A convergent Reinforcement Learning algorithm in the continuous case : the Finite-Element Reinforcement Learning

Rémi Munos [*]
DASSAULT-AVIATION,DGT-DTN-EL-Et.Avancées
78 quai Marcel Dassault, 92214 Saint-Cloud, FRANCE
Tel : (33-1) 47 11 52 82 Fax : (33-1) 47 11 52 83
E-mail : Remi.Munos@cemagref.fr

Abstract

This paper presents a direct reinforcement learning algorithm, called *Finite-Element Reinforcement Learning*, in the continuous case, i.e. continuous state-space and time.

The evaluation of the value function enables the generation of an optimal policy for reinforcement control problems, such as target or obstacle problems, viability problems or optimization problems.

We propose a continuous formalism for the studying of reinforcement learning using the continuous optimal control framework, then we state the associated Hamilton-Jacobi-Bellman equation.

First, we propose to approximate the value function by a numerical scheme based on a finite-element method. This generates a discrete Markov Decision Process, with finite state and control spaces, which can be solved by Dynamic Programming. The computation of this approximation scheme, in reinforcement learning terminology, belongs to the class of indirect learning methods.

Then we present our direct learning algorithm which approximates the previous finite-element scheme and prove its convergence to the value function of the continuous problem.

1 INTRODUCTION

In this paper, we present a theoretical study of Reinforcement Learning (RL) in which the state space and the time are continuous and we propose a reinforcement learning algorithm, called *Finite-Element Reinforcement Learning* (FERL), that converges to the optimal solution. An adequate formalism using the continuous optimal control framework is given. The Hamilton-Jacobi-Bellman (HJB) equation is stated. Then some problems due to the resolution of the HJB equation are underlined and references to viscosity solutions are indicated.

A number of reinforcement learning algorithms have been implemented with neural networks or other approximation systems (see Barto, Sutton and Anderson (1983), Barto (1990), Gullapalli (1992), Lin (1993) and many other). However, as it has been pointed out by Baird (1995), in general, these learning methods do not converge. The *Residual-gradient advantage updating* proposed by Harmon, Baird and Klopf (1996) is a convergent algorithm in the sense of the convergence of gradient descent methods. The main difficulty of these methods is to find how to design a suitable architecture so that the network could approximate the value function. Besides, gradient descent methods only insure local optimum.

Our approach is different. First, we study an approximation scheme that discretizes at some step δ the HJB equation using a finite-element method (with linear simplexes). The important feature of this scheme is that the discretized HJB equation is itself the Dynamic Programming (DP) equation for a suitable Markov Decision Process. This property has been extensively used to construct convergent numerical schemes (see Souganidis (1985), Kushner and Dupuis (1992), Fleming and Soner (1993)).

Secondly, we propose a direct RL algorithm that constructs a piecewise linear continuous function that $\epsilon-$converges to the previous approximation scheme. This is the main contribution of this paper.

Section 2 presents a formulation of reinforcement learning in the continuous case using the formalism of optimal control theory. We are interested in deterministic systems with discounted reinforcement and

[*] CEMAGREF, LISC, Parc de Tourvoie, B.P.121, 92185 Antony Cedex, FRANCE. Tel : (33-1) 40 96 61 21 Poste 64 14 Fax : (33-1) 40 96 60 80

infinite horizon.

Section 3 is a very brief study of some properties of the value function. It introduces the HJB equation and gives some conditions for the continuity of the value function.

Section 4 presents an approximation scheme using as a finite-element the linear simplex. The convergence of this scheme is stated but the proof is not given here.

Section 5 presents the Finite-Element Reinforcement Learning algorithm and the convergence is proven in appendix A.

2 REINFORCEMENT LEARNING, A CONTINUOUS FORMALISM

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

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

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

Starting from initial state x at time $t = 0$ and choosing 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 $r : \bar{O} \times U \to \mathbb{R}$ is the *running reinforcement* and $R : \Omega \setminus O \to \mathbb{R}$ the *terminal reinforcement.* γ is the *discount factor* $(0 \leq \gamma < 1)$.

Then, the *value function* is defined by :

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

In the following, we consider problems satisfying hypotheses :

Hyp. 1 *We assume that :*
- $f : \bar{O} \times U \to \mathbb{R}^n$ is bounded with M_f and Lipschitzian : $|f(x,u) - f(y,u)| \leq L_f ||y - x||$,
- $r : \bar{O} \times U \to \mathbb{R}$ is bounded with M_r and Lipschitzian : $|r(x,u) - r(y,u)| \leq L_r ||y - x||$,
- $R : \Omega \setminus O \to \mathbb{R}$ is Lipschitzian : $|R(x) - R(y)| \leq L_R ||y - x||$.
- The boundary ∂O is $\mathcal{C}^2$.

3 THEORETICAL STUDY OF THE VALUE FUNCTION

3.1 THE HAMILTON-JACOBI-BELLMAN EQUATION

The following theorem comes from Bellman principle in the continuous case (See Fleming and Soner (1993) for a complete survey).

Theorem 1 (Hamilton-Jacobi-Bellman) *If the value function V is differentiable at x, let $DV(x)$ be the gradient of V at x, then the Hamilton-Jacobi-Bellman equation :*

$$\boxed{V(x)\ln\gamma + \sup_{v \in U}[DV(x).f(x,v) + r(x,v)] = 0} \tag{1}$$

holds at $x \in O$. Besides, the value function satisfies the following boundary conditions :

$$V(x) \geq R(x) \text{ for } x \in \partial O \tag{2}$$

Reinforcement learning methods intend to approximate the value function in order to build an optimal policy, i.e. a feed-back control $\pi(x) : \bar{O} \to U$ that optimizes the reinforcement functional J. Indeed, knowing V enables the choose of an optimal control :

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

We consider the following hypothesis concerning the dynamics around the boundary ∂O and we state a theorem of continuity for V (whose proof is in Barles and Perthame (1990)).

Hyp. 2 *For all $x \in \partial O$, let $\vec{n}(x)$ be the outward normal of O at x,*
- If there exists $u \in U$ with $f(x,u).\vec{n}(x) \leq 0$ then there exists $v \in U$ with $f(x,v).\vec{n}(x) < 0$,
- If for all $u \in U$, $f(x,u).\vec{n}(x) \leq 0$ then for all $u \in U$, $f(x,u).\vec{n}(x) < 0$

Theorem 2 (Continuity) *Suppose that Hyp. 1 and Hyp. 2 are satisfied, then the value function is continuous in O.*

3.2 SOLUTIONS TO HJB EQUATION

If the value function is differentiable then it solves the HJB equation. However, in general, the value function is not smooth enough to satisfy the HJB equation everywhere, so there is no classical solutions (differentiable everywhere) to HJB equation. Besides, there are many functions other than the value function that satisfy (1) and (2) almost everywhere (i.e. many generalized solutions). An other problem is that the boundary conditions (2) may hold with a strict inequality.

A well suited method to overcome these problems is to define a weak formulation of solutions to HJB equation, which are called *viscosity solutions.* This notion has been introduced by Crandall and Lions (1983) (for a complete survey, see Crandall, Ishii and Lions (1992), Barles (1994) or Fleming and Soner (1993)). The definition and the usefulness of viscosity solutions are beyond the scope of this paper but a result is that if the value function is continuous, then it is the unique viscosity solution to HJB equation (1) with the boundary conditions (2).

4 STUDY OF A CONVERGENT APPROXIMATION SCHEME

The continuous optimal control problem is approximated by a controlled Markov process on the set of vertices of a triangulation upon the state space. The discretized HJB equation is a DP equation for a suitably defined stochastic control problem for Markov chains. We use a finite-element method (with linear-simplexes) derived from Kushner (1990) for approximating the value function.

For a maximal length δ of the sides of the simplexes, consider a triangulation of the state space (figure 1)

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

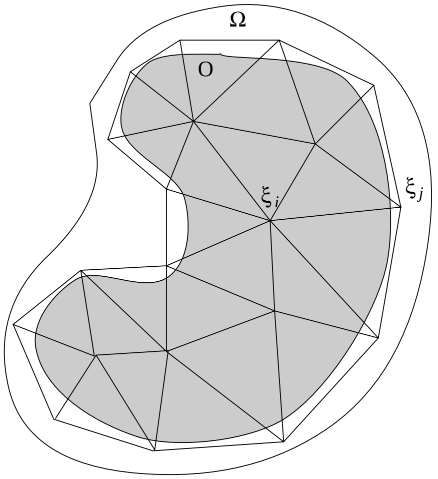

Figure 1: Triangulation $\Sigma^\delta = \{\xi_i\} \subset \Omega$ of the state space. We assume that : $\exists k_\Sigma$, $\text{dist}(\Sigma^\delta, O) \leq k_\Sigma \delta^2$ and that $\exists k_\rho$ s.t. the radius of the sphere inscribed in each simplex is $> k_\rho \delta$

We consider piecewise linear continuous functions defined upon Σ^δ : $\phi(x) = \sum_{i=0}^n \lambda_{\xi_i}(x)\phi(\xi_i)$ with $\lambda_{\xi_i}(x)$ the barycentric coordinates of x inside the simplex $(\xi_0, ..., \xi_n) \ni x$.

Thanks to a contraction property due to the discount factor γ (see Bertsekas (1987)), DP theory insures that there is a unique solution, called V^δ that satisfies equations :

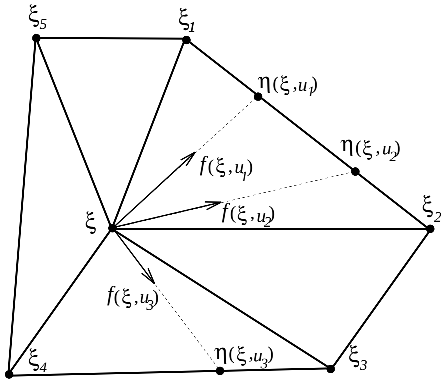

Figure 2: Approximation scheme in a simplex : the values $\eta(\xi, u_i)$ are the projection of vertex ξ in a parallel direction to $f(\xi, u_i)$

$$\begin{aligned} V^\delta(\xi) &= \sup_{u \in U^\delta} [\gamma^{\tau(\xi,u)}.V^\delta(\eta(\xi,u)) + \tau(\xi,u)r(\xi,u)] \\ &= \sup_{u \in U^\delta} [\gamma^{\tau(\xi,u)}.\sum_{i=1}^n \lambda_{\xi_i}(\eta(\xi,u))V^\delta(\xi_i) \\ &\quad +\tau(\xi,u)r(\xi,u)] \text{ for } \xi \in O \qquad (3) \\ V^\delta(\xi) &= R(\xi) \text{ for } \xi \in \Omega \setminus O \end{aligned}$$

where $\eta(\xi, u)$ is the projection of ξ in a parallel direction to $f(\xi, u)$ onto the opposite side of the simplex (see figure 2) and $\tau(\xi, u)$ is such that : $\eta(\xi, u) = \xi + \tau(\xi, u)f(\xi, u)$.

We need the following hypotheses for the convergence of the scheme V^δ :

Hyp. 3 *We assume that :*
- $\exists \nu_1 < 0$ *such that* $\forall x \in \partial O, \exists u \in U, f(x,u).\vec{n}(x) \leq \nu_1$
- $\exists \nu_2 > 0$ *such that* $\forall x \in \partial O, \exists u \in U, f(x,u).\vec{n}(x) \geq \nu_2$

Theorem 3 (Convergence of the scheme)
Suppose that Hyp. 1 and Hyp. 3 are satisfied, then the approximation scheme V^δ converges to the value function of the continuous problem as δ tends to 0 :

$$\lim_{\xi \xrightarrow{\delta \downarrow 0} x} V^\delta(\xi) = V(x) \text{ for all } x \in O$$

The demonstration of this theorem uses the general convergence results of approximation schemes for non-linear equations of Barles and Souganidis (1991) and a comparison principle of Barles and Perthame (1990) and Barles (1994) but we will not give the proof here. (similar approximation schemes have been proven in

Kushner (1990), Kushner and Dupuis (1992), Souganidis (1985), Fleming and Soner (1993) but without the boundary conditions used here).

Remark : Hyp. 3 is rather strong and can be weakened in some cases (see Barles and Perthame (1990)).

In the RL terminology, the computation of the value function with such an approximation scheme is called an *indirect learning* method, because first the system has to learn the dynamics f (thus $\eta(\xi_i, u)$) and secondly it computes the value function with DP algorithms (value iteration or policy iteration) (see Barto, Bradtke and Singh (1991) and Bertsekas (1987)).
On the contrary, a *direct learning* approach is a real time learning method that approximates the value function without learning any model of the dynamics. An example is the Q-learning algorithm (see Watkins (1989) and Watkins and Dayan (1992)). In the next section, we present a direct learning algorithm which is a kind of Real Time Dynamic Programming (RTDP) (see Barto et al. (1991)).

Remark : The equation (3) may be rewritten by defining upon Σ^δ the Q-values (this notation will be useful in the next section) :

$$Q^\delta(\xi, u) = \gamma^{\tau(\xi,u)}.V^\delta(\eta(\xi, u)) + \tau(\xi, u)r(\xi, u)$$
$$\text{and } V^\delta(\xi) = \sup_{u \in U^\delta} Q^\delta(\xi, u) \qquad (4)$$

5 THE FINITE-ELEMENT REINFORCEMENT LEARNING

5.1 PRESENTATION

Here we construct a direct reinforcement learning algorithm that ϵ-converges to the solution of the approximation scheme previously studied.

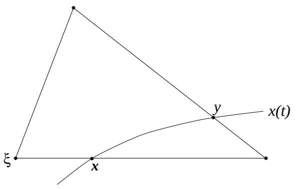

Figure 3: The trajectory exits at $y \in O$

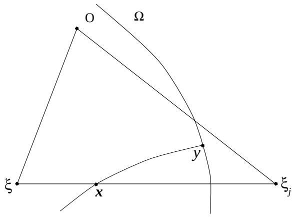

Figure 4: The trajectory hits the boundary at $y \in \partial O$

Let us consider a trajectory $x(t)$ which enters simplex T at point $x = x(t_1)$, then some control u is chosen and kept until the trajectory exits T at $y = x(t_2)$ (figure 3) or hits the boundary ∂O inside T (figure 4). This means that the choice of the control only occurs when the system crosses the border of simplexes.

Let $\tau_x = t_2 - t_1$. Let $T_{in} \ni x$ be the (n-1)-input-simplex and T_{out} the (n-1)-output-simplex. ($y \in T_{out}$ in the case of figure 3).

The algorithm is the iterated version of equation (4) and use the available knowledge for approximating $V^\delta(\eta(\xi, u))$ and $\tau(\xi, u)$: when the system exits from T, (figure 3), the algorithm updates the Q-value of the vertex ξ opposite to the exit side T_{out} according to rule (6) below.

The idea is that for small δ, thanks to Thales' theorem,
$\frac{\tau_x}{\lambda_\xi(x)}$ approximates $\tau(\xi, u)$, and
$\frac{V_n^\delta(y) - V_n^\delta(x)}{\lambda_\xi(x)} + V_n^\delta(\xi)$ approximates $V^\delta(\eta(\xi, u))$,

where as in the previous section, Q_n^δ and V_n^δ are the iterated values such that :
$V_n^\delta(\xi) = \sup_{u \in U^\delta} Q_n^\delta(\xi, u)$ and $V_n^\delta(x) = \sum \lambda_\xi(x) V_n^\delta(\xi)$ with $\{\lambda_\xi(x)\}$ the barycentric coordinates of $x \in T$.

In order to prevent the trajectories from staying infinitely inside a simplex, we assume that :

Hyp. 4 $\exists m_f > 0, \forall x \in \bar{O}, ||f(x, u)|| \geq m_f$

5.2 THE ALGORITHM

Let us choose a constant $\lambda \in (0, 1]$ close to zero. For any initial values $Q_0^\delta(\xi, u)$, we consider the controlled dynamical system described previously. For any trajectory going through a simplex T with control u, let consider the two cases :
- the trajectory exits from T, so $y \in O$ (figure 3).
- the trajectory hits ∂O inside T, so $y \in \partial O$ (figure 4).

- If the second case happens, the nearest vertex $\xi_j \in \Omega \setminus O$ from y is updated with :

$$V_{n+1}^\delta(\xi_j) = R(y) \qquad (5)$$

- In both cases, if Hyp. 5 and Hyp. 6 :

  Hyp. 5 $\lambda_\xi(x) \geq \lambda$
  (this relation eliminates cases for which $\lambda_\xi(x)$ is too small)

  Hyp. 6 $\forall \xi_i \in T_{in} \cap T_{out}$, $\lambda_{\xi_i}(y) > \lambda_{\xi_i}(x) + \lambda$
  (these relations imply that $y - x$ strictly belongs to the cone of vertex ξ and base T_{out} and insure that for small enough δ, $\eta(\xi, u) \in T_{out}$).

  are satisfied, then update the Q-value of vertex ξ opposite to the exit side T_{out} with :

$$Q_{n+1}^\delta(\xi, u) = \gamma^{\frac{\tau_x}{\lambda_\xi(x)}} \left(\frac{V_n^\delta(y) - V_n^\delta(x)}{\lambda_\xi(x)} + V_n^\delta(\xi) \right) + \frac{\tau_x}{\lambda_\xi(x)} r(x, u) \qquad (6)$$

The following hypothesis (similar to the one of RTDP) allows the statement of the convergence theorem whose proof is given in appendix A.

Hyp. 7 *When experimenting the FERL, we consider series of trajectories such that the algorithm leads to update (with rule(5)) every state $\xi \in \Omega \setminus O$ at least once and update (with rule(6)) every state $\xi \in O$ and control $u \in U^\delta$ infinitely often.*

Theorem 4 (Convergence of the algorithm) *Suppose that Hyp. 1, Hyp. 3 and Hyp. 4 are satisfied, for all $\epsilon > 0$, there exists Δ such that for all $\delta \leq \Delta$, for any triangulation Σ^δ satisfying hypotheses of section 4, if we use the FERL algorithm such that Hyp. 7 is satisfied, then there exists N, for all $n \geq N$,*

$$\sup_{x \in O} |V_n^\delta(x) - V(x)| \leq \epsilon.$$

Remark 1 : Rule (6) is similar to a RTDP iteration ; it could be replaced by a Q-learning like rule :

$$Q_{n+1}^\delta(\xi, u) = [1 - \alpha_n(\xi, x)] Q_n^\delta(\xi, u) + \alpha_n(\xi, x) \Bigg[\gamma^{\frac{\tau_x}{\lambda_\xi(x)}} \left(\frac{V_n^\delta(y) - V_n^\delta(x)}{\lambda_\xi(x)} + V_n^\delta(\xi) \right) + \frac{\tau_x}{\lambda_\xi(x)} r(x, u) \Bigg] \quad (7)$$

for appropriated $\alpha_n(\xi, x) \leq 1$. This rule enables the averaging of the Q-values, and the convergence of the algorithm is still insured.

Remark 2 : We can note that for a fixed δ, the V_n^δ-values do not converge as n tends to infinity (this is natural because the discretized process is not Markovian). Meanwhile, these values get closer to the value function V with some error ϵ. The convergence of the algorithm is insured as δ tends to zero.

Remark 3 : For computational aspects of this algorithm, we can start with an a-priori rough triangulation of the state space (high δ) and progressively decrease δ and modify the triangulation (for example splitting each simplex into $n + 2$ smaller ones) until the desired estimation error is reached.

6 CONCLUSION

This paper proposes an adequate formalism of Reinforcement Learning in the continuous case and states the Hamilton-Jacobi-Bellman equation. This is an important result in the field of reinforcement adaptive control.

After a brief study of some properties of the value function, a finite-element approximation scheme using linear simplexes is proposed. Then we define a direct RL algorithm that builds a piecewise linear continuous function using a triangulation of the state space. The function computed by the Finite-Element Reinforcement Learning is proven to converge to the value function of the continuous problem.

For computational aspects 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 multi-resolution simplexes (using the work of Akian (1990) for multigrid methods or Moore (1994) for variable resolution), or use other finite-element than the linear simplex. Following remark 3, we are currently working on a non-homogeneous splitting of the simplexes depending on a local estimation of the regularity of the value function.

An other improvement should be to consider the stochastic case where the dynamics is described by a stochastic differential equation.

A APPENDIX : CONVERGENCE OF THE ALGORITHM

The main idea of the demonstration is to prove that for all $\epsilon > 0$, for small enough values of δ, the approximation error between the computed V_n^δ and the V^δ values of the convergent approximation scheme is $< \epsilon_2$ (with ϵ_2 linear function of ϵ) for $n \geq N$, i.e. :

$$\sup_{\xi \in \Sigma^\delta} |V_n^\delta(\xi) - V^\delta(\xi)| \leq \epsilon_2 \quad (8)$$

The theorem is then easily proven : let ϵ_1 be such that $\epsilon_1 + \epsilon_2 = \epsilon$; thanks to theorem 3, for small δ,

$$\sup_{\xi \in \Sigma^\delta} |V^\delta(\xi) - V(\xi)| \leq \epsilon_1$$

Thus we have :

$$\sup_{\xi \in \Sigma^\delta} |V_n^\delta(\xi) - V(\xi)| \leq \epsilon_1 + \epsilon_2 \leq \epsilon$$

In order to prove (8), we need to evaluate the modification of the error $|Q_n^\delta(\xi, u) - Q^\delta(\xi, u)|$ after updating $Q_n^\delta(\xi, u)$ with rule (6). This is the object of section A.3. Section A.1 introduces some useful notations and Section A.2 gives some properties, a comparison of times τ_ξ and τ_x and a important majoration result.

A.1 SOME DEFINITIONS

Let T be the current n-simplex. The trajectory $x(.)$ enters T at $x = x(t_1) \in T_{in}$ the (n-1)-input-simplex (for example triangle (ξ, ξ_1, ξ_2) in figure 5) and exits

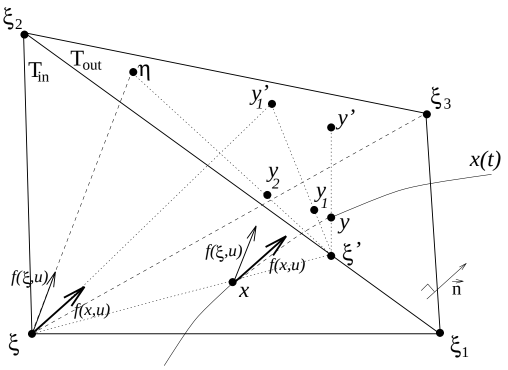

Figure 5: A 3-simplex $(\xi, \xi_1, \xi_2, \xi_3)$

from T at $y = x(t_2) \in T_{out}$ the (n-1)-output-simplex (for example triangle (ξ_1, ξ_2, ξ_3)).

Let P be the hyperplan of dimension $(n-1)$ parallel to T_{out} through y (in the case $y \in O$, see figure (6), P is the hyperplan generated by T_{out}). In the following, all the projections are made onto P.

Let η (resp. y_2) be the projection of ξ (resp. x) in a parallel direction to $f(\xi, u)$. Let y_1 be the projection of x in a parallel direction to $f(x, u)$, Let ξ' be the projection of ξ in a parallel direction to $x - \xi$. Let $y' = \xi' + \frac{1}{\lambda_\xi(x)}(y - \xi')$ and $y'_1 = \xi' + \frac{1}{\lambda_\xi(x)}(y_1 - \xi')$ (hence, from Thales' theorem, y'_1 is the projection of ξ in a parallel direction to $f(x, u)$).

Let $\tau_x = t_2 - t_1$. Let τ_ξ be such that : $\eta = \xi + \tau_\xi f(\xi, u)$, Let τ_1 be such that : $y_1 = x + \tau_1 f(x, u)$, Let τ_2 be such that : $y_2 = x + \tau_2 f(\xi, u)$, Let τ'_1 be such that : $y'_1 = \xi + \tau'_1 f(x, u)$. Let $y_3 = x + \tau_x f(x, u)$ (figure (6)).

Let M_{V^δ} be a majoration of V^δ.

A.2 SOME PROPERTIES

A.2.1 y', y'_1 and η belong to T_{out}

In this section, we study the case depicted in figure (6) and we prove that if Hyp. 6 and Hyp. 5 are satisfied, for small enough δ, we have y', y'_1 and $\eta \in T_{out}$.

From Hyp. 6 and the definition of y', we deduce that : $\forall \xi_i \in T_{in} \cap T_{out}$, $\lambda_{\xi_i}(y') > \lambda$.

Some geometrical considerations give $||y - y_1|| \leq \frac{||f(x,u)||}{f(x,u)\vec{n}} ||y - y_3||$. Besides, Taylor's majoration gives :

$$||y - y_3|| \leq \frac{1}{2} L_f \tau_x^2 \tag{9}$$

Since $\tau_x \leq \frac{\delta}{m_f}$ we have : $||y - y_1|| \leq \frac{L_f \delta^2}{2m_f^2} \frac{||f(x,u)||}{f(x,u)\vec{n}}$. Thus, for small enough values of δ, $y_1 \in T_{out}$.
So there exists a vertex $\xi_i \neq \xi$ of T such that : $\frac{f(x,u)\vec{n}}{||f(x,u)||} \geq \frac{(\xi_i - \xi)\vec{n}}{||\xi_i - \xi||}$. Let $\bar{\xi}$ be the orthogonal projection of ξ onto P, then : $\frac{f(x,u)\vec{n}}{||f(x,u)||} \geq \frac{||\bar{\xi} - \xi||}{||\xi_i - \xi||} \geq \frac{2k_\rho \delta}{\delta} \geq 2k_\rho$. Thus, we have :

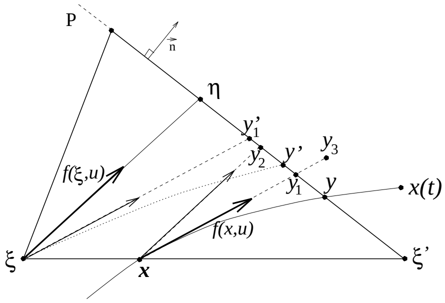

Figure 6: Useful notations in a 2-simplex. The trajectory enters at x and exits at $y \in O$

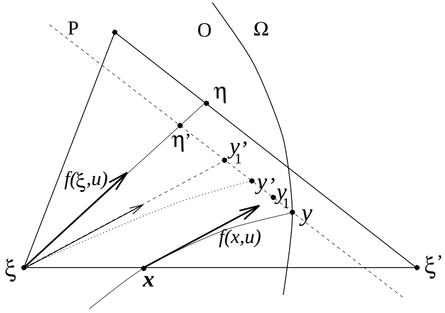

Figure 7: The trajectory hits the boundary : $y \in \partial O$

$$||y - y_1|| \leq \frac{1}{2k_\rho} ||y - y_3|| \leq \frac{L_f \delta^2}{4m_f^2 k_\rho} \tag{10}$$

From Thales' theorem : $||y' - y'_1|| = \frac{1}{\lambda_\xi(x)} ||y - y_1||$. So with Hyp. 5 we have :

$$||y' - y'_1|| \leq \frac{L_f \delta^2}{4\lambda m_f^2 k_\rho}$$

Thus for small δ, y'_1 is as close to y' as wished, so $\forall \xi_i \in T_{in} \cap T_{out}, \lambda_{\xi_i}(y'_1) \geq \frac{\lambda}{2}$.

Besides, $||\eta - y'_1|| = \frac{1}{\lambda_\xi(x)} ||y_2 - y_1||$, and using majoration (13) below, we deduce that :

$$||\eta - y'_1|| \leq \frac{(1-\lambda)}{\lambda} \frac{L_f}{m_f} \left(1 + \frac{1}{2k_\rho}\right) \delta^2$$

Thus, η is as close to y'_1 as wished, so there exists Δ_0 such that for any $\delta \leq \Delta_0$, we have y', y'_1 and $\eta \in T_{out}$.

A.2.2 Comparison of τ_ξ and τ_x

From Thales' theorem : $\frac{||\xi' - x||}{||\xi' - \xi||} = \frac{||y_2 - x||}{||\eta - \xi||}$, so $\lambda_\xi(x) = \frac{\tau_2}{\tau_\xi}$. Besides, some geometrical considerations give :

$\frac{\tau_2}{\tau_1} = \frac{f(x,u)\vec{n}}{f(\xi,u)\vec{n}}$. From the majorations (9) and (10) we have :

$$|\tau_1 - \tau_x|.||f(x,u)|| = ||y_1 - y_3|| \leq ||y_1 - y|| + ||y - y_3|| \leq \frac{1}{2} L_f \tau_x^2 \left(1 + \frac{1}{2k_\rho}\right)$$

Thus, from what precedes,
$\left|\lambda_\xi(x) \frac{f(x,u)\vec{n}}{f(\xi,u)\vec{n}} \tau_\xi - \tau_x\right| ||f(x,u)|| \leq \frac{1}{2} L_f \tau_x^2 \left(1 + \frac{1}{2k_\rho}\right)$

$$|\lambda_\xi(x)\tau_\xi - \tau_x|.||f(x,u)|| \leq \frac{1}{2} L_f \tau_x^2 \left(1 + \frac{1}{2k_\rho}\right) + \lambda_\xi(x)\tau_\xi ||f(\xi,u) - f(x,u)|| \frac{||f(x,u)||}{f(x,u)\vec{n}}$$

From the Lipschitz property of f, we have :

$$\left|\tau_\xi - \frac{\tau_x}{\lambda_\xi(x)}\right| \leq k_\tau \delta^2 \text{ with}$$

$$k_\tau = \frac{L_f}{m_f^2} \left[\frac{1}{2k_\rho} + \frac{1}{2m_f}\left(1 + \frac{1}{2k_\rho}\right)\right] \quad (11)$$

We deduce from this majoration and from the property of exponential function : $\gamma^x \geq 1 - x \ln \frac{1}{\gamma}$, that :

$$\left|\gamma^{\tau_\xi} - \gamma^{\frac{\tau_x}{\lambda_\xi(x)}}\right| \leq k_\tau \ln \frac{1}{\gamma}.\delta^2$$

A.2.3 Majoration of $V^\delta(\eta) - V^\delta(y')$

Some geometrical considerations show that the gradient of the linear function V^δ defined upon the (n-1)-simplex T_{out} is majorated by $\frac{n-1}{2k_\rho} \frac{|V^\delta(\xi_i) - V^\delta(\xi_j)|}{||\xi_i - \xi_j||}$, for some vertices ξ_i and ξ_j of T_{out}. Thus, we have :

$$\left|V^\delta(y') - V^\delta(\eta)\right| \leq ||y' - \eta|| \frac{(n-1)}{2k_\rho} \frac{\left|V^\delta(\xi_i) - V^\delta(\xi_j)\right|}{||\xi_i - \xi_j||}$$

Moreover, $\left|V^\delta(\xi_i) - V^\delta(\xi_j)\right| \leq \left|V^\delta(\xi_i) - V(\xi_i)\right| + \left|V(\xi_i) - V(\xi_j)\right| + \left|V(\xi_j) - V^\delta(\xi_j)\right|$.
We know that V^δ converges to V, so let consider any $\epsilon_1 > 0$, then there exists Δ_1, for any $\delta \leq \Delta_1$, we have : $|V^\delta - V| \leq \epsilon_1$.

Besides, thanks to the continuity of V, there exists Δ_2, for all $\delta \leq \Delta_2$, we have : $|V(\xi_i) - V(\xi_j)| \leq \epsilon_1$. Thus, for any $\delta \leq \min\{\Delta_1, \Delta_2\}$, we have : $\left|V^\delta(\xi_i) - V^\delta(\xi_j)\right| \leq 3\epsilon_1$. Knowing that $||\xi_i - \xi_j|| \geq 2k_\rho\delta$, we have :

$$\left|V^\delta(y') - V^\delta(\eta)\right| \leq ||y' - \eta|| \frac{3(n-1)\epsilon_1}{4k_\rho^2\delta} \quad (12)$$

Now, let's estimate $||y' - \eta||$. We have the relation :

$$\begin{aligned} ||\eta - y_1'|| &= ||\tau_\xi f(\xi,u) - \tau_1' f(x,u)|| \\ &= \tau_1' \left|\left|\frac{f(x,u).\vec{n}}{f(\xi,u).\vec{n}} f(\xi,u) - f(x,u)\right|\right| \end{aligned}$$

Using the lipschitz property of f, we have :

$$\begin{aligned} ||\eta - y_1'|| &\leq \tau_1' ||\xi - x|| L_f \left(1 + \frac{1}{2k_\rho}\right) \\ &\leq \frac{L_f}{m_f}\left(1 + \frac{1}{2k_\rho}\right)[1 - \lambda_\xi(x)]\delta^2 \quad (13) \end{aligned}$$

Remark : these majorations have been established in the case $y \in O$. But in the case $y \in \partial O$ (see figure (7)) the last one becomes :

$$\begin{aligned} ||\eta - y_1'|| &\leq ||\eta - \eta'|| + ||\eta' - y_1'|| \quad (14) \\ &\leq \frac{k_\Sigma \delta^2}{2k_\rho} + \frac{L_f}{m_f}\left(1 + \frac{1}{2k_\rho}\right)[1 - \lambda_\xi(x)]\delta^2 \end{aligned}$$

Besides, we have : $||y_1' - y'|| = \frac{||y_1 - y||}{\lambda_\xi(x)} \leq \frac{L_f \tau_x^2}{4\lambda_\xi(x) k_\rho}$, and : $\frac{\tau_x}{\lambda_\xi(x)} \leq \tau_\xi + k_\tau\delta^2$, so :

$$||y_1' - y'|| \leq \frac{L_f}{4k_\rho} \tau_x (\tau_\xi + k_\tau \delta^2) \leq \frac{L_f(1 + k_\tau m_f \delta)}{4m_f^2 k_\rho} \delta^2 \quad (15)$$

Putting together (12), (14) and (15), we get :

$$|V^\delta(\eta) - V^\delta(y')| \leq \left[\frac{L_f}{m_f}\left(1 + \frac{1}{2k_\rho}\right)[1 - \lambda_\xi(x)] + \frac{L_f(1 + k_\tau m_f \delta)}{4m_f^2 k_\rho} + \frac{k_\Sigma \delta}{2k_\rho}\right] \frac{3(n-1)\epsilon_1 \delta}{4k_\rho^2}$$

A.3 CONVERGENCE OF FERL

A.3.1 Majoration of $Q_{n+1}^\delta(\xi,u) - Q^\delta(\xi,u)$

Let $E_n^\delta = \sup_{\xi \in \Sigma^\delta \cap O, u \in U^\delta} |Q_n^\delta(\xi,u) - Q^\delta(\xi,u)|$.
After updating $Q_n^\delta(\xi,u)$ with rule (6), let Λ denote the value $Q_{n+1}^\delta(\xi,u) - Q^\delta(\xi,u)$. We have :

$$\begin{aligned} \Lambda &= \gamma^{\frac{\tau_x}{\lambda_\xi(x)}} \left[\frac{V_n^\delta(y) - V_n^\delta(x)}{\lambda_\xi(x)} + V_n^\delta(\xi)\right] - \gamma^{\tau_\xi} V^\delta(\eta) \\ &\quad + \frac{\tau_x}{\lambda_\xi(x)} r(x,u) - \tau_\xi r(\xi,u) \\ &= \gamma^{\frac{\tau_x}{\lambda_\xi(x)}} V_n^\delta(y') - \gamma^{\tau_\xi} V^\delta(\eta) + \frac{\tau_x}{\lambda_\xi(x)} r(x,u) - \tau_\xi r(\xi,u) \\ &= \gamma^{\frac{\tau_x}{\lambda_\xi(x)}} \left[V_n^\delta(y') - V^\delta(y')\right] + \gamma^{\tau_\xi}\left[V^\delta(y') - V^\delta(\eta)\right] \\ &\quad + \tau_\xi[r(x,u) - r(\xi,u)] + V^\delta(y')\left(\gamma^{\frac{\tau_x}{\lambda_\xi(x)}} - \gamma^{\tau_\xi}\right) \\ &\quad + \left(\frac{\tau_x}{\lambda_\xi(x)} - \tau_\xi\right) r(x,u) \\ |\Lambda| &\leq \gamma^{\frac{\tau_x}{\lambda_\xi(x)}} E_n^\delta + \left|V^\delta(y') - V^\delta(\eta)\right| \\ &\quad + \left(\frac{L_r}{m_f} + M_{V^\delta} k_\tau \ln \frac{1}{\gamma} + M_r k_\tau\right)\delta^2 \\ &\leq \gamma^{\frac{\tau_x}{\lambda_\xi(x)}} E_n^\delta + k_0\epsilon_1[1 - \lambda_\xi(x)]\delta + k_1\epsilon_1\delta + k_2\delta^2 \end{aligned}$$

with :

$$\begin{aligned} k_0 &= \frac{3(n-1)L_f}{4m_f k_\rho^2}\left(1+\frac{1}{2k_\rho}\right) \\ k_1 &= \frac{3(n-1)L_f}{16m_f^2 k_\rho^3} \\ k_2 &= \frac{L_r}{m_f} + M_{V^\delta} k_\tau \ln\frac{1}{\gamma} + M_r k_\tau + \frac{3(n-1)L_f k_\tau \epsilon_1}{16 m_f k_\rho^3} \end{aligned}$$

A.3.2 Hypotheses for $E_n^\delta \leq \epsilon_2$

Let us suppose that the following conditions are true for some $\alpha > 0$:

$$E_n^\delta > \epsilon_2 \quad \Rightarrow \quad |\Lambda| \leq E_n^\delta - \alpha \tag{16}$$
$$E_n^\delta \leq \epsilon_2 \quad \Rightarrow \quad |\Lambda| \leq \epsilon_2 \tag{17}$$

From Hyp. 7, all states $\xi \in \Sigma^\delta \cap O$ and controls are updated an infinite number of times, so there exists an integer m such that at stage $n+m$ all the states in O and controls have been iterated at least once since stage n. Then, from (16) and (17) we have :

$$\begin{aligned} E_n^\delta > \epsilon_2 &\quad \Rightarrow \quad E_{n+m}^\delta \leq E_n^\delta - \alpha \\ E_n^\delta \leq \epsilon_2 &\quad \Rightarrow \quad E_{n+m}^\delta \leq \epsilon_2 \end{aligned}$$

This implies that there exists N_1 such that :
$\forall n \geq N_1, E_n^\delta \leq \epsilon_2$.

Besides, from Hyp. 7, all states $\xi_j \in \Sigma^\delta \cap (\Omega \setminus O)$ are updated at least once with rule (5). Then :

$$\begin{aligned} |V_{n+1}^\delta(\xi_j) - V^\delta(\xi_j)| = |R(y) - R(\xi_j)| \leq L_R||y - \xi_j|| \\ \leq L_R \delta \leq \epsilon_2 \end{aligned}$$

for any $\delta \leq \Delta_3 = \frac{\epsilon_2}{L_R}$. Thus, $\exists N_2$, such that :
$\forall n \geq N_2, \sup_{\xi_j \in \Sigma^\delta \cap (\Omega \setminus O)} |V_n^\delta(\xi_j) - V^\delta(\xi_j)| \leq \epsilon_2$.

So : $\forall n \geq N = \max\{N_1, N_2\}$,

$$\sup_{\xi \in \Sigma^\delta} |V_n^\delta(\xi) - V^\delta(\xi)| \leq \epsilon_2$$

A.3.3 Sufficient condition

A sufficient condition for conditions (16) and (17) to be true is that the following inequality is true.

$$\gamma^{\frac{\tau_x}{\lambda_\xi(x)}} \frac{\epsilon_2}{2} + k_0 \epsilon_1 [1 - \lambda_\xi(x)]\delta + k_1 \epsilon_1 \delta + k_2 \delta^2 \leq \frac{\epsilon_2}{2} \tag{18}$$

Proof : if $E_n^\delta > \epsilon_2$ then

$$\begin{aligned} |\Lambda| &\leq E_n^\delta - (1 - \gamma^{\frac{\tau_x}{\lambda_\xi(x)}})\epsilon_2 + k_0\epsilon_1[1 - \lambda_\xi(x)] \\ &\quad + k_1 \epsilon_1 \delta + k_2 \delta^2 \\ &\leq E_n^\delta - (1 - \gamma^{\frac{\tau_x}{\lambda_\xi(x)}})\frac{\epsilon_2}{2} \end{aligned}$$

From a property of the exponential function, we have : $1 - \gamma^{\frac{\tau_x}{\lambda_\xi(x)}} \geq \frac{\tau_x}{2\lambda_\xi(x)} \ln \frac{1}{\gamma}$ *for any small enough value of* $\frac{\tau_x}{\lambda_\xi(x)}$*. Besides, from (11),* $\frac{\tau_x}{\lambda_\xi(x)} \geq \tau_\xi - k_\tau \delta^2$ *and* $\tau_\xi > \frac{2k_\rho\delta}{M_f}$*. So :* $1 - \gamma^{\frac{\tau_x}{\lambda_\xi(x)}} \geq \frac{1}{2} \ln\frac{1}{\gamma}\left(\frac{2k_\rho\delta}{M_f} - k_\tau\delta^2\right)$*. With* $\alpha = \frac{1}{2}\ln\frac{1}{\gamma}\left(\frac{2k_\rho\delta}{M_f} - k_\tau\delta^2\right)\frac{\epsilon_2}{2}$ *, there exists* Δ_4 *such that the first equation (16) is satisfied for any* $\delta \leq \Delta_4$*.*

The second condition (17) is true because :

$$\begin{aligned} |\Lambda| \leq \gamma^{\frac{\tau_x}{\lambda_\xi(x)}} \frac{\epsilon_2}{2} + \gamma^{\frac{\tau_x}{\lambda_\xi(x)}} \frac{\epsilon_2}{2} + k_0 \epsilon_1 [1 - \lambda_\xi(x)]\delta \\ + k_1 \epsilon_1 \delta + k_2 \delta^2 \leq \frac{\epsilon_2}{2} + \frac{\epsilon_2}{2} \end{aligned}$$

Let us find a sufficient condition on $\lambda_\xi(x)$ for condition (18) to be true :

$$1 - \lambda_\xi(x) \leq \frac{k_\rho \ln\frac{1}{\gamma}}{2k_0 M_f}\frac{\epsilon_2}{\epsilon_1} - \frac{k_1}{k_0} - \left(\frac{k_2}{k_0} + \frac{k_\tau \epsilon_2}{4k_0}\ln\frac{1}{\gamma}\right)\frac{\delta}{\epsilon_1}$$

For any $\delta \leq \Delta_5 = \epsilon_1^2$, it is sufficient that :

$$1 - \lambda_\xi(x) \leq \frac{k_\rho \ln\frac{1}{\gamma}}{2k_0 M_f}\frac{\epsilon_2}{\epsilon_1} - \frac{k_1}{k_0} - \left(\frac{k_2}{k_0} + \frac{k_\tau \epsilon_2}{4k_0}\ln\frac{1}{\gamma}\right)\sqrt{\delta} \tag{19}$$

A.3.4 Proof of the theorem

For any $\lambda \in (0, 1]$ and $\epsilon > 0$, let us choose $\epsilon_1 > 0$ and $\epsilon_2 > 0$ such that $\epsilon_1 + \epsilon_2 = \epsilon$, and

$$\frac{k_\rho \ln\frac{1}{\gamma}}{2k_0 M_f}\frac{\epsilon_2}{\epsilon_1} - \frac{k_1}{k_0} = 2(1 - \lambda)$$

That is :

$$\begin{aligned} \epsilon_1 &= \frac{k_\rho \ln\frac{1}{\gamma}}{2k_0 M_f\left(2(1-\lambda) + \frac{k_1}{k_0}\right) + k_\rho \ln\frac{1}{\gamma}}\epsilon \\ \epsilon_2 &= \epsilon - \epsilon_1 \end{aligned}$$

The sufficient condition (19) becomes :

$$1 - \lambda_\xi(x) \leq 2(1-\lambda) - \left(\frac{k_2}{k_0} + \frac{k_\tau \epsilon_2}{4k_0}\ln\frac{1}{\gamma}\right)\sqrt{\delta}$$

As $\sqrt{\delta}$ tends to 0, there exists Δ_6 such that for any $\delta \leq \Delta_6$, a sufficient condition is :

$$1 - \lambda_\xi(x) \leq 1 - \lambda$$

So, using the algorithm with points x satisfying Hyp. 5 with $\delta \leq \Delta = \min\{\Delta_0, ..., \Delta_6\}$ implies that condition (18) is true, so conditions (16) and (17) are true. Thus there exists an integer N, such that $\forall n \geq N, \sup |V_n^\delta - V^\delta| \leq \epsilon_2$. The theorem (4) is then easily proven. Indeed, for all $\lambda \in (0,1]$, for all $\epsilon > 0$, we define ϵ_1 and ϵ_2 as above. From the convergence of the approximation scheme V^δ, for $\delta \leq \Delta$, we have $\sup |V^\delta - V| \leq \epsilon_1$ and $\exists N, \forall n \geq N, \sup |V_n^\delta - V^\delta| \leq \epsilon_2$. Thus :

$$\sup_{x \in O} |V_n^\delta(x) - V(x)| \leq \sup_{x \in O} |V_n^\delta(x) - V^\delta(x)| + \sup_{x \in O} |V^\delta(x) - V(x)| \leq \epsilon_2 + \epsilon_1 \leq \epsilon$$

References

Akian, M.: 1990, *Méthodes multigrilles en contrôle stochastique*, PhD thesis, University Paris IX Dauphine.

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

Barles, G.: 1994, *Solutions de viscosité des équations de Hamilton-Jacobi*, Vol. 17 of *Mathématiques et Applications*, Springer-Verlag.

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

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

Barto, A. G.: 1990, *Neural networks for control*, W.T. Miller, R.S.Sutton, P.J. Werbos editors. MIT press, Cambridge, Massachussetts.

Barto, A. G., Bradtke, S. J. and Singh, S. P.: 1991, Real-time learning and control using asynchronous dynamic programming, *Technical Report 91-57*, Computer Science Department, University of Massachusetts.

Barto, A. G., Sutton, R. S. and Anderson, C. W.: 1983, *Neuronlike adaptive elements that can solve difficult learning control problems*, IEEE Transactions on Systems, Man and Sybernetics.

Bertsekas, D. P.: 1987, *Dynamic Programming : Deterministic and Stochastic Models*, Prentice Hall.

Crandall, M., Ishii, H. and Lions, P.: 1992, User's guide to viscosity solutions of second order partial differential equations, *Bulletin of the American Mathematical Society* **27**(1).

Crandall, M. and Lions, P.: 1983, Viscosity solutions of hamilton-jacobi equations, *Trans. of the American Mathematical Society* **277**.

Fleming, W. H. and Soner, H. M.: 1993, *Controlled Markov Processes and Viscosity Solutions*, Applications of Mathematics, Springer-Verlag.

Gullapalli, V.: 1992, *Reinforcement Learning and its application to control*, PhD thesis, University of Massachussetts, Amherst.

Harmon, M. E., Baird, L. C. and Klopf, A. H.: 1996, Reinforcement learning applied to a differential game, *Adaptive Behavior* **4**, 3–28.

Kushner, H. J.: 1990, Numerical methods for stochastic control problems in continuous time, *SIAM J. Control and Optimization* **28**, 999–1048.

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

Lin, L.-J.: 1993, *Reinforcement Learning for Robots using Neural Networks*, PhD thesis, Carnegie Mellon University, Pittsburg, Pennsylvania.

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

Souganidis, P. E.: 1985, Approximation schemes for viscosity solutions of hamilton-jacobi equations, *Journal of Differential Equations* **59**, 1–43.

Watkins, C. J.: 1989, *Learning from delayed reward*, PhD thesis, Cambridge University.

Watkins, C. J. and Dayan, P.: 1992, Q-learning, *Machine Learning* **8**, 279–292.