbb{R}$:

$$|V_\gamma^\pi(s,t)| < \frac{M}{1-\gamma^\alpha} \tag{5}$$

We will admit here that the set $\mathcal{V}$ of value functions (functions from $S \times \mathbb{R}$ to $\mathbb{R}$) is a complete metrizable space for the supremum norm $\|V\|_\infty = \sup_{(s,t)\in S\times\mathbb{R}} V(s,t)$.

An optimal policy is then defined as a policy π^* which verifies $V_\gamma^{\pi^*} = \sup_{\pi\in\mathcal{D}} V_\gamma^\pi$. The existence of such a policy is proven using the hypothesis of upper semi-continuity on the reward model which guarantees that there exists a parameter that reaches the $\sup$ of the reward function (such a proof was immediate in the classical MDP model because the action space was countable).

From here on we will omit the γ index on V .

On this basis, we look for a way of characterizing the optimal strategy. In a standard MDP resolution often uses dynamic programming (Bellman 1957) or linear programming (Guestrin, Hauskrecht, & Kveton 2004) techniques on the optimality equations. Here we concentrate on these optimality equations and prove the existence of a Bellman equation for the discounted criterion we have introduced.

4 Extending the Bellman equation

We introduce the policy evaluation operator L^π. Then we redefine the Bellman operator L for XMDPs and we prove that V^* is the unique solution to $V = LV$. Dealing with random decision times and parametric actions invalidates the proof of (Puterman 1994), we adapt it and emphasize the differences in section 4.2.

4.1 Policy evaluation

Definition (L^π operator)**.** *The policy evaluation operator L^π maps any element V of $\mathcal{V}$ to the value function:*

$$L^\pi V(s,t) = r(s,t,\pi(s,t)) + \int_{\substack{t'\in\mathbb{R}\\ s'\in S}} \gamma^{t'-t} p(s',t'|s,t,\pi(s,t)) V(s',t') ds' dt' \tag{6}$$

We note that for non-parametric actions and discrete state spaces, $p()$ is a discrete probability density function, the integrals turn to sums and the L^π operator above turns to the classical L^π operator for standard MDPs. This operator represents the one-step gain if we apply π and then get V. We now prove that this operator can be used to evaluate policies.

Proposition (Policy evaluation)**.** *Let π be a policy in $\mathcal{D}$. Then $V = V^\pi$ is the only solution of $L^\pi V = V$.*

Proof. In the following proofs $E_{a,b,c}^\pi$ denotes the expectation with respect to π, knowing the values of the random variables a, b and c. Namely, $E_{a,b,c}^\pi(f(a,b,c,d,e))$ is the expectation calculated with regard to d and e, and is therefore a function of a, b and c.

Our starting point is $(s_0,t_0) = (s,t)$:

$$\begin{aligned} V^\pi(s,t) &= E_{s_0,t_0}^\pi \left\{ \sum_{\delta=0}^{\infty} \gamma^{t_\delta - t} r_\pi(s_\delta,t_\delta) \right\} \\ &= r_\pi(s,t) + E_{s_0,t_0}^\pi \left\{ \sum_{\delta=1}^{\infty} \gamma^{t_\delta - t} r_\pi(s_\delta,t_\delta) \right\} \\ &= r_\pi(s,t) + E_{s_0,t_0}^\pi \left\{ E_{\substack{s_0,t_0\\ s_1,t_1}}^\pi \left(\sum_{\delta=1}^{\infty} \gamma^{t_\delta - t} r_\pi(s_\delta,t_\delta) \right) \right\} \end{aligned}$$

The inner mathematical expectation deals with random variables $(s_i,t_i)_{i=2\ldots\infty}$, the outer one deals with the remaining variables (s_1,t_1). We expand the outer expected value with $(s_1,t_1) = (s',t')$:

$$V^\pi(s,t) = r_\pi(s,t) + \int_{\substack{t'\in\mathbb{R}\\ s'\in S}} E_{\substack{s_0,t_0\\ s_1,t_1}}^\pi \left(\sum_{\delta=1}^{\infty} \gamma^{t_\delta - t} r_\pi(s_\delta,t_\delta) \right) \cdot p_\pi(s',t'|s,t) ds' dt'$$

$$V^\pi(s,t) = r_\pi(s,t) + \int_{\substack{t'\in\mathbb{R}\\ s'\in S}} \gamma^{t'-t} p_\pi(s',t'|s,t) \cdot E_{\substack{s_0,t_0\\ s_1,t_1}}^\pi \left(\sum_{\delta=1}^{\infty} \gamma^{t_\delta - t'} r_\pi(s_\delta,t_\delta) \right) ds' dt'$$

The expression inside the $E^{\pi}_{s_0,t_0,s_1,t_1}()$ deals with random variables (s_i, t_i) for $i \geq 2$. Because of the Markov property on the $p()$ probabilities, this expectation only depends on the (s_1, t_1) variables and thus:

$$E^{\pi}_{\substack{s_0,t_0\\ s_1,t_1}}\left(\sum_{\delta=1}^{\infty}\gamma^{t_\delta - t}r_\pi(s_\delta,t_\delta)\right) = V^\pi(s',t')$$

And we have:

$$V^\pi(s,t) = L^\pi V^\pi(s,t) \tag{7}$$

The solution is unique because L^π is a contraction mapping on $\mathcal{V}$ and we can use the Banach fixed point theorem (the proof of L^π being a contraction mapping is similar to the one we give for the L operator in the next section). □

4.2 Bellman operator

Introducing the L^π operator is the first step towards defining the dynamic programming operator L.

Definition (L operator)**.** *The Bellman dynamic programming operator L maps any element V of $\mathcal{V}$ to the value function: $LV = \sup_{\pi\in\mathcal{D}}\{L^\pi V\}$*

$$LV(s,t) = \sup_{\pi\in\mathcal{D}}\left\{r_\pi(s,t) + \int_{\substack{t'\in\mathbb{R}\\ s'\in S}}\gamma^{t'-t}p_\pi(s',t'|s,t)V(s',t')ds'dt'\right\} \tag{8}$$

This operator represents the one-step optimization of the current policy. We now prove that L defines the optimality equation equivalent to the discounted criterion (equation 4).

One can note that the upper semi-continuity of the rewards with regard to the parameter guarantees that such a supremum exists in equation 8, thus justifying this hypothesis which wasn't necessary in (Puterman 1994) because the action space was countable.

Proposition (Bellman equation)**.** *For an XMDP with a discounted criterion, the optimal value function is the unique solution of the Bellman equation $V = LV$.*

Proof. The proofs adapts (Puterman 1994) to the XMDP hypothesis. Namely, hybrid state space, parametric action space and semi-continuous action rewards. Our reasoning takes three steps:

1. We first prove that if $V \geq LV$ then $V \geq V^*$,
2. Then, we similarly prove that if $V \leq LV$ then $V \leq V^*$,
3. Lastly, we prove that there exists a unique solution to $V = LV$.

Suppose that we have a V such that $V \geq LV$. Therefore, with π a policy in $\mathcal{D}$, we have: $V \geq \sup_{\pi\in\mathcal{D}}\{L^\pi V\} \geq L^\pi V$.

Since L^π is positive, we have, recursively: $V \geq L^\pi V \geq L^\pi L^\pi V \ldots \geq L^{\pi(n+1)}V$. We want to find a $N \in \mathbb{N}$ such that $\forall n \geq N,\quad L^{\pi(n+1)}V - V \geq 0$.

$L^{\pi(n+1)}V$ corresponds to applying policy π for $n+1$ steps and then getting reward V.

$$L^{\pi(n+1)}V = r_\pi(s_0,t_0) + E^{\pi}_{s_0,t_0}\Bigg(\gamma^{t_1-t_0}r_\pi(s_1,t_1) + E^{\pi}_{s_1,t_1}\Bigg(\gamma^{t_2-t_0}r_\pi(s_2,t_2) + E^{\pi}_{s_2,t_2}\Bigg(\ldots + E^{\pi}_{s_{n-1},t_{n-1}}\Bigg(\gamma^{t_n-t_0}r_\pi(s_n,t_n) + E^{\pi}_{s_n,t_n}\Bigg(\gamma^{t_{n+1}-t_0}V(s_{n+1},t_{n+1})\Bigg)\Bigg)\ldots\Bigg)\Bigg)\Bigg)$$

$$V^\pi = r_\pi(s_0,t_0) + E^{\pi}_{s_0,t_0}\Bigg(\gamma^{t_1-t_0}r_\pi(s_1,t_1) + E^{\pi}_{s_1,t_1}\Bigg(\gamma^{t_2-t_0}r_\pi(s_2,t_2) + E^{\pi}_{s_2,t_2}\Bigg(\ldots + E^{\pi}_{s_{n-1},t_{n-1}}\Bigg(\gamma^{t_n-t_0}r_\pi(s_n,t_n) + E^{\pi}_{s_n,t_n}\Bigg(\sum_{\delta=n+1}^{\infty}\gamma^{t_\delta-t_0}r_\pi(s_\delta,t_\delta)\Bigg)\Bigg)\ldots\Bigg)\Bigg)\Bigg)$$

When writing $L^{\pi(n+1)}V - V^\pi$ one can merge the two expressions above in one big expectation over all random variables $(s_i,t_i)_{i=0\ldots\infty}$. Then all the first terms cancel each other and we can write:

$$L^{\pi(n+1)}V - V^\pi = E^{\pi}_{\substack{(s_i,t_i)\\ i=0\ldots n}}\Bigg(\gamma^{t_{n+1}-t_0}V(s_{n+1},t_{n+1}) - \sum_{\delta=n+1}^{\infty}\gamma^{t_\delta-t_0}r_\pi(s_\delta,t_\delta)\Bigg)$$

and thus:

$$L^{\pi(n+1)}V - V^\pi = E^{\pi}_{\substack{(s_i,t_i)\\ i=0\ldots n}}\Bigg(\gamma^{t_{n+1}-t_0}V(s_{n+1},t_{n+1})\Bigg) - E^{\pi}_{\substack{(s_i,t_i)\\ i=0\ldots n}}\Bigg(\sum_{\delta=n+1}^{\infty}\gamma^{t_\delta-t_0}r_\pi(s_\delta,t_\delta)\Bigg)$$

We write: $L^{\pi(n+1)}V - V^\pi = q_n - r_n$.

Since $\gamma < 1$, $r()$ bounded by M and for all $n \in \mathbb{N}$, $t_{n+1} - t_n \geq \alpha > 0$, we know $\|V\|$ is bounded (equation 5) and we have:
$E^{\pi}_{\substack{(s_i,t_i)\\ i=0\ldots n}}\left(\gamma^{t_{n+1}-t_0}V(s_{n+1},t_{n+1})\right) \leq \gamma^{(n+1)\alpha}\|V\|$.

So we can write $\lim_{n\to\infty} q_n = 0$.

On the other hand, r_n is the remainder of a convergent series. Thus we have: $\lim_{n\to\infty} r_n = 0$.

So $\lim_{n\to\infty} L^{\pi(n+1)}V - V^\pi = 0$.

We had $V \geq L^{\pi(n+1)}V$, so $V - V^\pi \geq L^{\pi(n+1)}V - V^\pi$. The left hand side expression doesn't depend on n and since the right hand side expression's limit is zero, we can write: $V - V^\pi \geq 0$.

Since this is true for any $\pi \in \mathcal{D}$, it is true for π^* and:

$$V \geq LV \Rightarrow V \geq V^*$$

Following a similar reasoning we can show that if $\pi' = \arg\sup_{\pi\in\mathcal{D}} L^\pi V$ and $V \leq LV$, then $V \leq L^{\pi'(n+1)}V$. Therefore $V - V^{\pi'} \leq L^{\pi'(n+1)}V - V^{\pi'}$ and so $V - V^{\pi'} \leq 0$. Since $V^{\pi'} \leq V^*$, we have:

$$V \leq LV \Rightarrow V \leq V^*$$

The two previous assertions show that if a solution to $V = LV$ exists, then this solution is equal to V^*.

In order to finish proving the proposition, we need to prove that there always is a solution to $V = LV$.

$\mathcal{V}$ is a metrizable space, complete for the supremum norm $\|V\|_\infty = \sup_{(s,t)\in S\times\mathbb{R}} V(s,t)$ (Bertsekas & Shreve 1996). If we show that L is a contraction mapping in $\mathcal{V}$, then we will be able to apply Banach fixed point theorem.

Let U and V be two elements of $\mathcal{V}$ with $LV \geq LU$.

Let (a^*, x^*) be the solution of:

$$a^*(x^*) = \underset{a(x)\in A(\mathbb{R})}{argsup} \left\{ r(s,t,a(x)) + \int_{\substack{t'\in\mathbb{R}\\ s'\in S}} \gamma^{t'-t} p_{a(x)}(s',t'|s,t)V(s',t')ds'dt' \right\}$$

(a^*, x^*) exists because of the upper semi-continuity hypothesis of the reward function which guarantees that, even at a discontinuity point in the reward function, the upper value is reachable.

For all (s,t) in $S \times \mathbb{R}$, we have:
$|LV(s,t) - LU(s,t)| = LV(s,t) - LU(s,t)$

$$LV(s,t) - LU(s,t) \leq r(s,t,a^*(x^*)) + \int_{\substack{t'\in\mathbb{R}\\ s'\in S}} \gamma^{t'-t} p_{a^*(x^*)}(s',t'|s,t)V(s',t')ds'dt' - r(s,t,a^*(x^*)) - \int_{\substack{t'\in\mathbb{R}\\ s'\in S}} \gamma^{t'-t} p_{a^*(x^*)}(s',t'|s,t)U(s',t')ds'dt'$$

Which yields:

$$LV(s,t) - LU(s,t) \leq \int_{\substack{t'\in\mathbb{R}\\ s'\in S}} \gamma^{t'-t} p_{a^*(x^*)}(s',t'|s,t)\cdot (V(s',t') - U(s',t'))\, ds'dt'$$

We have: $\begin{cases} V(s,t) - U(s,t) \leq \|V - U\| \\ t' - t \geq \alpha > 0 \\ p(s',t'|s,t,a(x)) \leq 1 \\ \gamma < 1 \end{cases}$, so we can write:

$$LV(s,t) - LU(s,t) \leq \|V - U\| \cdot \gamma^\alpha$$

and thus:

$$\|LV - LU\| \leq \|V - U\| \cdot \gamma^\alpha$$

Since $\gamma^\alpha < 1$, this proves L is a contraction mapping on $\mathcal{V}$. Banach fixed point theorem then tells us that there exists a fixed point $V' \in \mathcal{V}$ to the L operator such that $V' = LV'$.

The previous results allow us to conclude that under the general hypothesis mentioned above, the equation $LV = V$ has a unique solution and this solution is equal to V^*, the optimal value function with regard to the discounted criterion.

$$LV = V \Rightarrow V = V^* \tag{9}$$

□

One can rewrite the Bellman equation in the following way, making it more suitable for dynamic programming algorithms such as value or policy iteration:

$$LV(s,t) = \max_{a\in A} \sup_{x\in X} \left\{ r(s,t,a(x)) + \int_{\substack{t'\in\mathbb{R}\\ s'\in S}} \gamma^{t'-t} p(s',t'|s,t,a(x))V(s',t')ds'dt' \right\} \tag{10}$$

Using this formulation, we alternate an optimization on x of the value of each action which yields the optimal value of the parameter per action and a choice among the (discrete) set of possible actions (with their optimal parameter).

For a brief example giving the flavor of the next section, we can imagine a problem with a single continuous time variable factoring a discrete state space and a single continuous duration parameter τ affecting only the "wait" action. Then equation 10 can be straightforwardly implemented as a two-step value iteration algorithm. The first step calculates the optimal value of τ for any action that depends on it. The second step is a maximization step over all actions with their optimal parameter. This naive example shows the difficulties we can expect form designing algorithms to solve XMDPs. These difficulties deal with representing the continuous functions of the model's dynamics, solving the integrals in the Bellman equation and representing the continuous part of the policy. These prob-

lems have been encountered more generally when dealing with continuous variables in MDPs and various solutions for representing / approximating value functions have been proposed in (Boyan & Littman 2001; Liu & Koenig 2006; Li & Littman 2005; Marecki, Topol, & Tambe 2006; Hauskrecht & Kveton 2006) while there have been some attempts at dealing with continuous actions in reinforcement learning ((Hasselt & Wiering 2007)).

One can notice that if the state space is discrete, all probability density functions are discrete and integrals turn to sums. If the parameter space is discrete as well, by reindexing the actions in the action space, the $\sup$ operator turns to a $\max$ and the above Bellman equation (equation 9) is the standard dynamic programming equation characterizing the solutions of classical MDPs. Therefore we can conclude that the XMDP model and its optimality equation includes and generalizes the results presented in (Puterman 1994) for standard MDPs.

5 Illustration on the TMDP model

The TMDP model was introduced in (Boyan & Littman 2001). It is a modification of a standard MDP in order to take time-dependent dynamics into account in the transition and reward model. Even though they provide optimality equations, it is unclear to determine which criterion the authors actually optimized. We prove that the optimized criterion really was a total reward criterion similar to the discounted criterion we introduced above. If we suppose the existence of absorbing reachable states in the model, then we can relax the $\gamma < 1$ hypothesis and - with $\gamma = 1$ - introduce a total reward criterion for XMDPs.

A TMDP is composed of a discrete state space S factored by a continuous time variable t, a set A of discrete actions and a set M of action outcomes, each outcome being represented by the tuple $< s', T_\mu, P_\mu >$ with s' the resulting state, T_μ a flag indicating whether the probability density function P_μ describes the duration of the transition or the absolute arrival time of μ. Transitions are then described with a function $L(\mu|s,t,a)$ and the reward model is given through $R(\mu,t,\tau)$ and a cost of "dawdling" $K(s,t)$.

The optimality equations of (Boyan & Littman 2001) are:

$$V(s,t) = \sup_{t' \geq t} \left(\int_t^{t'} K(s,\theta)d\theta + \overline{V}(s,t') \right) \tag{11}$$

$$\overline{V}(s,t) = \max_{a \in A} Q(s,t,a) \tag{12}$$

$$Q(s,t,a) = \sum_{\mu \in M} L(\mu|s,t,a) \cdot U(\mu,t) \tag{13}$$

$$U(\mu,t) = \begin{cases} \int_{-\infty}^{\infty} P_\mu(t')[R(\mu,t,t') + V(s'_\mu,t')]dt' \\ \int_{-\infty}^{\infty} P_\mu(t'-t)[R(\mu,t,t') + V(s'_\mu,t')]dt' \end{cases} \tag{14}$$

Equation 14 is different whether $T_\mu = REL$ or ABS.

In the TMDP model, actions are defined as pairs "t', a", which mean "wait until time t' and then undertake action a". The optimality equations provided in (Boyan & Littman 2001) separate the action selection from the waiting duration, alternating a phase of action choice based on standard Q-values and a phase of "dawdling duration" optimization (this duration might be zero). With the XMDP formulation, it is pretty straightforward to see that the "wait" action really is a parametric action and that atomic action selection and waiting time determination can be separated because:

1. "wait" is the only parametric action,
2. "wait" is a deterministic action,
3. "wait" doesn't change the current discrete state of the process and only affects t.

If we write equation 10 for such a problem, the max and the sup operators appear to be independent, which allows for equations separation as in (Boyan & Littman 2001). Equation 10 becomes (with τ the parameter of the "wait" action):

$$V^*(s) = \max_{a \in A} \left\{ \sup_{\tau \in \mathbb{R}^+} \left(r(s,t,a(\tau)) + \iint\limits_{\substack{s' \in S \\ t' \in \mathbb{R}}} V^*(s')p(s',t'|s,t,a(\tau))ds'dt' \right) \right\}$$

$$V^*(s) = \max_{a \in A} \left\{ \sup_{\tau \in \mathbb{R}^+} \left(r(s,t,a(\tau)) + \iint\limits_{\substack{s' \in S \\ t' \in \mathbb{R}}} V^*(s') \sum_{\mu_{s'}} L(\mu_{s'}|s,a,t) \cdot P_{\mu_{s'}}(t'-t)ds'dt' \right) \right\}$$

$$V^*(s) = \max_{a \in A} \left\{ \sup_{\tau \in \mathbb{R}^+} \left(r(s,t,a(\tau)) + \sum_{s' \in S} L(\mu_{s'}|s,a,t) \cdot \int\limits_{t' \in \mathbb{R}} \gamma^{t'-t} V^*(s') \cdot P_{\mu_{s'}}(t'-t)dt' \right) \right\}$$

Therefore, if we separate $wait$ from the other actions:

$$V^*(s) = max \left\{ \max_{a \in A \setminus \{wait\}} \left\{ \sup_{\tau \in \mathbb{R}^+} \left(r(s,t,a(\tau)) + \sum_{s' \in S} L(\mu_{s'}|s,a,t) \int\limits_{t' \in \mathbb{R}} P_{\mu_{s'}}(t'-t)V^*(s')dt' \right) \right\}; \sup_{\tau \in \mathbb{R}^+} \Big(r(s,t,wait(\tau)) + V^*(s,t+\tau) \Big) \right\}$$

It is straightforward to see that there cannot be two successive $wait$ actions, thus we can consider a sequence of $wait$-$action$ actions since $wait$ is deterministic and only affects t. This yields:

$$V^*(s) = \sup_{\tau \in \mathbb{R}^+} \Bigg(r(s,t,wait(\tau)) + \max_{a \in A \setminus \{wait\}} \Bigg\{ r(s,t,a(\tau)) + \sum_{s' \in S} L(\mu_{s'}|s,a,t) \cdot \int_{t' \in \mathbb{R}} P_{\mu_{s'}}(t'-t) V^*(s') dt' \Bigg\} \Bigg)$$

This very last equation is finally the equivalent (condensed in one line) of equations 11 to 14 which proves that, in the end, TMDPs can be written as parametric action MDPs with total reward criterion and a single parametric $wait$ action. The proof above insures the validity of the Bellman equation in the general case.

Extensions to this example are the generalization of resolution methods for TMDPs. TMDPs are usually solved using piecewise constant L functions, discrete probability density functions and piecewise linear additive reward functions. (Boyan & Littman 2001) show that in this case, the value function can be computed exactly. Aside form this work on XMDPs, we developed a generalization of TMDPs in the more general case where all functions are piecewise polynomial and adapted the value iteration scheme introduced above to solve (with approximation) this class of XMDPs.

6 Conclusion

We have introduced an extension to the classical MDP model in three ways that generalize the kind of problems we can consider:

- We generalized (as was already done in previous work) MDPs to continuous and discrete state spaces
- We extended the standard discounted reward criterion to deal with a continuous observable time and random decision epoch dates
- We introduced factored continuous action spaces through the use of parametric actions

We called this extension XMDP, an on this basis we proved that our extended Bellman optimality equation (equation 10) characterized the optimal value function for XMDPs in the same way the standard value function characterizes the optimal value function for regular MDPs. More specifically, we defined the set of conditions under which this extended Bellman equation held, namely:

- action durations are strictly positive
- the reward model is a bounded upper semi-continuous function of the continuous variables.

Finally, in order to illustrate our approach, we showed how the TMDP model was actually an XMDP with a parametric "wait" action. This equivalence validates the optimality equations given in (Boyan & Littman 2001) for the total reward criterion.

This now allows the adaptation of standard value iteration algorithms as in (Boyan & Littman 2001) or (Feng *et al.* 2004) or the adaptation of heuristic search algorithms for MDPs as in (Barto, Bradtke, & Singh 1995) for example.

This paper's purpose was to provide proofs of optimality for a class of MDPs that has received attention recently in the AI community and to set a sound basis for the generalized parametric MDP framework. Our current work focuses on temporal planning under uncertainty and therefore makes use of the results introduced here, but we believe the scope of this paper goes beyond our research applications since it provides a general framework and optimality equations.

An other essential element when dealing with time and planning is the possibility of concurrency between actions / events. From this point of view, our future work will focus more on Generalized Semi-Markov Decision Processes optimization (Younes & Simmons 2004) .

References

Barto, A. G.; Bradtke, S. J.; and Singh, S. P. 1995. Learning to act using real-time dynamic programming. *Artificial Intelligence* 72(1-2):81–138.

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

Bertsekas, D. P., and Shreve, S. E. 1996. *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific.

Boyan, J. A., and Littman, M. L. 2001. Exact solutions to time dependent MDPs. *Advances in Neural Information Processing Systems* 13:1026–1032.

Bresina, J.; Dearden, R.; Meuleau, N.; Smith, D.; and Washington, R. 2002. Planning under continuous time and resource uncertainty: A challenge for ai. In *18th Conference on Uncertainty in Artificial Intelligence*.

Feng, Z.; Dearden, R.; Meuleau, N.; and Washington, R. 2004. Dynamic programming for structured continuous markov decision problems. In *20th Conference on Uncertainty in Artificial Intelligence*, 154–161.

Guestrin, C.; Hauskrecht, M.; and Kveton, B. 2004. Solving factored MDPs with continuous and discrete variables. In *20th Conference on Uncertainty in Artificial Intelligence*.

Hasselt, H., and Wiering, M. 2007. Reinforcement learning in continuous action spaces. In *IEEE Symposium on Approximate Dynamic Programming and Reinforcement Learning*.

Hauskrecht, M., and Kveton, B. 2006. Approximate linear programming for solving hybrid factored MDPs. In *9th International Symposium on Artificial Intelligence and Mathematics*.

Li, L., and Littman, M. 2005. Lazy approximation for solving continuous finite-horizon MDPs. In *National Conference on Artificial Intelligence*.

Liu, Y., and Koenig, S. 2006. Functional value iteration for decision-theoretic planning with general utility functions. In *National Conference on Artificial Intelligence*.

Marecki, J.; Topol, Z.; and Tambe, M. 2006. A fast analytical algorithm for markov decision process with continuous state spaces. In *AAMAS06*, 2536–2541.

Puterman, M. L. 1994. *Markov Decision Processes*. John Wiley & Sons, Inc.

Younes, H. L. S., and Simmons, R. G. 2004. Solving generalized semi-markov decision processes using continuous phase-type distributions. In *National Conference on Artificial Intelligence*.