Using Finite Differences methods for approximating the value function of continuous Reinforcement Learning problems

Rémi Munos*
DASSAULT-AVIATION, DGT-DTN-EL-Et.Avancées
78 quai Marcel Dassault, 92214 Saint-Cloud, FRANCE
Tel : (0)1 40 96 61 21 Poste : 64 14. Fax : (0)1 40 96 60 80
E-mail : Remi.Munos@cemagref.fr

Abstract

This paper presents a reinforcement learning method for solving continuous optimal control problems when the dynamics of the system is unknown. First, we use a Finite Differences method for discretizing the Hamilton-Jacobi-Bellman equation and obtain a finite Markovian Decision Process. This permits us to approximate the value function of the continuous problem with piecewise constant functions defined on a grid. Then we propose to solve this MDP on-line with the available knowledge using a direct and convergent reinforcement learning algorithm, called the *Finite-Differences Reinforcement Learning.*

1 INTRODUCTION

This paper proposes an adaptive method for solving optimal control problems, like target or obstacle problems, viability or optimization problems when the time and the state space are continuous variables. Reinforcement Learning (RL) techniques are kinds of Dynamic Programming (DP) methods which generate an optimal feed-back control by computing the *value function* (VF) defined as the best expected cumulative reinforcement for each state. A local condition for the VF is given by the Hamilton-Jacobi-Bellman (HJB) equation. Here we use Finite-Differences (FD) methods for discretizing the HJB equation. We obtain a DP equation for a finite Markovian Decision Process (MDP) whose solution generates a piecewise constant function, defined on a regular grid, that approximates the VF.
Section 2 describes a formalism for optimal control problems in the continuous case. *Section 3* defines the value function, presents the RL approach and states the HJB equation. *Section 4* describes the Finite-Differences approximation used here and states the associated DP equation. *Section 5* presents a direct learning algorithm for computing the approximated value function : the *Finite-Differences Reinforcement Learning* (FDRL).

* CEMAGREF, LISC, Parc de Tourvoie, BP 121, 92185 Antony Cadex, FRANCE

2 OPTIMAL CONTROL PROBLEMS IN THE CONTINUOUS CASE

In this paper, we are interested in *deterministic infinite time* problems with *discounted* reinforcement. We consider a controlled dynamical system whose state $x(t) \in \overline{O}$ the state space with O open subset of $\mathbb{R}^n$. Its evolution depends on a differential equation, called its dynamics :

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

where the control $u(t)$ is a function with values in a finite set 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 $r : \bar{O} \times U \to \mathbb{R}$ is the *running reinforcement* and $R : \partial O \to \mathbb{R}$ the *terminal reinforcement*. γ is the *discount factor* $(0 \leq \gamma < 1)$.

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

3 THE REINFORCEMENT LEARNING APPROACH

RL techniques belongs to the class of DP methods which compute the optimal control by the building of the *value function*, defined here by :

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

In the RL approach, the system tries to approximate this function without knowing the dynamics f. Following the DP principle, the value function satisfies a local condition called the *Hamilton-Jacobi-Bellman* equation. The following theorem comes from Bellman optimality principle in the continuous case (See [4] for a proof) :

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 following HJB equation holds at $x \in O$.*

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

The challenge of learning the VF is motivated by the fact that from the VF we can deduce the optimal control :

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

4 APPROXIMATION WITH FINITE-DIFFERENCES METHODS

Let $e_1, e_2, ..., e_n$ be a basis for $\mathbb{R}^n$. The dynamics is : $f = (f_1, ..., f_n)$. 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 lattice : $\Sigma^\delta = \{\delta . \sum_{i=1}^n j_i e_i\} \cap O$ where $j_1, ..., j_n$ are any integers. The *interior* of Σ^δ is the set of points $\xi \in \Sigma^\delta$ such that all nearest neighboor points $\xi \pm \delta e_i$ are in Σ^δ. 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. We assume that Σ^δ approximates O in the sense that : $\lim_{\delta \searrow 0} dist(\partial O, \partial\Sigma^\delta) = 0$.

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 :
- For ξ interior to Σ^δ,

$$V^\delta(\xi)\ln\gamma + \sup_{v\in U}\left\{\sum_{i=1..n}\left[f_i^+(\xi,v).\Delta_i^+ V^\delta(\xi) + f_i^-(\xi,v).\Delta_i^- V^\delta(\xi)\right] + r(\xi,v)\right\} = 0$$

- For $\xi \in \partial\Sigma^\delta$, $\qquad V^\delta(\xi) = R(\xi)$

In the following of the paper, we use the norm : $||u||_1 = \sum_{i=1}^n |u_i|$. Knowing that $(\Delta t \ln\gamma)$ is an approximation of $(\gamma^{\Delta t} - 1)$ as Δt tends to 0, this equation can be written as a DP equation for a finite MDP :

$$V^\delta(\xi) = \sup_{v\in U}\left\{\gamma^{\frac{\delta}{||f(x,v)||_1}} \sum_{\xi'\in\Sigma^\delta} p(\xi,v,\xi').V^\delta(y) + \frac{\delta}{||f(\xi,v)||_1} r(\xi,v)\right\} \tag{1}$$

whose *state space* is Σ^δ and whose *probabilities of transition* from :
(state ξ, control v) to the possible next states ξ' are :

$$\begin{aligned} \text{If } \xi \text{ is interior to } \Sigma^\delta, \qquad p(\xi,v,\xi') &= \frac{f_i^+(\xi,v)}{||f(\xi,v)||_1} \text{ if } \xi' = \xi + \delta e_i \\ &= \frac{f_i^-(\xi,v)}{||f(\xi,v)||_1} \text{ if } \xi' = \xi - \delta e_i \\ &= 0 \text{ otherwise} \\ \text{If } \xi \in \partial\Sigma^\delta, \qquad p(\xi,v,\xi') &= 1 \text{ if } \xi = \xi' \\ &= 0 \text{ otherwise} \end{aligned}$$

Thanks to a contraction property due to the discount factor γ (see [3]), DP theory insures that there exists a unique solution V^δ to equation 1. The following theorem insures that this solution V^δ is a good approximation of V.

Theorem 2 (Convergence of the Finite Differences scheme) *Let us assume that some smoothness assumptions described in [6] are satisfied, then the fixed-point V^δ of the DP equation (1) converges to the VF as δ tends to 0 :*

$$\lim_{\substack{\delta \searrow 0 \\ \xi \longrightarrow x}} V^\delta(\xi) = V(x) \text{ for all } x \in O$$

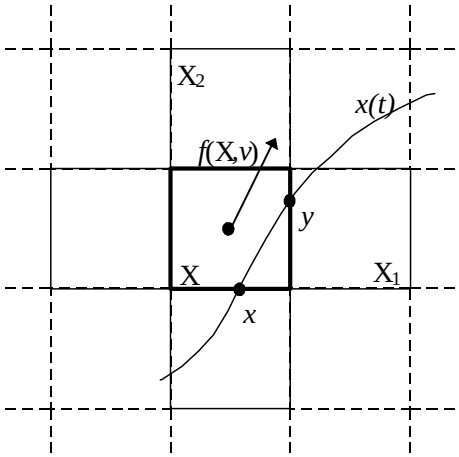
The proof of this theorem uses the general convergence results of [2].

5 THE FINITE-DIFFERENCES REINFORCEMENT LEARNING

Let us consider a grid G^δ composed of regular cells $\{X_i\}$ such that the center of the cells are the previously defined vertices of the lattice. Let us denote $f(X, v)$ the dynamics at the center of cell X for control v (see figure 1).

In the direct RL approach, the system does not know the dynamics f, so we need to approximate the probabilities of transition $p(\xi, v, \xi')$ with the available knowledge.

Figure 1: The grid G^δ is composed of regular cells X. The dynamics $f(X, v)$ is approximated by $\frac{y-x}{\tau_X}$ where x and y are the input and output points of a trajectory $x(t)$ crossing the cell X and τ_X is the running time of the trajectory inside X.



Let a trajectory $x(t)$ enters cell X at point x ; then a control v is chosen and kept until the trajectory exits at some point y. Let $r(X, v)$ be the cumulative current reinforcement obtained inside cell X. Let X_i be the next cell (for example cell X_1 in figure 1).

We use the value $\frac{y-x}{\tau_X}$ for approximating $f(X, v)$. Let τ_X be the running time of the trajectory inside X. The DP equation leads to the FDRL updating rules :

$$Q^\delta_{n+1}(X, v, X_i) = \gamma^{\frac{\delta}{||y-x||_1}.\tau_X}.\frac{|y_i - x_i|}{||y-x||_1}.V^\delta_n(X_i) \quad (2)$$

$$r^\delta_{n+1}(X, v, X_i) = \frac{1}{n}.\frac{\delta}{||y-x||_1}.r(X, v) \quad (3)$$

$$\text{with the values :} \qquad Q_n^\delta(X,v) = \sum_{i=1..n} \left[Q_n^\delta(X,v,X_i) + r_n^\delta(X,v,X_i) \right]$$
$$V_n^\delta(X) = \sup_{v \in U} Q_n^\delta(X,v)$$

The algorithm is the following : when the system crosses cell X with a control v and enters X_i then update the Q^δ and r^δ values with rules (2) and (3). Then the current optimal control v^* in cell X is the one that optimizes $Q_n^\delta(X,v)$.
We have the following theorem that states that the values computed by the algorithm converge to the VF of the continuous problem :

Theorem 3 (Convergence of the algorithm) *When experimenting the FDRL, we consider series of trajectories such that the algorithm leads to update every combination (X,v,X_i) of current cell X, control v and next cell X_i infinitely often. Suppose that some smoothness assumptions are satisfied (see [6]), then :*
$\forall \epsilon > 0$, $\exists \Delta$ s.t. $\forall \delta \leq \Delta$, for any grid G^δ, using the FDRL algorithm, $\exists N, \forall n \geq N$,

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

The proof of this theorem is very similar to the one given in [6] for the Finite-Element method.

Remark: for a particular grid G^δ, the V_n^δ-values do not converge. The theorem states that the convergence occurs as the number of iterations tends to infinity and the discretization step δ tends to 0. Thus, for computational aspects, the learning process (which computes iteratively the V_n^δ-values with rules (2) and (3)) has to be combinate with a grid refinement process.

6 CONCLUSION

This paper proposes an algorithm for solving optimal control problems in the continuous case using RL techniques. We use Finite-Differences method for approximating the HJB equation by a DP equation for a MDP. This equation is solved iteratively by a direct and convergent RL algorithm. In practical use of this algorithm, we are faced to the combinatorial explosion of the number of values to be estimated, in particular when the dimension of the state space is high. Future work should consider adaptive multigrid methods (like the parti-game algorithm of [5] or the multigrid methods of [1]) which use a local refinement process for the grid. An other improvement should be the study of the stochastic case.

References

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

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

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

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

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

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