Variable resolution discretization for high-accuracy solutions of optimal control problems

Rémi Munos and Andrew Moore
Robotics Institute, Carnegie Mellon University,
5000 Forbes Ave, Pittsburgh, PA 15213, USA
E-mail : {munos, awm}@cs.cmu.edu
Web page : http://www.cs.cmu.edu/~AUTON/

Abstract

State abstraction is of central importance in reinforcement learning and Markov Decision Processes. This paper studies the case of variable resolution state abstraction for continuous-state, deterministic dynamic control problems in which near-optimal policies are required. We describe variable resolution policy and value function representations based on Kuhn triangulations embedded in a kd-tree. We then consider top-down approaches to choosing which cells to split in order to generate improved policies. We begin with local approaches based on value function properties and policy properties that use only features of individual cells in making splitting choices. Later, by introducing two new non-local measures, *influence* and *variance*, we derive a splitting criterion that allows one cell to efficiently take into account its impact on other cells when deciding whether to split. We evaluate the performance of a variety of splitting criteria on many benchmark problems (published on the web), paying careful attention to their number-of-cells versus closeness-to-optimality tradeoff curves.

1 Introduction

This paper is about non-uniform discretization of state spaces when finding controllers for continuous optimal control problems. Uniform discretizations suffer from impractical computational requirements when the size of the discretization step is small. In this paper we try to keep the convergence properties of the discretization methods while introducing a variable resolution approximation. Here, we only consider the "general towards specific" approach : an initial coarse grid is successively refined at some areas of the state space by using a splitting process, until some desired measure of accuracy is reached.

We consider discounted deterministic control problems, which include the well-known reinforcement learning (RL) benchmarks of "car on the hill" [Moore, 1991], Cart-Pole [Barto *et al.*, 1983] and Acrobot [Sutton, 1996a]. Let $x(t) \in X \subset \mathbb{R}$ be the *state* of the system whose evolution is described by the controlled differential equation :

$$dx/dt = f(x(t), u(t)) \tag{1}$$

The objective of the control problem is to find, for any initial state x, the control $u^*(t)$ that optimizes the gain :

$$J(x; u(t)) = \int_0^\tau \gamma^t r(x(t), u(t))dt + \gamma^\tau R(x(\tau)) \tag{2}$$

where $r(x,u)$ is the *current reinforcement*, $R(x)$ the *boundary reinforcement* (obtained at the boundary of the state-space), γ the *discount factor* ($0 \leq \gamma < 1$), and τ the exit time from X. The *value function* (VF), maximal value of the gain, is :

$$V(x) = \sup_{u(t)} J(x; u(t)).$$

It is known (see [Fleming and Soner, 1993]) that V satisfies a first-order non-linear differential equation, called the *Hamilton-Jacobi-Bellman* (HJB) equation :

$$V(x)\ln\gamma + \max_{u\in U}[DV(x).f(x,u) + r(x,u)] = 0 \tag{3}$$

with DV being the gradient of V.

Section 2 introduces the discretization process ; *section* 3 describes several local splitting criteria, based on the value function and on the policy ; and *section* 4 proposes a heuristic for a global splitting criterion.

2 The discretization process

The state-space is discretized into a variable resolution grid using a structure of a kd-tree. The root of the tree covers the whole state space, supposed to be a (hyper) rectangle. Each node (except for the leaf nodes) splits in some direction the rectangle it covers at its middle into two nodes of half area. For each leaf, we use a Kuhn triangulation to linearly interpolate inside the rectangle (see the triangulation of figure 1). See [Munos and Moore, 1999] for more details. This defines a class of functions known as *barycentric interpolators* [Munos and Moore, 1998] which are piecewise linear, continuous inside each rectangle, but may be discontinuous at the boundary between two rectangles.

The discretization process is based on the finite-element methods of [Kushner and Dupuis, 1992] (extended to RL in [Munos, 1999]) which approximate, for any given discretization of the state space, the continuous deterministic control process by a Markov Decision Process (MDP). Note that the stochastic aspect of the MDP does not come from the continuous problem itself (which is deterministic here), but from the discretization process used.

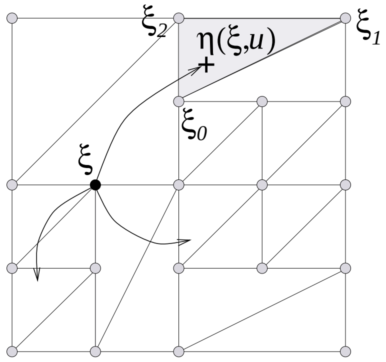


Figure 1: An example of discretization of the state-space. Values are stored at the corners ξ_i and the value for a non-corner point is linearly interpolated from the simplex of which it is a member. This is a fast process above 2 dimensions.

2.1 Building the discretized MDP

For a given discretization, we build the corresponding MDP in the following way. The state-space is the set Ξ of corners of the tree. For every corner ξ and control u, we approximate a part of the corresponding trajectory $x(t)$ by integrating the state dynamics (1) from initial state ξ for a constant control u, during some time $\tau(\xi,u)$ until it enters inside a new simplex at some point $\eta(\xi,u)$ (see Figure 1). At the same time, we also compute the integral of the current reinforcement : $R(\xi,u)=\int_{t=0}^{\tau(\xi,u)}\gamma^t.r(x(t),u).dt$.

Then we compute the vertices $(\xi_0,...,\xi_d)$ of the simplex containing $\eta(\xi,u)$ and the corresponding barycentric coordinates $\lambda_{\xi_0}(\eta(\xi,u)),...,\lambda_{\xi_d}(\eta(\xi,u))$ (which, by definition, satisfy : $\sum_{\xi_i}\lambda_{\xi_i}(\eta)=1$ and $\sum_{\xi_i}\lambda_{\xi_i}(\eta)(\eta-\xi_i)=0$). The interpolated value at $\eta(\xi,u)$ is thus just a linear combination of the values at the vertices $\xi_0,...,\xi_d$ of the simplex it belongs to, with positive coefficients that sum to one. Doing this interpolation is mathematically equivalent to probabilistically jumping to a vertex. Thus, we define the probabilities of transition $p(\xi_i|\xi,u)$ of the MDP from state ξ and control u to states ξ_i as these barycentric coordinates : $p(\xi_i|\xi,u)=\lambda_{\xi_i}(\eta(\xi,u))$, and the dynamic programming (DP) equation corresponding to this MDP is :

$$V(\xi)=\max_{u\in U}\left[\gamma^{\tau(\xi,u)}\sum_{i=0}^{d}\lambda_{\xi_i}(\eta(\xi,u)).V(\xi_i)+R(\xi,u)\right] \quad (4)$$

2.2 The "Car on the Hill" example

This problem, which will be made available on the web, is of 2 dimensions : position and velocity of the car. Here, the current reinforcement $r(x,u)$ is zero everywhere. The terminal reinforcement $R(x)$ is -1 if the car exits from the left side of the state-space, and varies linearly between $+1$ and -1 depending on the velocity of the car when it exits from the right side of the state-space. The best reinforcement $+1$ occurs when the car reaches the right boundary with a null velocity (see Figures 2 and 4). The control u has only 2 possible values : maximal positive or negative thrust. Because of the discount factor, we are encouraged to get to the goal as quickly as possible.

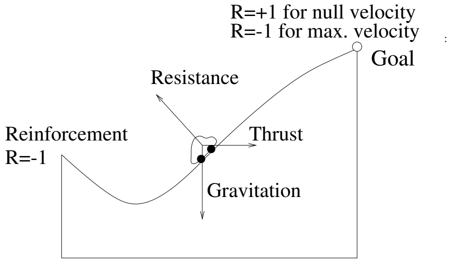


Figure 2: The "Car on the Hill" control problem.

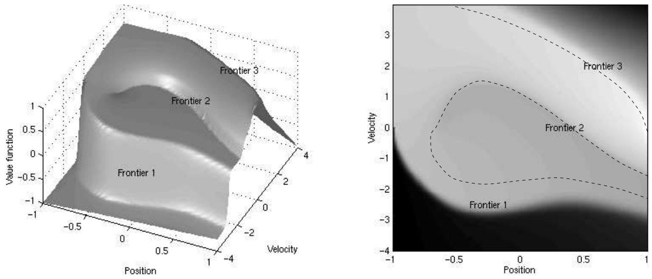


Figure 3: The value function of the Car-on-Hill. Frontier 1 shows the discontinuity of the VF. Frontiers 2 and 3 (the dash lines) stand where there is a change in the optimal control.

Figure 3 shows the value function of this problem. The discontinuity along Frontier 1 happens because a point beginning just above the frontier can eventually get a positive reward whereas any point below is doomed to exit on the left. Note that there is no change in the optimal control around this frontier. There is a discontinuity in the gradient of the VF along the upper part of Frontier 2 and along Frontier 3 because of a change in the optimal control.

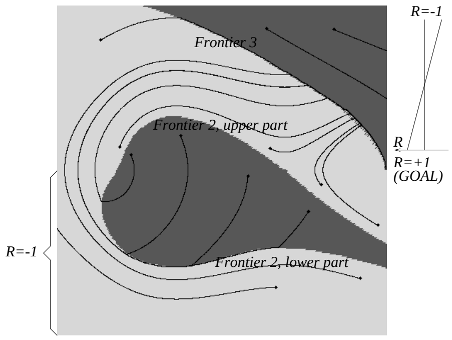


Figure 4: The optimal policy is represented by several gray levels (light gray for the positive control, dark gray for the negative one). Several optimal trajectories are drawn for different initial starting points.

2.3 The variable resolution approach

The basic idea is to start with an initial coarse discretization, build the corresponding MDP, solve it in order to have a (coarse) approximation of the value function ; then refine locally the discretization by splitting rectangles according to some "splitting criterion", build the new MDP, and so on (see the splitting process in Figure 5), until we estimate that the approximation of the value function or the optimal control is precise enough. The central purpose of this paper is the study of several possible splitting criteria.

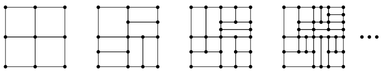


Figure 5: Several discretizations resulting of successive splitting operations.

3 Local splitting criteria

3.1 First criterion : average corner-value difference

For every hyper-rectangle in the tree we can ask the question "do the values on the left-side of the rectangle tend to have a significantly different value on average than those on the right-side?". The notion of left and right depend on which axis we are considering splitting on, and in fact we iterate over all axes, finding the one with the most significant difference, which we will call the "corner-value difference".

Having computed this for all cells in the current tree, we must choose which cells to actually split. We select some fraction (e.g. 50%) of the cells with the highest criterion values and break them in half along their most significant splitting directions. Figure 6 represents the discretization obtained using this criterion.

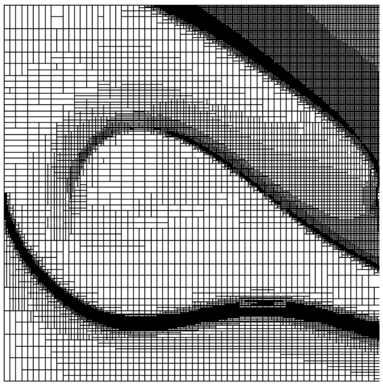


Figure 6: The discretization of the state-space for the Car-on-Hill problem obtained after 15 splitting iterations, starting with a 9 by 9 initial grid and using the *corner-value difference* criterion with a splitting rate of 50% of the rectangles at each iteration.

3.2 Second criterion : value non-linearity

For every rectangle, we compute the variance of the absolute increase of the values at the corners of the edges for all directions $i = 0..d$. This criterion is similar to the previous criterion except that it measures the extent to which the value throughout the cell is non-linear instead of non-constant. Figure 7 shows the corresponding discretization.

The *value non-linearity* criterion splits more parsimoniously than *corner-value difference* because it intends to

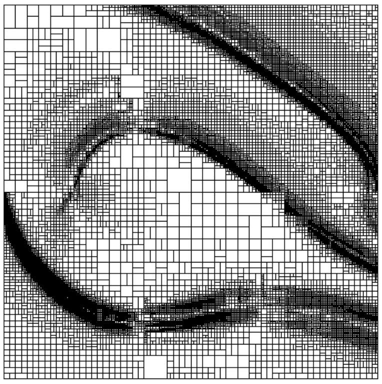


Figure 7: The discretization of the state-space for the Car-on-Hill problem using the *value non-linearity* criterion with a splitting rate of 50% after 15 iterations.

refine whenever the approximated function is non-linear instead of non-constant (see the difference of splitting in the area above the frontier 3, where the value function is almost linear).

We observe that in both cases, the refinement process does not split around the bottom part of frontier 2, because the VF is almost constant in this area although there is a change in the optimal control.

We observe that in both of these splitting processes based on the approximation of the VF, there is a huge amount of memory spent for the approximation of the discontinuity despite the optimal control being constant in this area. This is undesirable and so in the next section we will pay attention to the policy when splitting.

3.3 The policy disagreement criterion

Figure 4 shows the optimal policy and several optimal trajectories for different starting points. We would like to define a process that could refine around the areas of change in the optimal control, that is around frontiers 2 and 3, but not around frontier 1.

When we solve the MDP and compute the value function of the DP equation, we can compare the policy given by the $\arg\max_u$ of equation (4) with the optimal control law derived from the local gradient of V (given by the $\arg\max_u$ in equation 3). We define the states of *policy disagreement* being those where these two measures of the optimal control diverge.

Figure 8 shows the discretization obtained by splitting the rectangles containing states of policy disagreement.

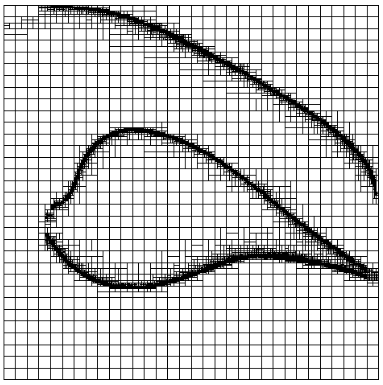


Figure 8: The discretization of the state-space for the Car-on-Hill using the *policy disagreement* criterion (Here we used an initial grid of 33×33 and a splitting rate of 20% to obtain this discretization).

This criterion is interesting since it splits at the places where there is a change in the optimal control, thus refining the resolution at the most important parts of the state-space for the approximation of the optimal control. However, as we can expect, if we only use this criterion, the value function will not be well approximated, thus this process may converge to a sub-optimal performance (see section 3.5). Indeed, we can observe that on Figure 8, the bottom part of frontier 2 is situated higher than its optimal position, illustrated on Figure 4. This results in an underestimation of the value function at this area because of the lack of precision around the discontinuity (frontier 1).

3.4 Combination of several criteria

We can combine *policy disagreement* with the *corner-value difference* or *value non-linearity* criterion in order to take the advantages of both methods : a good approximation of the value function on the whole state-space and an increase of the resolution around the areas of change in the optimal control. We can combine the previous criteria in several ways, for example by a weighted sum of the respective criteria, or by a logical operation (split if an and/or combination of these criteria is satisfied). Figure 9 shows the discretization obtained by alternatively, between iterations, using the *value non-linearity* criterion and the *policy disagreement* criterion. We notice that the lower part of frontier 2 is well refined.

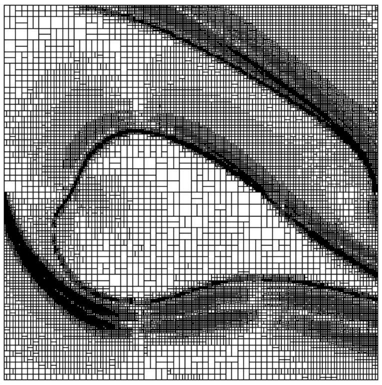


Figure 9: The discretization of the state-space for the "Car on the Hill" problem using the combination of the *value non-linearity* and the *policy disagreement* criterion.

3.5 Comparison of the performances

To compare the discretizations, we ran a set (here 256) of trajectories starting from initial states regularly situated in the state-space, using the policies resulting from the discretizations. The *performance* of a discretization is defined as the sum of the gain (defined by equation (2)) of these trajectories. Figure 10 shows the respective performances of several splitting criteria as a function of the number of states.

We notice the following points :

- Both the *corner-value difference* and *value non-linearity* splitting processes perform better than the uniform grids. For example, in order to obtain an almost optimal performance of 80, the variable resolution grids resulting of these criteria need around four times fewer states than the uniform one.

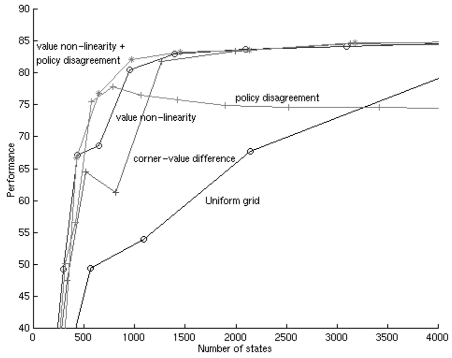


Figure 10: The performance for the uniform versus variable resolution grids for several splitting criterion

- The *value non-linearity* splitting performs slightly better than the *corner-value difference* one.
- The *policy disagreement* splitting is very good for a small number of states but does not improve after, and leads to a sub-optimal performance.
- The *policy disagreement* combined with the *value non-linearity* gives the best performance.

Can we do better ? So far, we have only considered local splitting criteria, in the sense that we decide whether or not to split a rectangle according to information (value function and policy) relative to the rectangle itself. However, the effect of the splitting is not local : it has an influence on the whole state-space.

4 A global splitting criterion

In the following sections we introduce two new measures of an Markov chain : *influence* and *variance* that have been designed to be efficient to provide useful indexes of which parts of state-space are most important to devote data gathering or computational resources to. We derive a global splitting heuristic.

4.1 Influence of a Markov chain

Let us consider a Markov chain whose set of states is $\Xi = \{\xi_1, ..., \xi_n\}$ and probabilities of transition from state ξ to state ξ_i are $p(\xi_i|\xi)$. We assume that the discount factor is a function of the state, and is written $\gamma^{\tau(\xi)}$ with $\gamma < 1$ for some holding time $\tau(\xi)$. When the system gets in state ξ, it receives a reinforcement $R(\xi)$.

We wish to define the influence $I(\xi_i|\xi)$ of a state ξ_i on another state ξ as a measure of how much the state ξ_i "contributes" to the value function of state ξ.

For that purpose, let us define the discounted cumulative $k-$chained probabilities $p_k(\xi_i|\xi)$, which represent the sum of the discounted transition probabilities of all sequences of k states from ξ to ξ_i :

$$\begin{aligned}
p_0(\xi_i|\xi) &= 1 \text{ (if } \xi = \xi_i) \text{ or } 0 \text{ (if } \xi \neq \xi_i) \\
p_1(\xi_i|\xi) &= \gamma^{\tau(\xi)} p(\xi_i|\xi) \\
p_2(\xi_i|\xi) &= \sum_{\xi_j \in \Xi} p_1(\xi_i|\xi_j).p_1(\xi_j|\xi) \\
&\dots \\
p_k(\xi_i|\xi) &= \sum_{\xi_j \in \Xi} p_1(\xi_i|\xi_j).p_{k-1}(\xi_j|\xi) \qquad (5)
\end{aligned}$$

Definition 1 (Influence) *Let $\xi \in \Xi$. We define the* influence of a state ξ_i on the state ξ *as the quantity :*

$$I(\xi_i|\xi) = \sum_{k=0}^{\infty} p_k(\xi_i|\xi)$$

Let Ω be a subset of Ξ. We define the influence of a state ξ_i on the subset Ω *as* $I(\xi_i|\Omega) = \sum_{\xi \in \Omega} I(\xi_i|\xi)$.

We notice that if the holding times $\tau(\xi)$ are > 0, then the influence is well defined (and is bounded by $\frac{1}{1-\gamma^{\tau_{\min}}}$ with $\tau_{\min} = \min_\xi \tau(\xi)$). The intuitive idea that $I(\xi_i|\xi)$ represents the "contribution" of state ξ_i on the VF of state ξ is formalized by the result that $I(\xi_i|\xi)$ is the partial derivative of $V(\xi)$ by $R(\xi_i)$: $I(\xi_i|\xi) = \frac{\partial V(\xi)}{\partial R(\xi_i)}$.

Moreover, we prove that the influence satisfies the following property :

$$I(\xi_i|\xi) = \sum_{\xi_j} \gamma^{\tau(\xi_j)} p(\xi_i|\xi_j).I(\xi_j|\xi) + \left\{ \begin{array}{l} 1 \text{ if } \xi_i = \xi \\ 0 \text{ if } \xi_i \neq \xi \end{array} \right. \quad (6)$$

Computation of the influence. Equation (6) is not a Bellman's equation since the probabilities $\sum_{\xi_j} p(\xi_i|\xi_j)$ do not sum to 1, so we cannot deduce that the successive iterations :

$$I_{n+1}(\xi_i|\xi) = \sum_{\xi_j} \gamma^{\tau(\xi_j)} p(\xi_i|\xi_j).I_n(\xi_j|\xi) + \left\{ \begin{array}{l} 1 \text{ if } \xi_i = \xi \\ 0 \text{ if } \xi_i \neq \xi \end{array} \right. \quad (7)$$

converge to the influence by using the classical contraction property in max-norm (see [Puterman, 1994]).

However, we have the following property :

$$\begin{aligned} \textstyle\sum_{\xi_i} I(\xi_i|\xi) &= \textstyle\sum_{\xi_i} \sum_{\xi_j} \gamma^{\tau(\xi_j)} p(\xi_i|\xi_j).I(\xi_j|\xi) + 1 \\ &= \textstyle\sum_{\xi_j} \gamma^{\tau(\xi_j)}.I(\xi_j|\xi) + 1 \end{aligned}$$

Thus, by denoting $I(\Xi|\xi)$ the vector whose components are the $I(\xi_i|\xi)$ and by introducing the 1-norm $||I(\Xi|\xi)||_1 = \sum_i |I(\xi_i|\xi)|$, we deduce that :

$$||I_{n+1}(\Xi|\xi) - I(\Xi|\xi)||_1 \leq \gamma^{\tau_{\min}}.||I_n(\Xi|\xi) - I(\Xi|\xi)||_1$$

and we have the contraction property for the 1-norm which insures convergence of the iterated $I_n(\xi_i|\xi)$ to the unique solution (the fixed point) $I(\xi_i|\xi)$ of (6).

Illustration on the "Car on the Hill". For a given discretization of the state space, we approximate the optimal control problem by building the corresponding MDP through the process described in section 2.1. Once the MDP is solved, we consider the Markov chain resulting from the choice of the control for the optimal policy $u^* = \pi^*(\xi)$. Let us denote $R(\xi) = R(\xi, \pi^*(\xi))$, $p(\xi_i|\xi) = p(\xi_i|\xi, \pi^*(\xi))$, and $\tau(\xi) = \tau(\xi, \pi^*(\xi))$.

Figure 11 shows (in gray levels) the influence $I(\xi|\Omega)$ of the states ξ on $\Omega = \{3 \text{ points}\}$.

Now define Σ the subset of states of policy disagreement (in the sense of section 3.3). Figure 12(a) shows Σ for a regular grid (of 129×129). Then, the influence $I(\xi|\Sigma)$ of the states ξ on Σ is plotted in Figure 12(b).

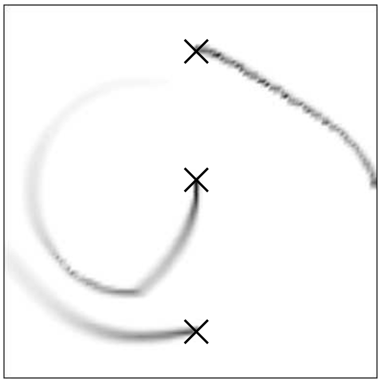
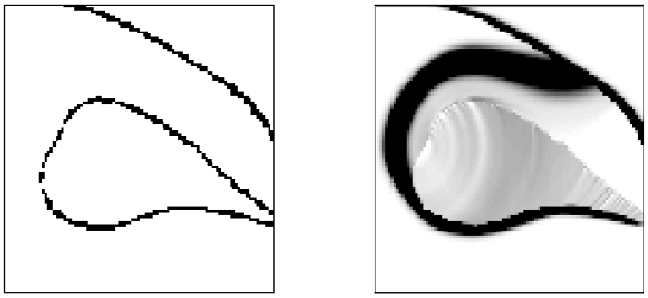


Figure 11: Influence $I(\xi|\Omega)$ on 3 points (the crosses). The influence on a state "follows" the direction of the optimal trajectory starting from that state through some kind of "diffusion process".



(a) States of policy disagreement

(b) Influence on these states

Figure 12: (a) The set Σ of policy disagreement and (b) the influence $I(\xi|\Sigma)$ on Σ.

The darkest zones are the places that "contribute" the most to the value function of states in Σ, thus the areas that are expected to affect the VF at the places of change in the optimal control.

In the following section, we introduce the *variance* of the Markov chain in order to get an estimation of how accurate the approximation of the VF is, for a given discretization.

4.2 Variance of a Markov chain

By using the notation of the previous section, we have : for any ξ, $V(\xi)$ satisfies the Bellman's equation :

$$V(\xi) - R(\xi) = \sum_{\xi_i \in \Xi} p(\xi_i|\xi) \gamma^{\tau(\xi)} V(\xi_i) \quad (8)$$

which states that $V(\xi) - R(\xi)$ is the discounted average of the next values $V(\xi_i)$ weighted by the probabilities of transition $p(\xi_i|\xi)$. We are interested in computing the variance of these values in order to have an estimation of the range of the values averaged by $V(\xi)$. We define :

$$e(\xi) = \sum_{\xi_i \in \Xi} p(\xi_i|\xi) \left[\gamma^{\tau(\xi)} V(\xi_i) - V(\xi) + R(\xi) \right]^2 \quad (9)$$

However, the values $V(\xi_i)$ also average some successive values $V(\xi_j)$, so we would like that the variance also takes into account this averaging, as well as all the following ones. In what follows, we define the value function as an averager of the reinforcements and give the definition of the variance of a Markov chain.

The value function as an averager of the reinforcements. Let us denote $s_k(\xi)$ a sequence of $(k+1)$ states $(\xi_0 = \xi, \xi_1, \xi_2, ..., \xi_k)$ whose first one is ξ. Let $S_k(\xi)$ be the set of all possible sequences $s_k(\xi)$. Let us denote $p(s_k(\xi)) = \prod_{i=1}^{k} p(\xi_i|\xi_{i-1})$, the product of the probabilities of transition of the successive states in a sequence, and for $i \leq k$, $\tau_i(s_k(\xi)) = \sum_{j=0}^{i-1} \tau(\xi_j)$ the cumulative time of the i^{th} first states of the sequence (with by definition $\tau_0(s_k(\xi)) = 0$).

We have the property that : $\sum_{s_k(\xi)\in S_k(\xi)} p(s_k(\xi)) = 1$. We can prove that the value function satisfies the following equation :

$$V(\xi) = \lim_{k\to\infty} \sum_{s_k(\xi)\in S_k(\xi)} p_k(s(\xi)) \left[\sum_{i=0}^{k} \gamma^{\tau_i(s_k(\xi))} R(\xi_i)\right] \qquad (10)$$

Thus, the value function is expressed as an average of the discounted reinforcements. Now, we can define the variance of these values.

Definition 2 (Variance) *We define the variance $\sigma^2(\xi)$ of any state $\xi \in \Xi$:*

$$\sigma^2(\xi) = \lim_{k\to\infty} \sum_{s_k(\xi)\in S_k(\xi)} p(s_k(\xi)) \left[\sum_{i=0}^{k} \gamma^{\tau_i(s_k(\xi))} R(\xi_i) - V(\xi)\right]^2$$

We can prove that the limit exists, thus the variance is well defined. Moreover, the variance satisfies :

$$\sigma^2(\xi) = e(\xi) + \gamma^{2\tau(\xi)} \sum_{\xi_i} p(\xi_i|\xi)\sigma^2(\xi_i)$$

with $e(\xi)$ defined by (9). Thus the variance $\sigma^2(\xi)$ is the sum of an immediate contribution $e(\xi)$ that takes into account the variation in the values of the immediate successors ξ_i, plus the discounted average of the variance $\sigma^2(\xi_i)$ of these successors.

This is a Bellman equation and it can be solved by DP methods (thanks to a contraction property in max-norm).

Notice that the variance defined here has no relationship to the variance of the gain obtained by following any specific policy for the real problem, which is always 0 since the continuous problem is deterministic. Here, the variance indicates the extent of uncertainty on the value function because of the discretization process, and thus gives an estimation of the quality of approximation of the value function for a given discretization.

By using some geometrical considerations (see [Munos and Moore, 1999]), we can prove that $e(\xi)$ is close to 0 in two specific cases : either if the gradient at the iterated point $\eta(\xi, \pi^*(\xi))$ is low (i.e. the values of the successive states ξ_i are almost constant) or if $\eta(\xi, \pi^*(\xi))$ is close to one corner ξ_i (then the barycentric coordinate λ_{ξ_i} is close to 1 and the λ_{ξ_j} (for $j \neq i$) are close to 0). In both of these cases, $e(\xi)$ is low and implies that the iteration of ξ does not lead to a degradation of the quality of approximation of the value function (the variance does not increase).

Figure 14a shows the standard-deviation $\sigma(\xi)$ for the Car-on-the-Hill problem for a uniform grid (of 257 by 257). We notice the following points :

- The standard deviation is very high around the discontinuity of the value function.
- There is a noticeable positive standard deviation around the 2 frontiers of discontinuity of the gradient of the value function, which correspond to a change in the optimal control. Indeed, in these areas, the VF is the average of the discounted reinforcement for different exit times (depending on whether the car can reach the goal directly or has to do one more loop).
- Apart from these areas, the standard deviation is very low, which means that the discounted reinforcements averaged by the value function are almost constant. *A refinement of the resolution in these areas has no chance of producing an important change in the value function.*

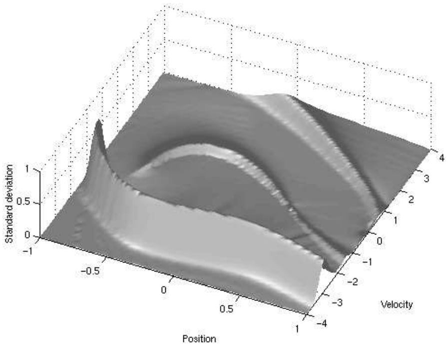


Figure 13: The standard deviation function for the "Car on the Hill".

Thus it appears that the areas where a splitting might affect the most the approximation of the value function are the cells whose corners have the highest standard deviations.

4.3 A global splitting heuristic

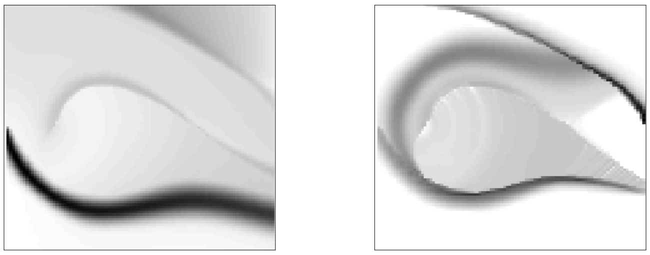
Now we combine the notions of *influence* and *variance* described in the previous sections, in order to define a non-local splitting criterion. We have seen that :

- The states ξ of highest standard deviation $\sigma(\xi)$ are the states of highest uncertainty on the quality of approximation of the VF, thus the states that could improve the most their approximation accuracy when split (see figure 14(a) for an illustration on the Car-on-the-Hill).
- The states ξ of highest influence (see figure 12(b)) on the set Ω of states of policy disagreement (figure 12(a)) are the states whose value function affects the area where there is a change in the optimal control.

Thus, in order to improve the precision of approximation at the most relevant areas of the state-space (i.e. where there is a change in the optimal control), an heuristic should be to split the states of highest *Stdev_Inf* criterion (see figure 14) where :

$$Stdev\_Inf(\xi) = \sigma(\xi).I(\xi|\Omega)$$

Figure 15 shows the discretization obtained by splitting the states of highest *Stdev_Inf*.



(a) Standard deviation (b) Influence x Standard deviation

Figure 14: (a) The standard-deviation $\sigma(\xi)$ for the "Car on the Hill" and (b) The *Stdev_Inf* criterion $= \sigma(\xi).I(\xi|\Sigma)$.

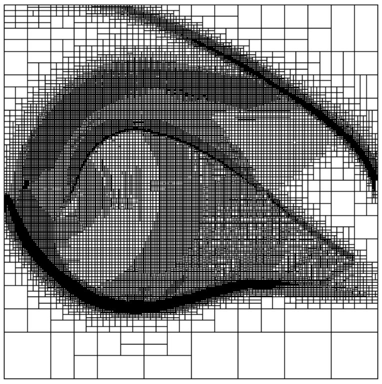


Figure 15: The discretization resulting of the splitting of the states of highest standard deviation that have an influence on the states of policy disagreement (*Stdev_Inf* criterion.)

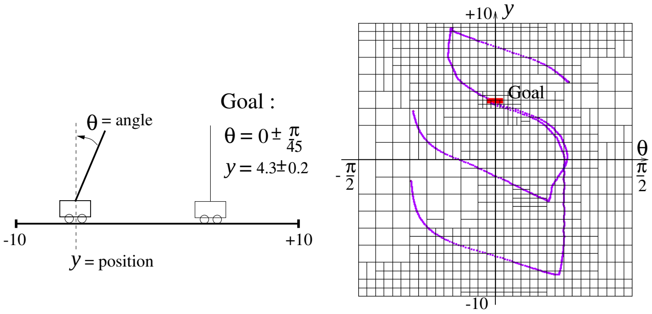
We observe that the *Stdev_Inf* criterion does not split the areas where the VF is discontinuous unless some refinement is necessary to get a better approximation of the optimal control. This turns out to be critical in higher dimensions, where the cost to get an accurate approximation of the VF is too high.

5 Illustration on other control problems

5.1 The Cart-Pole problem

The dynamics of this 4-dimensional physical system (illustrated in figure 16(a)) are described in [Barto *et al.*, 1983]. The goal is defined by the area : $y = 4.3 \pm 0.2$, $\theta = 0 \pm \frac{\pi}{45}$, (and no limits on $\dot{y}$ and $\dot{\theta}$). This is a much narrower goal than in previous cart-pole RL work. Notice that "minimum time maneuver to a small goal region" from an arbitrary start state is much harder than merely balancing the pole without falling. The current reinforcement is zero everywhere and the terminal reinforcement is -1 if the system exits from the state-space, and $+1$ if it reaches the goal.

Figure 17 shows the performance obtained for several splitting criteria previously defined. Note that the local criteria do not perform better than the uniform grids because the VF is discontinuous at several parts of the



(a) The "Cart-pole" (b) The projection of the state space

Figure 16: (a) Description of the cart-pole 4d problem. (b) The projection of the discretization (onto the plane (θ,y)) obtained by the *Stdev_Inf* criterion and some trajectories for several initial points.

state-space (because of boundary problems similar to the frontier 1 of the "Car on the Hill" problem) and the value-based criteria spend too many resources on approximating these useless areas. But note too that the *Stdev_Inf* criterion performs very well because it avoids this problem.

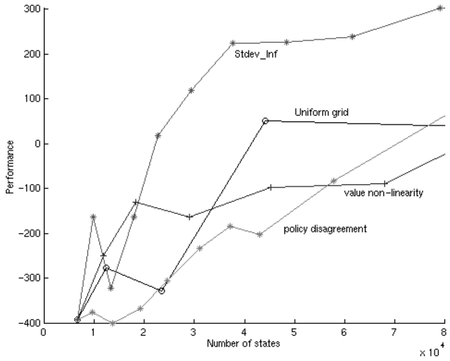


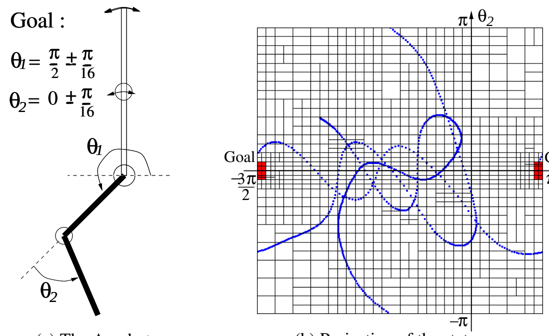
Figure 17: Comparison of the performances on the Cart-Pole problem for several splitting criteria.

5.2 The Acrobot

The Acrobot, which has a 4-d state-space, is a two-link arm with a single actuator at the elbow (see figure 18). Here, the goal of the controller is to balance the Acrobot at its unstable, inverted vertical position, in the minimum time.

This actuator exerts a torque between the links. In [Sutton, 1996b] the goal was to lift the hand to a given height, but here we perform the harder task of balancing the Acrobot at its unstable, inverted vertical position, in the minimum time. The goal is defined by a very narrow range of $\frac{\pi}{16}$ on both angles around the vertical position $\theta_1 = \frac{\pi}{2}, \theta_2 = 0$ (figure 18(b)), for which the system receives a reinforcement of $R = +1$. Anywhere else, the reinforcement is zero. The two first dimensions (θ_1, θ_2) of the state space have a structure of a torus (because of the 2π modulo on the angles), which is implemented in our structure by having the vertices of the 2 first dimensions being angle 0 and 2π pointing to the same entry for the value function. Figure 19 shows the performance.

Interpretation of the results : As we noticed for



(a) The Acrobot (b) Projection of the state space

Figure 18: (a) Description of the Acrobot. (b) Projection of the discretization (onto the plane (θ_1,θ_2)) obtained by the *Stdev_Inf* criterion, and one trajectory.

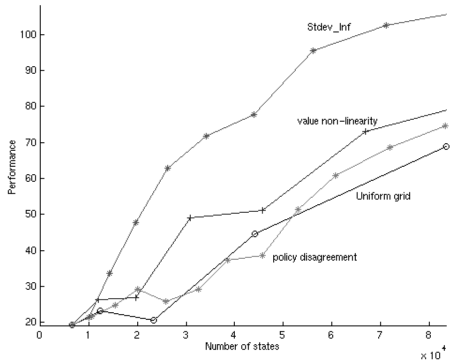


Figure 19: Comparison of the performances on the acrobot problem for several splitting criteria.

the two previous 4d problems, the local splitting criteria fail to improve the performance of the uniform grids because they spend too many resources on local considerations (either approximating the value function or the optimal policy). For example, on the Cart-Pole problem, the *value non-linearity* criterion will concentrate on approximating the VF mostly at parts of the state space where there is already no chance to rebalance the pole. And the areas around the vertical position (low θ), which are the most important areas, will not be refined in time (however, if we continue the simulations after about 90000 states, the local criteria start to perform better than the uniform grids, because these areas get eventually refined).

The *Stdev_Inf* criterion, which takes into account global consideration for the splitting, performs very well for all the problems described above.

6 Conclusion

In this paper we proposed a variable resolution discretization approach to solve continuous time and space control problems. We described several local splitting criteria, based on the VF or the policy approximation. Local value-based splitting is an efficient, model-based, relative of the Q-learning-based tree splitting criteria used, for example, by [Chapman and Kaelbling, 1991; Simons *et al.*, 1982; McCallum, 1995]. But it is only when combined with new measures based on policy and on global considerations (influence and variance) that we are able to get truly effective, near-optimal performance on our control problems. The tree-based state-space partitions in [Moore, 1991; Moore and Atkeson, 1995] were produced by different criteria (of empirical performance), and produced far more parsimonious trees, but no attempt was made to minimize cost: merely to find a valid path. Our planned future work includes pruning, in order to include "specific towards general" grouping of areas that have been over-refined, and studying behavior on continuous stochastic systems such as production scheduling.

Acknowledgments.
This research was sponsored by DASSAULT-AVIATION and CMU.

References

[Barto *et al.*, 1983] A. G. Barto, R. S. Sutton, and C. W. Anderson. Neuronlike Adaptive elements that that can learn difficult Control Problems. *IEEE Trans. on Systems Man and Cybernetics*, 13(5):835–846, 1983.

[Chapman and Kaelbling, 1991] D. Chapman and L. P. Kaelbling. Learning from Delayed Reinforcement In a Complex Domain. In *IJCAI-91*, 1991.

[Fleming and Soner, 1993] Wendell H. Fleming and H. Mete Soner. *Controlled Markov Processes and Viscosity Solutions*. Applications of Mathematics. Springer-Verlag, 1993.

[Kushner and Dupuis, 1992] Harold J. Kushner and Dupuis. *Numerical Methods for Stochastic Control Problems in Continuous Time*. Applications of Mathematics. Springer-Verlag, 1992.

[McCallum, 1995] A. McCallum. Instance-Based Utile Distinctions for Reiforcement Learning with Hidden State. In *Machine Learning (proceedings of the twelfth international conference)*, San Francisco, CA, 1995. Morgan Kaufmann.

[Moore and Atkeson, 1995] A. W. Moore and C. G. Atkeson. The Parti-game Algorithm for Variable Resolution Reinforcement Learning in Multidimensional State-spaces. *Machine Learning*, 21, 1995.

[Moore, 1991] A. W. Moore. Variable Resolution Dynamic Programming: Efficiently Learning Action Maps in Multivariate Real-valued State-spaces. In L. Birnbaum and G. Collins, editors, *Machine Learning: Proceedings of the Eighth International Workshop*. Morgan Kaufmann, June 1991.

[Munos and Moore, 1998] Rémi Munos and Andrew Moore. Barycentric interpolators for continuous space and time reinforcement learning. *Neural Information Processing Systems*, 1998.

[Munos and Moore, 1999] Rémi Munos and Andrew Moore. Variable resolution discretization in optimal control. *Robotics Institute Technical Report, CMU*, 1999.

[Munos, 1999] Rémi Munos. A study of reinforcement learning in the continuous case by the means of viscosity solutions. *To appear in Machine Learning Journal*, 1999.

[Puterman, 1994] Martin L. Puterman. *Markov Decision Processes, Discrete Stochastic Dynamic Programming*. A Wiley-Interscience Publication, 1994.

[Simons *et al.*, 1982] J. Simons, H. Van Brussel, J. De Schutter, and J. Verhaert. A Self-Learning Automaton with Variable Resolution for High Precision Assembly by Industrial Robots. *IEEE Trans. on Automatic Control*, 27(5):1109–1113, October 1982.

[Sutton, 1996a] R. S. Sutton. Generalization in Reinforcement Learning: Successful Examples Using Sparse Coarse Coding. In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Neural Information Processing Systems 8*, 1996.

[Sutton, 1996b] Richard S. Sutton. Generalization in reinforcement learning : Successful examples using sparse coarse coding. *Advances in Neural Information Processing Systems*, 8, 1996.