A convergent Reinforcement Learning algorithm in the continuous case based on a Finite Difference method

Rémi Munos*
CEMAGREF, LISC, Parc de Tourvoie,
BP 121, 92185 Antony Cedex, FRANCE.
Tel : (0)1 40 96 61 79. Fax : (0)1 40 96 60 80
E-mail : Remi.Munos@cemagref.fr

Abstract

In this paper, we propose a convergent Reinforcement Learning algorithm for solving optimal control problems for which the state space and the time are continuous variables.

The problem of computing a good approximation of the value function, which is essential because this provides the optimal control, is a difficult task in the continuous case. Indeed, as it has been pointed out by several authors, the use of parameterized functions such as neural networks for approximating the value function may produce very bad results and even diverge.

In fact, we show that classical algorithms, like Q-learning, used with a simple look-up table built on a regular grid, may fail to converge. The main reason is that the discretization of the state space implies a lost of the Markov property even for deterministic continuous processes.

We propose to approximate the value function with a convergent numerical scheme based on a Finite Difference approximation of the Hamilton-Jacobi-Bellman equation. Then we present a model-free reinforcement learning algorithm, called *Finite Difference Reinforcement Learning*, and prove its convergence to the value function of the continuous problem.

1 Introduction

This paper is concerned with convergence results of Reinforcement Learning (RL) algorithms in the continuous-time, continuous-state-space case. We discuss the problem of the necessary discretization of the state space and propose a RL algorithm that converges to the optimal solution.

*DASSAULT-AVIATION, DGT-DTN-EL-Et. Avancées, 78 quai Marcel Dassault, 92214 Saint-Cloud, FRANCE

The objective of RL is to find -thanks to a reinforcement signal- an optimal strategy for solving a dynamical control problem, such as target or obstacle problems, viability or optimization problems. The technique used belongs to the Dynamic Programming (DP) methods which define an optimal feed-back control by approximating the *value function* (VF), which is the best expected cumulative reinforcement as a function of initial state.

In the continuous case, the VF has to be represented with a function approximator with a finite number of parameters. Several author have pointed out that the combination of RL algorithms with general approximation systems (such as neural networks, fuzzy sets, polynomial approximators, etc.) may produce unstable or divergent results even for very simple problems (see [Boyan and Moore, 1995], [Baird, 1995], [Gordon, 1995]). Here we show that classical RL algorithms, like Q-learning (see [Watkins, 1989]), used with a look-up table built from a simple discretization of the state space may produce a very bad approximation of the value function. The main reason is that the discretization of a deterministic continuous process is not Markovian. So algorithms such as Q-learning which estimate the value of a state as an average of the values of successive states according to their occurrence will not converge. We propose as an alternative an algorithm that averages the values of the next states according to the state dynamics.

Section 2 proposes a formalism for optimal control problems in the continuous case. The VF is introduced and the Hamilton-Jacobi-Bellman (HJB) equation is stated. *Section 3* discusses the lost of the Markov property with the discretization of the state space and studies the Q-learning algorithm with a look-up table. *Section 4* describes the discretization of the HJB equation by a Finite Difference (FD) method, which leads to a DP equation for a finite Markov Decision Process (MDP) and whose solution approximates the VF. *Section 5* presents the algorithm, called *Finite Difference Reinforcement Learning* (FDRL), that converges to the value function of the continuous process. *Appendix A* gives the proof of convergence of the algorithm.

2 A formalism for Reinforcement Learning in the continuous case

In this paper, we consider *deterministic* controlled systems with *infinite time horizon* and *discounted reinforcement*. Let $x(t) \in \bar{O}$ be the state of the system with $O \subset \mathbb{R}^d$ an open and bounded subset. The evolution of the system (the *state dynamics* f) depends on the *current state* $x(t)$ and *control* $u(t)$; it is defined by a controlled differential equation :

$$\frac{d}{dt}x(t) = f(x(t), u(t)) \qquad (1)$$

where the control $u(t)$ is a bounded, Lebesgue measurable function with values in a compact U.

From any initial state x, the choice of a control $u(t)$ leads to a unique *trajectory* $x(t)$. Let τ be the *exit time* of $x(t)$ from $\bar{O}$ (with the convention that if $x(t)$ always stays in $\bar{O}$, then $\tau = \infty$). Then, we define the discounted reinforcement functional of state x, control $u(.)$:

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

Where $r(x, u)$ is the *running reinforcement* and $R(x)$ the *boundary reinforcement*. γ is the *discount factor* ($0 \leq \gamma < 1$).

The **objective of the control problem** is to find the optimal feed-back control $u^*(x)$ that optimizes the reinforcement functional for any state x.

2.1 The Reinforcement Learning approach

RL techniques belongs to the class of DP methods which compute the optimal control by the means of the *value function*, which is the maximum value of the functional as a function of initial state x :

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

In the RL approach, the system tries to approximate this function without knowing the state dynamics f nor the reinforcement functions r, R. RL appears as a constructive and iterative process, based on experience, that estimates the value function by successive approximations.

2.2 The Hamilton-Jacobi-Bellman equation

Following the dynamic programming principle, the value function satisfies a first-order nonlinear partial differential equation called the *Hamilton-Jacobi-Bellman* equation (see [Fleming and Soner, 1993] for a survey).

Theorem 1 (Hamilton-Jacobi-Bellman) *If V is differentiable at $x \in O$, let $DV(x)$ be the gradient of V at x, then the following HJB equation holds at x.*

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

Besides, V satisfies the following boundary condition :

$$V(x) \geq R(x) \text{ for } x \in \partial O$$

Remark 1 *The challenge of learning the value function is motivated by the fact that from V, we can deduce the following optimal feed-back control policy:*

$$u^*(x) = \arg\sup_{u \in U}[DV(x).f(x, u) + r(x, u)]$$

In the following, we assume that:
- f and r are bounded with M_f (respectively M_r) and Lipschitzian: $|f(x, u) - f(y, u)| \leq L_f \|x - y\|_1$ (resp. $|r(x, u) - r(y, u)| \leq L_r \|x - y\|_1$).
- R is Lipschitzian: $|R(x) - R(y)| \leq L_R \|x - y\|_1$.
with the norm $||x||_1 = \sum_{i=1}^d |x_i|$.
- The boundary ∂O is $\mathcal{C}^2$.

Besides, we consider the following hypothesis concerning the state dynamics f around the boundary ∂O and we state a theorem of continuity whose proof is in [Barles and Perthame, 1990]. For all $x \in \partial O$, let $\vec{n}(x)$ be the outward normal of O at x, we assume that:
- *If* there exists $u \in U$, such that $f(x, u).\vec{n}(x) \leq 0$ *then* there exists $v \in U$, such that $f(x, v)\vec{n}(x) < 0$.
- *If* there exists $u \in U$, such that $f(x, u).\vec{n}(x) \geq 0$ *then* there exists $v \in U$, such that $f(x, v)\vec{n}(x) > 0$.

Theorem 2 *Suppose that these hypotheses hold, then the value function is continuous in O.*

3 The discretization implies a lost of the Markov property

Let us discretize the state space into a regular grid and define a finite discretized state space composed of cells X. Consider a trajectory $x(t)$ and the corresponding sequence of cells X_i containing it: $x(t) \in X_i$ for $t \in [t_i, t_{i+1}]$. Suppose that the control u_i is kept constant inside X_i. We observe that the transition from a cell X to an adjacent cell X_1 not only depends on X but also on the place from which the trajectory enters inside X (see figure 1). Thus, in general, the succession of cells X_i does not provide a MDP even when the continuous process is deterministic.

Q-learning with a look-up table:

RL algorithms such as Q-learning (see [Watkins, 1989]) are classically used in order to approximate the value function. Here, the updating rule could be:

$$\begin{aligned} \Delta Q(X_i, u) &= \alpha_i [\gamma^{\tau_i} V(X_{i+1}) - Q(X_i, u) + \tau_i r(X_i)] \\ \text{with } V(X) &= \sup_{u \in U} Q(X, u) \end{aligned}$$

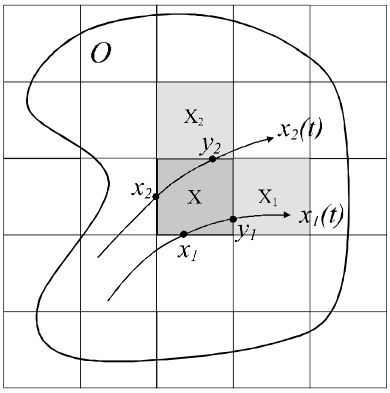

Figure 1: *Discretization of the state space.* The serie of the successive cells X_i containing the current state $x(t)$ is not a MDP. Here, the transition from X to X_1 or X_2 (occuring for some trajectories $x_1(t)$ and $x_2(t)$) depends on the input points (x_1 and x_2).

for some time τ_i and some decreasing learning rate α_i.

This is an iterative recursive equation that incrementally computes the average of the values of successive cells according to their occurrence. Meanwhile, from the non-Markovian aspect of the succession of cells, in general, this algorithm (whose convergence is proved for finite MDP) has no chance to converge. For example, suppose that in figure 1, most of the trajectories comes inside X from its bottom side thus leaves X from its right side. Then from the algorithm, $Q(X,u)$ will almost exclusively depend on $V(X_1)$. In fact, the values computed by such algorithms will depend on the exploration strategies and will not converge to the value function.

As the Q-learning with look-up table do not converge, the combination of similar RL algorithms with function approximators appears unlikely to converge. As an alternative, we propose that the updating rule should take into account the state dynamics f in order to approximate a convergent FD scheme studied next section.

4 A Finite Difference scheme

Let $e_1, e_2, ..., e_d$ be a basis for $\mathbb{R}^d$. The state dynamics is : $f = (f_1, ..., f_d)$. Let the positive and negative parts of f_i be : $f_i^+ = \max(f_i, 0)$, $f_i^- = \max(-f_i, 0)$. For any discretization step δ, let us consider the lattices : $\delta\mathbb{Z}^d = \left\{\delta.\sum_{i=1}^d j_i e_i\right\}$ where $j_1, ..., j_d$ are any integers, and $\Sigma^\delta = \{\xi \in \delta\mathbb{Z}^d$ such that at least one adjacent points $\xi \pm \delta e_i \in O\}$. The *interior* of Σ^δ is $\Sigma^\delta \cap O$. Let $\partial\Sigma^\delta$, the *frontier* of Σ^δ, denote the set of points of Σ^δ which are not in the interior of Σ^δ.

Let $U^\delta \subset U$ be a finite control set that approximates U in the sense : $\delta \leq \delta' \Rightarrow U^{\delta'} \subset U^\delta$ and $\overline{\cup_\delta U^\delta} = U$.

By replacing the gradient $DV(\xi)$ by the forward and backward difference quotients of V in ξ :

$$\begin{aligned}\Delta_i^+ V(\xi) &= \frac{1}{\delta}\left[V(\xi + \delta e_i) - V(\xi)\right] \\ \Delta_i^- V(\xi) &= \frac{1}{\delta}\left[V(\xi - \delta e_i) - V(\xi)\right]\end{aligned}$$

we can approximate the HJB equation by the following equation :

$$\begin{aligned}V^\delta(\xi)\ln\gamma \;+\; \sup_{u\in U^\delta}&\left\{\sum_{i=1}^d \left[f_i^+(\xi,u).\Delta_i^+ V^\delta(\xi)\right.\right. \\ &\left.\left.+ f_i^-(\xi,u).\Delta_i^- V^\delta(\xi)\right] + r(\xi,u)\right\} = 0 \quad (2)\end{aligned}$$

Knowing that $(\Delta t \ln\gamma)$ is an approximation of $(\gamma^{\Delta t} - 1)$ as Δt tends to 0, and by introducing the qualities $Q^\delta(\xi,u)$ such that $V^\delta(\xi) = \sup_{u\in U^\delta} Q^\delta(\xi,u)$, (2) gives :

$$\begin{aligned}Q^\delta(\xi,u) &= \gamma^{\frac{\delta}{||f(\xi,u)||_1}} \sum_{i=1}^d \frac{|f_i(\xi,u)|}{||f(\xi,u)||_1}.V^\delta(\xi_i) \\ &\quad + \frac{\delta}{||f(\xi,u)||_1} r(\xi,u) \qquad (3)\end{aligned}$$

with $\xi_i = \xi + \delta e_i$ if $f_i(\xi,u) > 0$ and $\xi_i = \xi - \delta e_i$ if $f_i(\xi,u) < 0$.

This equation can be interpreted as a DP equation for a finite MDP (see [Fleming and Soner, 1993]) whose *state space* is Σ^δ, the *control space* is U^δ and the *probabilities of transition* $p(\xi,u,\xi_i)$ from state ξ, control u to next state ξ_i are the coordinates $\frac{|f_i(\xi,u)|}{||f(\xi,u)||_1}$ (see figure 2 for a geometrical interpretation).

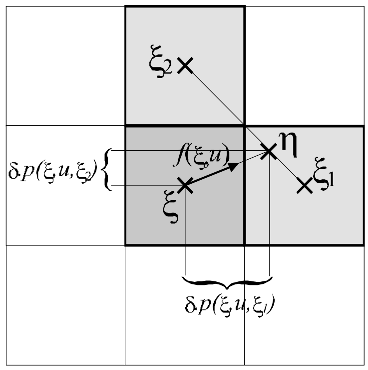

Figure 2: *A geometrical interpretation of the FD scheme.* The transition probabilities $p(\xi,u,\xi_i)$ of the corresponding MDP are the coordinates of the vector $\frac{1}{\delta}\xi\eta$ with η the projection of ξ onto the segment $(\xi_1\xi_2)$ in a direction parallel to $f(\xi,u)$.

Thanks to a contraction property due to the discount factor γ (see [Bertsekas, 1987]), DP theory insures that there exists a unique solution (the fixed-point) V^δ to equation (3) for ξ interior to Σ^δ with the boundary condition $V^\delta(\xi) = R(\xi)$ for $\xi \in \partial\Sigma^\delta$.

The following theorem, whose proof uses the general convergence result of [Barles and Souganidis, 1991] and the strong comparison result between sub- and superviscosity solution (see [Crandall *et al.*, 1992]) of HJB equations of [Barles, 1994], insures that V^δ is a convergent approximation of V.

Theorem 3 (Convergence of the FD scheme) *Let us assume that the hypotheses of section 2 hold, then V^δ converges to V as δ tends to 0 :*

$$\lim_{\substack{\delta\downarrow 0 \\ \xi\to x}} V^\delta(\xi) = V(x) \text{ uniformly on any compact } \Omega \subset O$$

Remark 2 *The MDP built here depends on the state dynamics f and is independent of the running of trajectories which generates a non-Markov succession of cells as it has been seen in section 3.*

In the next section, we propose a RL algorithm that approximates this FD scheme.

5 The FDRL algorithm

Let us consider a grid $G^\delta \supset O$ composed of regular cells X such that the center of the cells are the previously defined vertices ξ of the lattice. Let $\partial G^\delta = \{X \in G^\delta, X \cap \partial O \neq \emptyset\}$ its boundary. Let $Q_n^\delta(X,u)$ and $V_n^\delta(X)$, the values of cell X computed by the algorithm at stage n, intend to approximate $Q^\delta(\xi,u)$ and $V^\delta(\xi)$. We have the relation $V_n^\delta(X) = \sup_u Q_n^\delta(X,u)$. Let $\lambda > 0$ be any (small) constant.

Let a trajectory $x_i(t)$ enters a cell X at some point x_i ; then a control $u \in U^\delta$ is chosen and kept constant.

- If $X \notin \partial G^\delta$ then the trajectory exits at some point $y_i \in X \cap X_i$ for some adjacent cell X_i (see figure 1). Let $r(z_i,u)$ be the current reinforcement obtained at some point z_i of the trajectory inside X. Let τ_i be the running time of the trajectory inside X.

The algorithm is inspired by the DP equation (3) for which $f(\xi,u)$ is approximated by $\frac{\overrightarrow{x_iy_i}}{\tau_i}$:

If $\|\overrightarrow{x_iy_i}\|_1 \geq \lambda.\delta$ then update some value $Q_n^\delta(X,u,X_i)$:

$$Q_n^\delta(X,u,X_i) = \gamma^{\frac{\delta.\tau_i}{\|\overrightarrow{x_iy_i}\|_1}} V_n^\delta(X_i) + \frac{\delta.\tau_i}{\|\overrightarrow{x_iy_i}\|_1} r(z_i,u) \quad (4)$$

Then we consider an other trajectory $x_j(t)$ going through X from x_j till $y_j \in X \cap X_j$ with control u, which leads to update $Q_n^\delta(X,u,X_j)$, and the process is repeated until all possible transitions $(X,u) \to X_k$ for $k = 1..d$ are carried out at least once. Then we consider a vector $\overrightarrow{xy} = \overrightarrow{x_ky_k}$ for some k (for example, corresponding to the most recent trajectory) and compute :

$$Q_{n+1}^\delta(X,u) = \sum_{i=1}^{d} \frac{|(\overrightarrow{xy})_i|}{\|\overrightarrow{xy}\|_1} Q_n^\delta(X,u,X_i) \quad (5)$$

where $(.)_i$ denote the i^{th} coordinate. For example, in figure 1, if $x_1(t)$ occurs first, $Q_n^\delta(X,u,X_1)$ is computed, then when $x_2(t)$ occurs, $Q_n^\delta(X,u,X_2)$ and $Q_{n+1}^\delta(X,u)$ are updated).

- If $X \in \partial G^\delta$ and if the trajectory reaches the boundary $y_i \in \partial O$ inside X, update :

$$V_{n+1}^\delta(X) = R(y_i) \quad (6)$$

The following theorem states that with the following hypothesis of *exploring-every-possible-transition :*

> We consider series of trajectories such that the FDRL algorithm leads to the updating of every cell $X \notin \partial G^\delta$ with rule (4) for all control u and all possible successive cell X_k any (finite) number of times and every cell $X \in \partial G^\delta$ with rule (6) at least once.

the values computed by the algorithm converge to the VF of the continuous problem :

Theorem 4 (Convergence of the algorithm) *Suppose that the hypotheses of section 2 and the one of exploring-every-possible-transition hold, then:*
For any compact $\Omega \subset O, \forall\varepsilon > 0$, $\exists\Delta$ st. $\forall\delta \leq \Delta$, by using the FDRL algorithm, $\exists N, \forall n \geq N$,

$$\sup_{x\in\Omega} |V_n^\delta(X \ni x) - V(x)| \leq \varepsilon.$$

Remark 3 *The FDRL is a kind of Real Time (or asynchronous) DP and not a kind of Q-learning (there is no learning rate α_n that averages the successive values according to their occurrence). This seems more relevant here because the averaging (with rule (5)) of the values of next cells comes from the discretization process itself and not from the state dynamics of the continuous process which is deterministic.*

Remark 4 *Once the $Q_n^\delta(X,u)$ values have been computed, the current optimal control in cell X is:*
$u^* = \arg\sup_{u\in U^\delta} Q_n^\delta(X,u)$.

6 Conclusion

This paper uses FD methods for approximating the VF and generating a convergent RL algorithm. It need to be compared to the Finite Element method used in [Munos, 1996] that approximates the VF with piecewise linear functions defined on a triangulation of the state space.

In practical use of this algorithm, and in general, for all approximation systems of continuous functions, we are faced to the combinatorial explosion of the number of values to be estimated. Future work should consider adaptive multi-resolutions techniques (like the parti-game algorithm of [Moore, 1994] or the multigrid methods of [Akian, 1990]).

An other improvement should be to study the stochastic case for which a Q-learning version of FDRL could be relevant.

A Appendix: proof of theorem 4

A.1 Idea of the proof

The idea of the demonstration is to prove that for any $\varepsilon_2 > 0$, for small enough values of δ,

$$\sup_{\xi \in \Sigma^\delta} |V_n^\delta(X \ni \xi) - V^\delta(\xi)| \leq \varepsilon_2 \qquad (7)$$

Then for all $\varepsilon > 0$, we can find $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ such that $\varepsilon_1 + \varepsilon_2 = \varepsilon$ and from the convergence of the scheme V^δ (theorem 3), $\sup_{x\in\Omega} |V^\delta(x) - V(x)| \leq \varepsilon_1$ for small δ. Thus:

$$\sup_{x\in\Omega} |V_n^\delta(X \ni x) - V(x)| \leq \varepsilon_1 + \varepsilon_2 = \varepsilon$$

In order to prove (7), we need to estimate the difference $\left|Q_{n+1}^\delta(X,u) - Q^\delta(\xi,u)\right|$ after having updated $Q_n^\delta(X,u)$ with rule (5).

With $E_n^\delta = \sup_{\xi\in\Sigma^\delta} \left|V_n^\delta(X \ni \xi) - V^\delta(\xi)\right|$, we prove in *section A.2* that:

$$\left|Q_{n+1}^\delta(X,u) - Q^\delta(\xi,u)\right| \leq (1 - k.\delta)E_n^\delta + e(\delta).\delta$$

for some constant k and some function $e(\delta)$ that tends to 0. Then we give in *section A.3* a sufficient condition for $E_n^\delta \leq \varepsilon_2$ and summarize the proof in *section A.4*. In the following of this section, we give some comparisons.

Comparison of the times $\frac{\delta}{||f(\xi,u)||_1}$ **and** $\frac{\delta}{\|\overrightarrow{xy}\|_1}.\tau$

From Taylor's theorem, $||\overrightarrow{xy} - f(x,u).\tau||_1 \leq \frac{1}{2}L_f.\tau^2$. As the state dynamics is bounded from below, the time τ is bounded by $\frac{d.\delta}{m_f}$. From the Lipschitz property of f, $||f(x,u) - f(\xi,u)||_1 \leq L_f.||x-\xi||_1 \leq \frac{d.\delta}{2}L_f$. But $||\overrightarrow{xy} - f(\xi,u).\tau||_1 = ||\overrightarrow{xy} - f(x,u).\tau + \tau[f(x,u) - f(\xi,u)]||_1$, thus: $||\overrightarrow{xy} - f(\xi,u).\tau||_1 \leq \frac{d^2\delta^2}{2m_f}L_f.(1+\frac{1}{m_f})$. As $||\overrightarrow{xy}||_1 \geq \lambda.\delta$, we have:

$$\left|\frac{\delta}{\|\overrightarrow{xy}\|_1}.\tau - \frac{\delta}{\|f(x,u)\|_1}\right| \leq k_\tau \delta^2 \qquad (8)$$

with: $k_\tau = \frac{d^2}{2\lambda.m_f^2}L_f.(1+\frac{1}{m_f})$. We deduce from a property of the exponential function that:

$$\left|\gamma^{\frac{\delta}{\|\overrightarrow{xy}\|_1}.\tau} - \gamma^{\frac{\delta}{\|f(x,u)\|_1}}\right| \leq k_\tau \ln\frac{1}{\gamma}.\delta^2 \qquad (9)$$

Comparison of $\frac{(\overrightarrow{xy})_i}{||xy||_1}$ **and** $\frac{f_i(\xi,u)}{||f(\xi,u)||}$:

From the fact that for any couple of vector a and b such that $||a - k.b|| < \varepsilon$, we have $|\frac{|a_i|}{||a||} - \frac{|b_i|}{||b||}| \leq \frac{2\varepsilon}{k.|b_i|-2\varepsilon}$, we deduce from (8) that

$$\begin{aligned}\left|\frac{|(\overrightarrow{xy})_i|}{\|\overrightarrow{xy}\|_1} - \frac{|f_i(\xi,u)|}{\|f(\xi,u)\|_1}\right| &\leq \frac{2m_f.k_\tau.\delta}{1-2m_f.k_\tau.\delta} \\ &\leq 4m_f.k_\tau.\delta \qquad (10)\end{aligned}$$

for any $\delta \leq \Delta_1 = \frac{1}{4m_f.k_\tau}$.

A.2 Estimation of $|Q_{n+1}^\delta(X,u) - Q^\delta(\xi,u)|$

After having updated $Q_n^\delta(X,u)$ with rule (5), let Λ denote the difference $\left|Q_{n+1}^\delta(X,u) - Q^\delta(\xi,u)\right|$. From (3), (4) and (5),

$$\begin{aligned}\Lambda \leq\ & \left|\gamma^{\frac{\delta}{||f(\xi,u)||_1}}.\sum_{i=1}^d\left(\frac{|f_i(\xi,u)|}{||f(\xi,u)||_1} - \frac{|(\overrightarrow{xy})_i|}{\|\overrightarrow{xy}\|_1}\right)V^\delta(\xi_i)\right| \\ &+ \left|\left(\gamma^{\frac{\delta}{||f(\xi,u)||_1}} - \gamma^{\frac{\delta}{\|\overrightarrow{xy}\|_1}.\tau}\right).\sum_{i=1}^d\frac{|(\overrightarrow{xy})_i|}{\|\overrightarrow{xy}\|_1}V^\delta(\xi_i)\right| \\ &+ \left|\gamma^{\frac{\delta}{\|\overrightarrow{xy}\|_1}.\tau}.\sum_{i=1}^d\frac{|(\overrightarrow{xy})_i|}{\|\overrightarrow{xy}\|_1}\left[V^\delta(\xi_i) - V_n^\delta(\xi_i)\right]\right| \\ &+ \sum_{i=1}^d\frac{|(\overrightarrow{xy})_i|}{\|\overrightarrow{xy}\|_1}\frac{\delta.\tau_i}{\|\overrightarrow{x_iy_i}\|_1}|r(\xi,u) - r(z_i,u)| \\ &+ \sum_{i=1}^d\frac{|(\overrightarrow{xy})_i|}{\|\overrightarrow{xy}\|_1}\left|\left(\frac{\delta.\tau_i}{\|\overrightarrow{x_iy_i}\|_1} - \frac{\delta}{||f(\xi,u)||_1}\right)r(\xi,u)\right|\end{aligned}$$

Let us consider a linear function $\widetilde{V} : \mathbb{R}^d \to \mathbb{R}$ such that $\widetilde{V}(\xi_i) = V^\delta(\xi_i)$ for $i = 1..d$. As $\frac{|f_i(\xi,u)|}{||f(\xi,u)||_1}$ and $\frac{|(\overrightarrow{xy})_i|}{\|\overrightarrow{xy}\|_1}$ can be interpreted as barycentric coordinates, we have:

$$\sum_{i=1}^d\left(\frac{|f_i(\xi,u)|}{||f(\xi,u)||_1} - \frac{|(\overrightarrow{xy})_i|}{\|\overrightarrow{xy}\|_1}\right)V^\delta(\xi_i) = \widetilde{V}(\eta) - \widetilde{V}(\eta')$$

with $\eta = \xi + \delta.\frac{f(\xi,u)}{||f(\xi,u)||_1}$ (see figure 2) and $\eta' = \xi + \delta.\frac{\overrightarrow{xy}}{\|\overrightarrow{xy}\|_1}$.

Geometrical considerations give: $\exists\ k,l \in [1,d]$ st.:

$$\left|\widetilde{V}(\eta) - \widetilde{V}(\eta')\right| \leq d.\frac{||\eta - \eta'||_1.|V^\delta(\xi_k) - V^\delta(\xi_l)|}{||\xi_k - \xi_l||_1} \qquad (11)$$

Moreover,

$$\begin{aligned}|V^\delta(\xi_k) - V^\delta(\xi_l)| \leq\ & |V^\delta(\xi_k) - V(\xi_k)| \\ &+ |V(\xi_k) - V(\xi_l)| + |V(\xi_l) - V^\delta(\xi_l)|\end{aligned}$$

Let $E_\Omega^\delta = \sup_{x\in\Omega}|V^\delta(x) - V(x)|$.

Thanks to the continuity of V, there exists Δ_2, for $\delta \leq \Delta_2$, $|V(\xi_k) - V(\xi_l)| \leq E_\Omega^\delta$. Thus, $|V^\delta(\xi_k) - V^\delta(\xi_l)| \leq 3E_\Omega^\delta$. From (11) and (10), we have:

$$\left|\sum_{i=1}^d\left(\frac{|f_i(\xi,u)|}{||f(\xi,u)||_1} - \frac{|(\overrightarrow{xy})_i|}{\|\overrightarrow{xy}\|_1}\right)V^\delta(\xi_i)\right| \leq 6.d^2.m_f.k_\tau.E_\Omega^\delta.\delta$$

Thus, from (9), (8) and the Lipschitz property of r :

$$\begin{aligned}\Lambda \leq\ & 6d^2m_fk_\tau E_\Omega^\delta\delta + k_\tau\ln\frac{1}{\gamma}\delta^2M_{V^\delta} + \frac{\delta^2}{m_f}L_r + k_\tau L_r\delta^3 \\ &+ k_\tau\delta^2M_r + \left|\gamma^{\frac{\delta}{\|\overrightarrow{xy}\|_1}.\tau}.\sum_{i=1}^d\frac{(\overrightarrow{xy})_i}{\|\overrightarrow{xy}\|_1}\left[V^\delta(\xi_i) - V_n^\delta(\xi_i)\right]\right|\end{aligned}$$

Let $E_n^\delta = \sup_{\xi\in\Sigma^\delta} \left|V_n^\delta(X \ni \xi) - V^\delta(\xi)\right|$.

As $\gamma^{\frac{\delta}{\|x\dot{y}\|_1}.\tau} \leq 1 - \frac{\delta.\tau}{\|x\dot{y}\|_1}\ln\frac{1}{\gamma} \leq 1 - (\frac{\delta}{M_f} - k_\tau\delta^2)\ln\frac{1}{\gamma}$, we have:

$$\Lambda \leq (1-k.\delta)E_n^\delta + e(\delta).\delta \qquad (12)$$

with $k = \frac{1}{M_f}\ln\frac{1}{\gamma}$ and $e(\delta) = 6.d^2.m_f.k_\tau.E_\Omega^\delta + k_\tau\ln\frac{1}{\gamma}.\delta.M_{V^\delta} + \frac{\delta}{m_f}L_r + k_\tau L_r\delta^3 + k_\tau\delta M_r + k_\tau\delta\ln\frac{1}{\gamma}E_n^\delta$.

A.3 Condition for $E_n^\delta \leq \varepsilon_2$

Let us suppose that for cells $X \notin \partial G^\delta$, the following conditions hold for some $\alpha > 0$

$$\begin{array}{lll} E_n^\delta & > & \varepsilon_2 \Rightarrow \left|Q_{n+1}^\delta(X,u) - Q^\delta(\xi,u)\right| \leq E_n^\delta - \alpha \quad (13) \\ E_n^\delta & \leq & \varepsilon_2 \Rightarrow \left|Q_{n+1}^\delta(X,u) - Q^\delta(\xi,u)\right| \leq \varepsilon_2 \quad (14) \end{array}$$

From the hypothesis of exploring-every-possible-transition, there exists an integer m such that at stage $n+m$ all the cells $X \in G^\delta$ have been updated at least once since stage n. Since cells $X \in \partial G^\delta$ are updated with rule (6), $|V_{n+1}^\delta(X) - V^\delta(\xi)| = |R(y_i) - R(\xi)| \leq L_R.\delta \leq \varepsilon_2$ for any $\delta \leq \Delta_3 = \frac{\varepsilon_2}{L_R}$. Thus, from (13) and (14) we have:

$$\begin{array}{lll} E_n^\delta & > & \varepsilon_2 \Rightarrow E_{n+m}^\delta \leq E_n^\delta - \alpha \\ E_n^\delta & \leq & \varepsilon_2 \Rightarrow E_{n+m}^\delta \leq \varepsilon_2 \end{array}$$

Thus there exists N such that : $\forall n \geq N$, $E_n^\delta \leq \varepsilon_2$.

A sufficient condition: let us suppose that

$$(1-k.\delta).\frac{\varepsilon_2}{2} + e(\delta)\delta \leq \frac{\varepsilon_2}{2} \qquad (15)$$

holds, then conditions (13) and (14) are true.

Indeed, assume (15) is true. Let $E_n^\delta > \varepsilon_2$, then from (12), $\Lambda \leq E_n^\delta - k.\delta.\varepsilon_2 + e(\delta)\delta \leq E_n^\delta - k.\delta.\frac{\varepsilon_2}{2}$. Thus (13) holds for $\alpha = k.\delta.\frac{\varepsilon_2}{2}$.

Now suppose that $E_n^\delta \leq \varepsilon_2$. From (12), $\Lambda \leq (1-k.\delta)\varepsilon_2 + e(\delta)\delta \leq \frac{\varepsilon_2}{2} + \frac{\varepsilon_2}{2}$ and condition (14) is true.

A.4 Convergence of the algorithm

Let us prove theorem 4. For any compact $\Omega \subset O$, for all $\varepsilon > 0$, let us consider $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ such that $\varepsilon_1 + \varepsilon_2 = \varepsilon$. As $e(\delta)$ tends to 0, there exists Δ_4 for $\delta \leq \Delta_4$, (15) which is equivalent to: $e(\delta) - k\frac{\varepsilon_2}{2} \leq 0$ holds.

Thus for $\delta \leq \min\{\Delta_1, \Delta_2, \Delta_3, \Delta_4\}$, the sufficient condition (15) is satisfied and conditions (13) and (14) are true. So there exists N, for all $n \geq N$, $E_n^\delta \leq \varepsilon_2$. Besides, from the convergence of the scheme (theorem (3)), there exists Δ_0 st. $\forall\delta \leq \Delta_0, \sup_{x\in\Omega}|V^\delta(x) - V(x)| \leq \varepsilon_1$.

Thus for $\delta \leq \min\{\Delta_0, \Delta_1, \Delta_2, \Delta_3, \Delta_4\}$, $\exists N, \forall n \geq N$,

$$\sup_{x\in\Omega}|V_n^\delta(X \ni x) - V(x)| \leq \sup_{\xi\in\Sigma^\delta}|V_n^\delta(X \ni \xi) - V^\delta(\xi)| + \sup_{x\in\Omega}|V^\delta(x) - V(x)| \leq \varepsilon_1 + \varepsilon_2 = \varepsilon.$$

References

[Akian, 1990] Marianne Akian. *Méthodes multigrilles en contrôle stochastique.* PhD thesis, University Paris IX Dauphine, 1990.

[Baird, 1995] Leemon C. Baird. Residual algorithms : Reinforcement learning with function approximation. *Machine Learning : proceedings of the Twelfth International Conference*, 1995.

[Barles and Perthame, 1990] Guy Barles and B. Perthame. Comparison principle for dirichlet-type hamilton-jacobi equations and singular perturbations of degenerated elliptic equations. *Applied Mathematics and Optimization*, 21:21–44, 1990.

[Barles and Souganidis, 1991] Guy Barles and P.E. Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic Analysis*, 4:271–283, 1991.

[Barles, 1994] Guy Barles. *Solutions de viscosité des équations de Hamilton-Jacobi*, volume 17 of *Mathématiques et Applications.* Springer-Verlag, 1994.

[Bertsekas, 1987] Dimitri P. Bertsekas. *Dynamic Programming : Deterministic and Stochastic Models.* Prentice Hall, 1987.

[Boyan and Moore, 1995] J.A. Boyan and A.W. Moore. Generalization in reinforcement learning : Safely approximating the value function. *Advances in Neural Information Processing Systems*, 7, 1995.

[Crandall *et al.*, 1992] M.G. Crandall, Hitoshi Ishii, and P.L. Lions. User's guide to viscosity solutions of second order partial differential equations. *Bulletin of the American Mathematical Society*, 27(1), 1992.

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

[Gordon, 1995] G. Gordon. Stable function approximation in dynamic programming. *International Conference on Machine Learning*, 1995.

[Moore, 1994] Andrew W. Moore. The parti-game algorithm for variable resolution reinforcement learning in multidimensional state space. *Advances in Neural Information Processing Systems*, 6, 1994.

[Munos, 1996] Rémi Munos. A convergent reinforcement learning algorithm in the continuous case : the finite-element reinforcement learning. *International Conference on Machine Learning*, 1996.

[Watkins, 1989] Christopher J.C.H. Watkins. *Learning from delayed reward.* PhD thesis, Cambridge University, 1989.