Influence and Variance of a Markov Chain : Application to Adaptive Discretization in Optimal Control

Rémi Munos and Andrew Moore
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

This paper addresses the difficult problem of deciding where to refine the resolution of adaptive discretizations for solving continuous time-and-space deterministic optimal control problems. We introduce two measures, *influence* and *variance* of a Markov chain. Influence measures the extent to which changes of some state affect the value function at other states. Variance measures the heterogeneity of the future cumulated active rewards (whose mean is the value function). We combine these two measures to derive a non-local efficient splitting criterion that takes into account the impact of a state on other states when deciding whether to split. We illustrate this method on the non-linear, two dimensional "Car on the Hill" and the 4d "space-shuttle" and "airplane-meeting" control problems.

1 Introduction

In this paper, we introduce the theory behind two notions related to a Markov chain (MC) : **influence** and **variance**. We combine them to obtain an efficient splitting criterion for refining the resolution of adaptive discretizations of continuous time-and-space deterministic control problems.

Influence is a measure of the extent to which a state "contributes" to the value function of other states. It will be used to find out the states that have an influence on the most important areas of the state space for designing an accurate controller : the boundaries of change in the optimal control.

In the process of discretizing the continuous process into a finite Markov Decision Process (MDP), we introduce a bias which depends on the structure of the discretization. The variance of the MC derived from this MDP provides a good estimation of this bias, thus indicating the parts of the discretization that need to be refined to improve the accuracy of the approximations.

The global splitting heuristic we propose in order to improve the controller is to refine the discretization at states of high variance (the states that could improve the accuracy of the value function the most when split) that have an influence on the boundaries of change in the optimal control.

Section 2 introduces the notion of influence and provides an algorithm to compute it. *Section 3* considers MCs obtained through a discretization of continuous processes, illustrated on the "Car on the Hill" (see [6]). *Section 4* introduces the variance of a MC and shows that it satisfies a Markov chain. Finally in *section 5* we state our global splitting heuristic.

2 Influence of a Markov chain

Let us consider a Markov chain (MC) with state space $\Xi = \{\xi_1, ..., \xi_n\}$ whose probabilities of transition from state ξ to state ξ_i are $p(\xi_i|\xi)$. Here 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 $\tau(\xi) > 0$. We also assume that the rewards are deterministic and only depend on the state : when the system is in state ξ it receives a reward (or reinforcement) $R(\xi)$. Extensions to general MCs are possible. Let $\{\xi(t)\}_{t\geq 0}$ be a sequence of states whose discounted cumulative reward is : $J(\{\xi(t)\}_{t\geq 0}) = \sum_{t\geq 0} \gamma^{\sum_{s=0}^{s<t} \tau(\xi(s))} R(\xi(t))$. Then the value function (VF) of a state ξ is defined by the expectation :

$$V(\xi) = E\left[J(\{\xi(t)\}_{t\geq 0})|\xi(0) = \xi\right] \quad (1)$$

2.1 Definition of the influence

The intuition behind the notion of influence is to measure the extent to which a state ξ_i "contributes" to the VF of another state ξ. This is done by estimating the change in the VF at ξ resulting from a modification of the reinforcement $R(\xi_i)$.

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

$$\begin{aligned}
p_0(\xi_i|\xi) &= 1 \text{ (if } \xi = \xi_i) \text{ or } 0 \text{ (if } \xi \neq \xi_i) \\
p_1(\xi_i|\xi) &= \gamma^{\tau(\xi)} p(\xi_i|\xi) \\
p_2(\xi_i|\xi) &= \sum_{\xi_j \in \Xi} p_1(\xi_i|\xi_j).p_1(\xi_j|\xi) \\
&\dots \\
p_k(\xi_i|\xi) &= \sum_{\xi_j \in \Xi} p_1(\xi_i|\xi_j).p_{k-1}(\xi_j|\xi) \quad (2)
\end{aligned}$$

Definition 1 (Influence) *Let $\xi \in \Xi$. We define the influence $I(\xi_i|\xi)$ 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 $\tau_{\min} = \min_\xi \tau(\xi)$). This definition is related to the intuitive idea stated above that **the influence $I(\xi_i|\xi)$ is the partial derivative of $V(\xi)$ by $R(\xi_i)$** :

$$I(\xi_i|\xi) = \frac{\partial V(\xi)}{\partial R(\xi_i)} \qquad (3)$$

Indeed, by applying Bellman's equation :
$V(\xi) = R(\xi) + \sum_{\xi_i} p_1(\xi_i|\xi).V(\xi_i)$ to ξ and ξ_i we obtain :

$$V(\xi) = R(\xi) + \sum_{\xi_i} p_1(\xi_i|\xi)\left[R(\xi_i) + \sum_{\xi_j} p_1(\xi_j|\xi_i).V(\xi_j)\right]$$

From the definition of p_2 we can rewrite this as :

$$V(\xi) = R(\xi) + \sum_{\xi_i} p_1(\xi_i|\xi).R(\xi_i) + \sum_{\xi_i} p_2(\xi_i|\xi).V(\xi_i)$$

Again we can apply several times Bellman equation to $V(\xi_i)$ and deduce that for any n,

$$V(\xi) = \sum_{k=0}^{n-1}\sum_{\xi_i} p_k(\xi_i|\xi).R(\xi_i) + \sum_{\xi_i} p_n(\xi_i|\xi).V(\xi_i)$$

and at the limit when n tends to infinity, we obtain :

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

Then, we deduce the contribution of $R(\xi_i)$ to $V(\xi)$:

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

2.2 Computation of the influence

For any states ξ and ξ_i, we have the property :

$$I(\xi_i|\xi) = \sum_{\xi_j} p_1(\xi_i|\xi_j).I(\xi_j|\xi) + \begin{cases} 1 & \text{if } \xi_i = \xi \\ 0 & \text{if } \xi_i \neq \xi \end{cases} \qquad (4)$$

This is deduced from the very definition of the influence and the chained probability property (2) since for all ξ,

$$\begin{aligned} I(\xi_i|\xi) &= \sum_{k=0}^{\infty} p_k(\xi_i|\xi) = \sum_{k=0}^{\infty} p_{k+1}(\xi_i|\xi) + p_0(\xi_i|\xi) \\ &= \sum_{k=0}^{\infty}\sum_{\xi_j} p_1(\xi_i|\xi_j).p_k(\xi_j|\xi) + p_0(\xi_i|\xi) \\ &= \sum_{\xi_j} p_1(\xi_i|\xi_j).I(\xi_j|\xi) + \begin{cases} 1 & \text{if } \xi_i = \xi \\ 0 & \text{if } \xi_i \neq \xi \end{cases} \end{aligned}$$

Equation (4) is not a Bellman equation since the sum of the probabilities $\sum_{\xi_j} p_1(\xi_i|\xi_j)$ may be greater than 1, so we cannot deduce that the successive iterations :

$$I_{n+1}(\xi_i|\xi) = \sum_{\xi_j} p_1(\xi_i|\xi_j).I_n(\xi_j|\xi) + \begin{cases} 1 & \text{if } \xi_i = \xi \\ 0 & \text{if } \xi_i \neq \xi \end{cases} \qquad (5)$$

converge to the influence by using the classical contraction property in max-norm (see [9]). However, we have the following property :

$$\begin{aligned} \textstyle\sum_{\xi_i} I(\xi_i|\xi) &= \textstyle\sum_{\xi_i}\sum_{\xi_j} p_1(\xi_i|\xi_j).I(\xi_j|\xi) + 1 \\ &= \textstyle\sum_{\xi_j} \gamma^{\tau(\xi_j)}.I(\xi_j|\xi) + 1 \end{aligned}$$

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 in 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 (4).

Remark 1 *As pointed out by Geoffrey Gordon, the influence is closely related to the* dual variables *(or* shadow prices *in economics) of the Linear Program equivalent to the Bellman equation (see [2]). This property has already been used in [11] to derive an efficient adaptive grid generation.*

Remark 2 *A possible extension is to define the* **influence of a MDP** *as the* infinitesimal *change in the value function of a state resulting from an* infinitesimal *modification of the reward at another state. Since the value function is a* max *of linear expressions, the influence on states with multiple optimal actions (thus for which the value function is not differentiable) is defined (as a set-valued map) by taking the partial sub-gradient instead of the regular gradient (3).*

3 Influence on discretized continuous problems

Here, we illustrate this notion of influence on a particular class of Markov chains derived from a discretization process of a continuous deterministic control problem. We use the method described in [8] based on Finite-Element methods (see [4, 7]) for adaptive triangulations.

We consider a deterministic control system whose state $x(t) \in X \subset \mathbb{R}^d$ is described by the controlled differential equation :

$$\frac{dx}{dt} = f(x(t), u(t)) \qquad (6)$$

where u is the control (chosen among a finite number of possible values U). The objective of the control problem is to find, for any initial state x, the control $u(t)$ that optimizes the (discounted) gain :

$$J(x; u(t)) = \int_0^\tau \gamma^t r(x(t), u(t))dt + \gamma^\tau R(x(\tau)) \qquad (7)$$

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

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

We know (see [1] for example) that V satisfies a first-order non-linear differential equation, called the *Hamilton-Jacobi-Bellman* (HJB) equation :

$$V(x)\ln\gamma + \max_{u\in U}[DV(x).f(x,u) + r(x,u)] = 0 \qquad (8)$$

with DV being the gradient of V.

3.1 A discretization process

In order to approximate the value function of the continuous process, we discretize the state-space using a variable resolution grid structured as a tree. The root of the tree covers the whole state space, assumed to be a (hyper) rectangle. Each node (except for the leaf nodes) splits in some direction (parallel to the axes) 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).

For a given triangulation, we build the following Markov Decision Process (MDP) : the state-space of the MDP is the set of corners of the tree. For every corner ξ and control $u \in U$ we approximate a part of the corresponding trajectory $x(t)$ by integrating the state dynamics (6) from initial state ξ for a constant control u, during some time $\tau(\xi,u)$ until it enters inside a new rectangle at some *iterated point* $\eta(\xi,u)$ (see Figure 1). At the same time we compute the cumulated reinforcement : $R(\xi,u) = \int_{t=0}^{\tau(\xi,u)} \gamma^t.r(x(t),u).dt$.

Then we find out the corners $(\xi_0,\ldots,\xi_d)$ of the simplex containing $\eta(\xi,u)$ and the corresponding barycentric coordinates $\lambda_{\xi_0}(\eta(\xi,u)),\ldots,\lambda_{\xi_d}(\eta(\xi,u))$ (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 $\xi_0,\ldots,\xi_d$ 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 : we create a stochastic discrete MDP from a deterministic continuous process**. The probabilities of transition 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 :

$$V(\xi) = \max_u \left[\gamma^{\tau(\xi,u)} \sum_{i=0}^{d} \lambda_{\xi_i}(\eta(\xi,u)).V(\xi_i) + R(\xi,u) \right] \qquad (9)$$

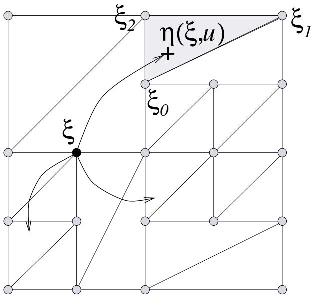


Figure 1: Triangulation of the state-space. The continuous *deterministic* control problem is discretized into a *stochastic* MDP.

3.2 Illustration of the influence on the "Car on the Hill"

This is a non-linear control problem of dimension 2 (the position and the velocity of the car). For a description of the dynamics of this problem see [6] and for the reinforcement functions used here see [8] (and figure 2). The control u has 2 possible values : maximal positive or negative thrust.

Figure 3 shows the interpolated value function of the MDP

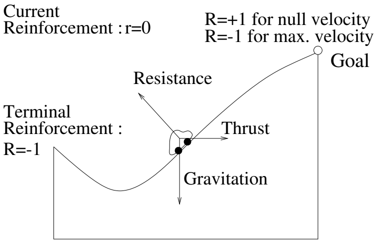


Figure 2: The "Car on the Hill" control problem.

obtained by a uniform discretization of 257 by 257 states.

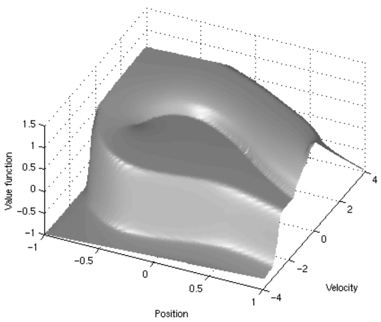
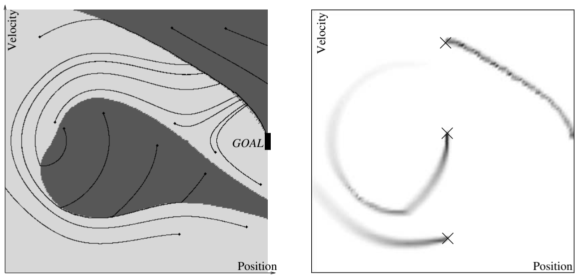


Figure 3: The value function of the "Car on the Hill"

Once the MDP is solved, we consider the Markov chain resulting from the choice of the optimal action u^*. Let us denote $R(\xi) = R(\xi,u^*)$, $p(\xi_i|\xi) = p(\xi_i|\xi,u^*)$, and $\tau(\xi) = \tau(\xi,u^*)$. Then we can compute the influence of the MC on any subset Ω. As an example, figure 4(b) shows the influence $I(\xi_i|\Omega)$ on the subset Ω composed of 3 points (the crosses), as a function of ξ_i. We notice that the influence on a state "follows" the direction of the optimal trajectory starting from that state (see figure 4(a)) through some "diffusion process" (introduced by the discretization process).



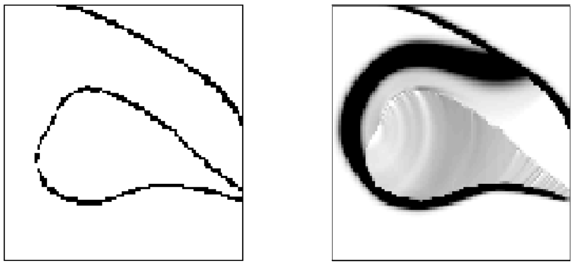
(a) The optimal policy and several trajectories (b) Influence on 3 points

Figure 4: (a) The optimal policy is indicated by 2 gray levels, and several trajectories are drawn for different starting points. (b) Influence $I(\xi_i|\Omega)$ on Ω composed of 3 points (the crosses).

In order to improve the resolution at the areas of the state space where there is a change in the optimal control, we compute the influence $I(\xi_i|\Sigma)$ on the subset Σ of the states of *policy disagreement* for which the policy given by the $\arg\max_u$ of equation (9) differs from the policy derived by the gradient of the value function (the $\arg\max_u$ of equation (8)). See figure 5.

The darkest zones in Figure 5(b) are the areas that contribute the most to the switching boundary of the optimal control.

Now, we would like to define the areas whose refinement



(a) States of policy disagreement (b) Influence on these states

Figure 5: (a) The set Σ of states of policy disagreement and (b) the influence $I(\xi_i|\Sigma)$ on Σ.

could increase the quality of approximation of the value function. In the following section, we introduce the *variance* of the Markov chain in order to estimate the bias introduced by the discretization process and thus derive an approximation error of the discretized VF.

4 Variance of a Markov chain

Whereas in [8] we define the variance as an averaging process, here we follow the advice of Csaba Szepesvári (personal communication) and define it as the expectation :

$$\sigma^2(\xi) = E\left[(J(\{\xi(t)\}_{t\geq 0}) - V(\xi))^2 | \xi(0) = \xi\right]$$

Now we prove that the variance satisfies a Bellman equation. Indeed, we have :

$$\sigma^2(\xi) = E\left[[J(\{\xi(t)\}_{t\geq 1}) - (V(\xi) - R(\xi))]^2 | \xi(0) = \xi\right]$$

From Bellman's equation :
$V(\xi) - R(\xi) = E[J(\{\xi(t)\}_{t\geq 1}) | \xi(0) = \xi]$ we deduce that :

$$\sigma^2(\xi) = E\left[J(\{\xi(t)\}_{t\geq 1})^2 - (V(\xi) - R(\xi))^2 | \xi(0) = \xi\right]$$

Let us introduce a conditional expectation with respect to the next state $\xi' = \xi(1)$:

$$\sigma^2(\xi) = \sum_{\xi'} p(\xi'|\xi).E\Big[J(\{\xi(t)\}_{t\geq 1})^2 - [\gamma^{\tau(\xi)}V(\xi')]^2 +$$
$$[\gamma^{\tau(\xi)}V(\xi')]^2 - (V(\xi) - R(\xi))^2 | \xi(0) = \xi, \xi(1) = \xi'\Big]$$

From Bellman's equation :
$V(\xi) = R(\xi) + \sum_{\xi'} p(\xi'|\xi).\gamma^{\tau(\xi)}V(\xi')$ we deduce that :
$E\left[[\gamma^{\tau(\xi)}V(\xi')]^2 - (V(\xi) - R(\xi))^2 | \xi(0) = \xi, \xi(1) = \xi'\right] = e(\xi)$
with :

$$e(\xi) = \sum_{\xi'} p(\xi'|\xi).\left[\gamma^{\tau(\xi)}V(\xi') - V(\xi) + R(\xi)\right]^2 \qquad (10)$$

and also that :
$\gamma^{\tau(\xi)}V(\xi') = E[J(\{\xi(t)\}_{t\geq 1}) | \xi(0) = \xi, \xi(1) = \xi']$, thus that :
$E\left[J(\{\xi(t)\}_{t\geq 1})^2 - [\gamma^{\tau(\xi)}V(\xi')]^2 | \xi(0) = \xi, \xi(1) = \xi'\right] =$
$E\left[[J(\{\xi(t)\}_{t\geq 1}) - \gamma^{\tau(\xi)}V(\xi')]^2 | \xi(1) = \xi'\right] = \gamma^{2\tau(\xi)}.\sigma^2(\xi')$

We then deduce that the variance satisfies the Bellman equation :

$$\sigma^2(\xi) = \gamma^{2\tau(\xi)} \sum_{\xi'} p(\xi'|\xi).\sigma^2(\xi') + e(\xi) \qquad (11)$$

and it can be solved by value iteration. The variance $\sigma^2(\xi)$ is equal to the 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.

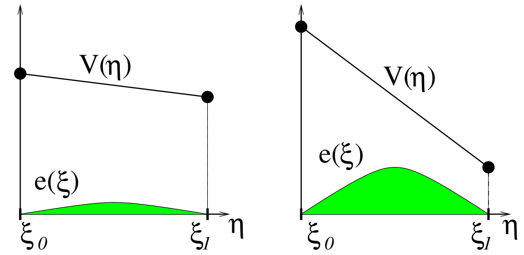


Figure 6: The term $e(\xi)$ as a function of the interpolated point η for low-(left) and high-(right) gradient value functions.

Remark 3 *We can give a geometric interpretation of the term $e(\xi)$ related to the gradient of the value function at the iterated point $\eta = \eta(\xi, u^*)$ (see figure 1) and to the barycentric coordinates $\lambda_{\xi_i}(\eta)$. Indeed, from the definition of the discretized MDP (section 3.1), we have $V(\xi) = R(\xi) + \gamma^{\tau(\xi)}V(\eta)$ and from the linearity of the interpolation we have $V(\xi_i) = V(\eta) + DV(\eta).(\xi_i - \eta)$, thus : $e(\xi) = \sum_{\xi_i} \lambda_{\xi_i}(\eta).\gamma^{2\tau(\xi)}[DV(\eta).(\xi_i - \eta)]^2$, which can be expressed as :*

$$e(\xi) = \gamma^{2\tau(\xi)}.DV(\eta)^T.Q(\eta).DV(\eta)$$

with the matrix $Q(\eta)$ defined by its elements $q_{jk}(\eta) = \sum_{\xi_i} \lambda_{\xi_i}(\eta).(\xi_i - \eta)_j.(\xi_i - \eta)_k$. Thus, $e(\xi)$ is close to 0 in two specific cases : when the gradient at the iterated point η is low (i.e. the values are almost constant) and when η is close to a 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, thus $Q(\eta)$ is low). In both cases, $e(\xi)$ is low and implies that the interpolation at ξ does not introduce a high bias in the approximated value function. Figure 6 shows $e(\xi)$ on a one-dimensional space.

4.1 Variance of the "Car on the Hill"

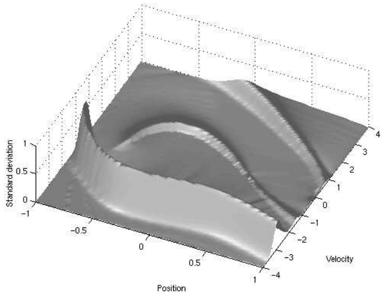


Figure 7: The standard deviation of the "Car on the Hill" problem for a uniform grid of 257 by 257

Figure 7 shows the standard deviation, square root of the variance, of the "Car on the Hill". We notice that it is very high around the discontinuity of the value function (indeed, a discontinuity is impossible to approximate perfectly by discretization methods, whatever the resolution is) and noticeably high around the discontinuities of the gradient of V (which correspond to boundaries of change in the optimal control, as shown by figure 4(a)). Indeed, in these areas the VF averages heterogeneous reinforcements.

Remark 4 *The continuous process itself has a zero variance since it is deterministic. The variance computed here* *reflects the bias introduced by the grid approximation as well as the discretization method used here, therefore it shows the parts of the discretization that would decrease the approximation error when refined.*

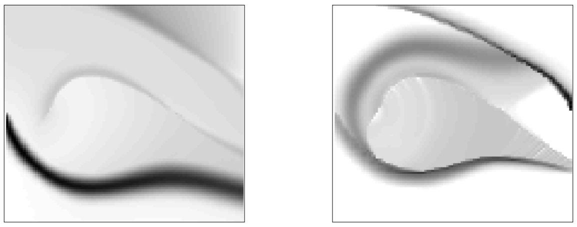
Thus it appears that the areas where a splitting might affect the most the approximation of the value function are the cells whose corners have the highest standard deviation.

5 A global splitting heuristic

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

- The states ξ of highest standard deviation $\sigma(\xi)$ are the states affected the most by the bias introduced by the grid-approximation, thus the states that could decrease the most their approximation error when split (see figure 8(a) for an illustration).
- The states ξ of highest influence $I(\xi|\Omega)$ (see figure 5(b)) on Ω, the set of states of policy disagreement (figure 5(a)) are the states whose value function affects the area of change in the optimal control.

Thus, in order to improve the quality of approximation at the most relevant areas of the state-space for the controller (i.e. at the boundary of switch in the optimal control) our heuristic is to split the states ξ with highest values of $\sigma(\xi).I(\xi|\Omega)$.



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

Figure 8: (a) The standard deviation $\sigma(\xi)$ (equivalent to figure 7). (b) The global criterion : $\sigma(\xi).I(\xi|\Omega)$, product of the two functions shown in figures 8(a) and figure 5(b)

5.1 A variable resolution discretization

In [8], we describe in detail a top-down approach to learning variable resolution discretizations, which will simply be summarized here. We start with a initial coarse discretization of the state-space, build the corresponding discrete MDP (see section 3.1), solve it, and compute the influence and the variance of the Markov chain associated to the optimal policy. Then we locally refine the discretization by splitting (a given rate of) cells whose corners are of highest $\sigma(\xi).I(\xi|\Omega)$ criterion (see figure 9).

Note that this variable resolution discretization is self-bootstrapping : it does not need to be formed by beginning with a high-resolution discretization for which influence and variance are calculated. Instead, we build the discretization top-down (reminiscent of the ID3 decision tree supervised learning method [10]).

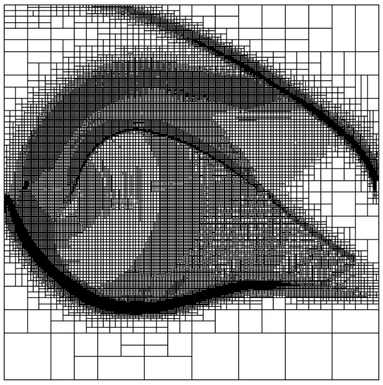


Figure 9: The discretization resulting of the splitting of the cells of highest $\sigma(\xi).I(\xi|\Omega)$ values. It is obtained by splitting 18 times a uniform grid of 9 by 9 states. Each iteration splits those 30% of cells with highest $\sigma(\xi).I(\xi|\Omega)$ values.

5.2 Comparison of the performance

To compare this adaptive resolution discretization to uniform grids, we ran a set (here 256) of trajectories starting from initial states regularly situated in the state-space, using the policy resulting from the discretizations. The *performance* is defined as the sum of the gain (defined by equation (7)) of these trajectories. Figure 10 shows the respective performance.

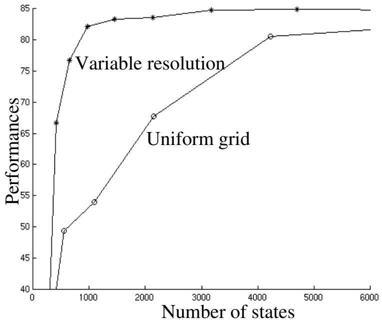


Figure 10: The performance for the uniform versus variable resolution grids as a function of the number of states of the discrete MDP.

In [8], we give a detailed empirical study of several splitting criteria and compare the relative performance for the "Car on the Hill" as well as other control problems in higher dimension (the "Acrobot", the "Cart-Pole"). The discretizations obtained by the method described in this paper give the best results.

6 More complex control problems

The "space-shuttle" control problem. We consider the 4-dimensional "space-shuttle" control problem defined by the position (x,y) and velocity (v_x,v_y) of a point (the shuttle) in a 2d-plane. There are 5 possible controls : do nothing or thrust to one of the 4 cardinal directions. The dynamics follow the laws of Newtonian physics where the shuttle is attracted by the gravitation of a planet (dark gray circle in figure 11) and some intergalactic dust (light gray circle). The goal is to reach some position in space (the square) by minimizing a cost (function of the time to reach the target and the fuel consumption). Figure 11 shows some trajectories.

The "airplane meeting" control problem. This is also a 4-dimensional control problem in which we consider one (or several) airplane(s) flying at constant altitude and velocity. They try to reach a target defined by a position

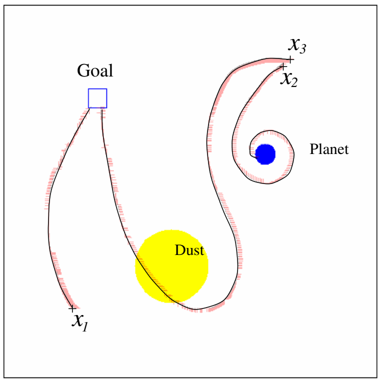
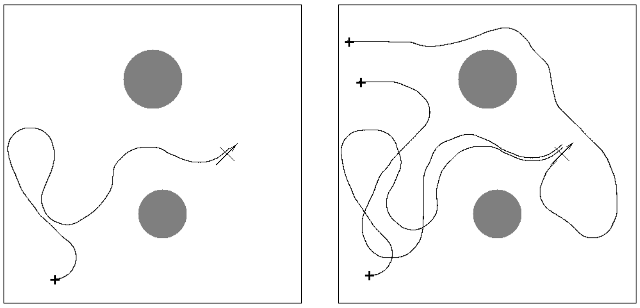


Figure 11: The "space-shuttle" trajectories for 3 different starting positions. From x_1 the goal is directly reachable (the gravitation is low). From x_2 the collision is unavoidable whatever the thrust (represented by small gray segments) to avoid the planet is. From x_3 the controller uses the gravitation forces to reach the goal.

x_G, y_G and an angle θ_G (the arrow in figure 12) at a *precise time* t_G. Each plane is defined at any time t by its position $x(t), y(t)$ and angle $\theta(t)$. There are 3 possible controls for each plane : turn left or right or go straight. The state space is of dimension 4 : the position x, y, the angle θ and the time t. The dynamics are : $\frac{dx}{dt} = \cos(\theta)$, $\frac{dy}{dt} = \sin(\theta)$, $\frac{d\theta}{dt} = \{-1, 0, +1\}.v_\theta$ and $\frac{dt}{dt} = 1$. Here, the terminal cost is : $(x - x_G)^2 + (y - y_G)^2 + k_\theta(\theta - \theta_G)^2 + k_t(t - t_G)^2$ and there is a small constant current cost if a plane is in a gray area (some clouds that the planes should avoid). Figure 12 shows some trajectories for one and 3 planes when there is more time than necessary to reach the target directly (the planes have to loop).



(a) One airplaine (b) Meeting of 3 airplanes

Figure 12: The "airplane meeting" control problem

The variable approach described in this paper is critical for these 4d complex problems : other splitting methods that do not take into account the global influence of the splitting process (some of which are described in [8]) as well as uniform grids fail to provide an accurate controller.

7 Conclusion and Future work

In this paper we introduced two useful measures of a Markov chain, influence and variance, and combined them to propose an efficient splitting heuristic that locally refines a discretization for continuous control problems. These measure could be used to solve large (discrete) MDPs by selecting which initial (coarse) features (or categories) one has to refine to provide a relevant partition of the state space.

Another extension could be to learn these measures through interactions with the environment in order to design efficient exploration policies in reinforcement learning. Our notion of variance could be used with "Interval Estimation" heuristic [3], to permit "optimism-in-the-face-of-uncertainty" exploration, or with the "back-propagation of exploration bonuses" of [5] for exploration in continuous state-spaces. Indeed, if we observe that the learned variance of a state ξ is high, then a good exploration strategy could be to inspect the states that have a high expected influence on ξ.

Also, the notion of variance might be useful to provide a safe controller for which choosing a sub-optimal action would be preferable if it leads to states of lower variance than when taking the optimal action.

Finally, our next focus will be to consider stochastic control problems (Markov Diffusion Processes) for which our splitting method will have to be reconsidered since in that case the variance would reflect two components : the bias introduced by the grid-approximation but also the intrinsic stochasticity of the continuous process. The latter is not relevant to our splitting method since a refinement around areas of high variance of the process will not result in an improvement of the approximations.

Acknowledgements To Csaba Szepesvári, Geoffrey Gordon, Jeff Schneider, Nicolas Meuleau for their very useful comments and advice.

References

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

[2] Geoffrey J. Gordon. *Approximate solutions to Markov Decision Processes*. PhD thesis, Carnegie Mellon University, 1999.

[3] Leslie P. Kaelbling. *Learning in Embedded Systems*. MIT Press, Cambridge MA, 1993.

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

[5] Nicolas Meuleau and Paul Bourgine. Exploration of multi-state environments: Local measures and back-propagation of uncertainty. *To appear in Machine Learning Journal*, 1999.

[6] Andrew W. Moore and C.G. Atkeson. The parti-game algorithm for variable resolution reinforcement learning in multidimensional state space. *Machine Learning Journal*, 21, 1995.

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

[8] Rémi Munos and Andrew Moore. Variable resolution discretization for high-accuracy solutions of optimal control problems. *International Joint Conference on Artificial Intelligence*, 1999.

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

[10] J. R. Quinlan. Learning Efficient Classification Procedures and their Application to Chess End Games. In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, editors, *Machine Learning—An Artificial Intelligence Approach (I)*. Tioga Publishing Company, Palo Alto, 1983.

[11] Michael A. Trick and Stanley E. Zin. A linear programming approach to solving stochastic dynamic programs. *Unpublished manuscript*, 1993.