Reinforcement Learning for Continuous Stochastic Control Problems

Rémi Munos
CEMAGREF, LISC, Parc de Tourvoie,
BP 121, 92185 Antony Cedex, FRANCE.
Remi.Munos@cemagref.fr

Paul Bourgine
Ecole Polytechnique, CREA,
91128 Palaiseau Cedex, FRANCE.
Bourgine@poly.polytechnique.fr

Abstract

This paper is concerned with the problem of Reinforcement Learning (RL) for continuous state space and time stochastic control problems. We state the Hamilton-Jacobi-Bellman equation satisfied by the value function and use a Finite-Difference method for designing a convergent approximation scheme. Then we propose a RL algorithm based on this scheme and prove its convergence to the optimal solution.

1 Introduction to RL in the continuous, stochastic case

The objective of RL is to find -thanks to a reinforcement signal- an optimal strategy for solving a dynamical control problem. Here we study the continuous time, continuous state-space stochastic case, which covers a wide variety of control problems including target, viability, optimization problems (see [FS93], [KP95]) for which a formalism is the following. The evolution of the current state $x(t) \in \mathcal{O}$ (the state-space, with $\mathcal{O}$ open subset of $\mathbb{R}^d$), depends on the control $u(t) \in U$ (compact subset) by a stochastic differential equation, called the state dynamics:

$$dx = f(x(t), u(t))dt + \sigma(x(t), u(t))dw$$

where $f$ is the local drift and $\sigma dw$ (with $w$ a brownian motion of dimension $r$ and $\sigma$ a $d \times r$-matrix) the stochastic part (which appears for several reasons such as lack of precision, noisy influence, random fluctuations) of the diffusion process.

For initial state $x$ and control $u(t)$, (1) leads to an infinity of possible trajectories $x(t)$. For some trajectory $x(t)$ (see figure 1), let $\tau$ be its exit time from $\mathcal{O}$ (with the convention that if $x(t)$ always stays in $\mathcal{O}$, then $\tau = \infty$). Then, we define the functional $J$ of initial state $x$ and control $u(.)$ as the expectation for all trajectories of the discounted cumulative reinforcement:

$$J(x; u(.)) = E_{x, u(.)} \left\{ \int_0^\tau \gamma^t r(x(t), u(t))dt + \gamma^\tau R(x(\tau)) \right\}$$
where \( r(x, u) \) is the running reinforcement and \( R(x) \) the boundary reinforcement. \( \gamma \) is the discount factor \((0 \leq \gamma < 1)\). In the following, we assume that \( f, \sigma \) are of class \( C^2 \) and \( R \) are Lipschitzian (with constants \( L_r \) and \( L_R \)) and the boundary \( \partial O \) is \( C^2 \).

Figure 1: The state space, the discretized \( \Sigma^d \) (the square dots) and its frontier \( \partial \Sigma^d \) (the round ones). A trajectory \( x_k(t) \) goes through the neighbourhood of state \( \xi \).

RL uses the method of Dynamic Programming (DP) which generates an optimal (feed-back) control \( u^*(x) \) by estimating the value function (VF), defined as the maximal value of the functional \( J \) as a function of initial state \( x \):

\[
V(x) = \sup_{u(\cdot)} J(x; u(\cdot)).
\]  

In the RL approach, the state dynamics is unknown from the system; the only available information for learning the optimal control is the reinforcement obtained at the current state. Here we propose a model-based algorithm, i.e. that learns on-line a model of the dynamics and approximates the value function by successive iterations.

Section 2 states the Hamilton-Jacobi-Bellman equation and use a Finite-Difference (FD) method derived from Kushner [Kus90] for generating a convergent approximation scheme. In section 3, we propose a RL algorithm based on this scheme and prove its convergence to the VF in appendix A.

2 A Finite Difference scheme

Here, we state a second-order nonlinear differential equation (obtained from the DP principle, see [FS93]) satisfied by the value function, called the Hamilton-Jacobi-Bellman equation.

Let the \( d \times d \) matrix \( a = \sigma \sigma' \) (with ' the transpose of the matrix). We consider the uniformly parabolic case, i.e. we assume that there exists \( c > 0 \) such that \( \forall x \in \bar{O}, \forall u \in U, \forall y \in \mathbb{R}^d, \sum_{i,j=1}^d a_{ij}(x, u) y_i y_j \geq c ||y||^2 \). Then \( V \) is \( C^2 \) (see [Kry80]). Let \( V_x \) be the gradient of \( V \) and \( V_{xixj} \) its second-order partial derivatives.

**Theorem 1 (Hamilton-Jacobi-Bellman)** The following HJB equation holds:

\[
V(x) \ln \gamma + \sup_{u \in U} \left[ r(x, u) + V_x(x).f(x, u) + \frac{1}{2} \sum_{i,j=1}^n a_{ij}(x) V_{xixj}(x) \right] = 0 \text{ for } x \in \bar{O}
\]

Besides, \( V \) satisfies the following boundary condition: \( V(x) = R(x) \) for \( x \in \partial O \).
Remark 1 The challenge of learning the VF is motivated by the fact that from \( V \), we can deduce the following optimal feed-back control policy:

\[
  u^*(x) \in \arg \sup_{u \in U} \left[ r(x, u) + V(x).f(x, u) + \frac{1}{2} \sum_{i,j=1}^{n} a_{ij} V(x, x_j)(x) \right]
\]

In the following, we assume that \( O \) is bounded. Let \( e_1, \ldots, e_d \) be a basis for \( \mathbb{R}^d \). Let the positive and negative parts of a function \( \phi \) be : \( \phi^+ = \max(\phi, 0) \) and \( \phi^- = \max(-\phi, 0) \). For any discretization step \( \delta \), let us consider the lattices : \( \delta \mathbb{Z}^d = \{ \delta \sum_{i=1}^{d} j_i e_i \} \) where \( j_1, \ldots, j_d \) are any integers, and \( \Sigma^d = \delta \mathbb{Z}^d \cap O \). Let \( \partial \Sigma^d \), the frontier of \( \Sigma^d \) denote the set of points \( \{ \xi \in \delta \mathbb{Z}^d \setminus O \) such that at least one adjacent point \( \xi \pm \delta e_i \in \Sigma^d \} \) (see figure 1).

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 \( \cup \delta \delta U^\delta = U \). Besides, we assume that: \( \forall i = 1..d, \\
  a_{ii}(x, u) - \sum_{j \neq i} |a_{ij}(x, u)| \geq 0. \) (3)

By replacing the gradient \( V_x(\xi) \) by the forward and backward first-order finite-difference quotients: \( \Delta_{x_i}^\pm V(\xi) = \frac{1}{\delta} [V(\xi + \delta e_i) + V(\xi - \delta e_i) - 2V(\xi)] \)

\( \Delta_{x_i x_j}^\pm V(\xi) = \frac{1}{2\delta^2} [V(\xi + \delta e_i \pm \delta e_j) + V(\xi - \delta e_i \mp \delta e_j) - V(\xi + \delta e_i) - V(\xi - \delta e_i) - V(\xi + \delta e_j) - V(\xi - \delta e_j) + 2V(\xi)] \)

in the HJB equation, we obtain the following : for \( \xi \in \Sigma^d \),

\[
  V^\delta(\xi) \ln \gamma + \sup_{u \in U^\delta} \left\{ r(\xi, u) + \sum_{i=1}^{d} [f_{i}^+(\xi, u).\Delta_{x_i}^+ V^\delta(\xi) - f_{i}^-(\xi, u).\Delta_{x_i}^- V^\delta(\xi)] + \frac{a_{ii}(\xi, u)}{2} \Delta_{x_i x_i} V^\delta(\xi) + \sum_{j \neq i} \left( \frac{a_{ij}(\xi, u)}{2} \Delta_{x_i x_j}^+ V^\delta(\xi) - \frac{a_{ij}(\xi, u)}{2} \Delta_{x_i x_j}^- V^\delta(\xi) \right) \right\} = 0
\]

Knowing that \( (\Delta t \ln \gamma) \) is an approximation of \( (\gamma^{\Delta t} - 1) \) as \( \Delta t \) tends to 0, we deduce:

\[
  V^\delta(\xi) = \sup_{u \in U^\delta} \left[ \gamma r(\xi, u) \sum_{\xi \in \Sigma^d} p(\xi, u, \xi) V^\delta(\xi) + \tau(\xi, u) r(\xi, u) \right]
\]

with \( \tau(\xi, u) = \frac{\delta^2 \sum_{i=1}^{d} \left[ |f_i(\xi, u)| + a_{ii}(\xi, u) - \frac{1}{2} \sum_{j \neq i} |a_{ij}(\xi, u)| \right]}{2} \)

which appears as a DP equation for some finite Markovian Decision Process (see [Ber87]) whose state space is \( \Sigma^d \) and probabilities of transition:

\[
  p(\xi, u, x) = \frac{\tau(\xi, u)}{2 \delta^2} \left[ 2|f_i^+(\xi, u)| + a_{ii}(\xi, u) - \sum_{j \neq i} |a_{ij}(\xi, u)| \right],
\]

\[
  p(\xi, u, x) = \frac{\tau(\xi, u)}{2 \delta^2} a_{ij}^+(\xi, u) \text{ for } i \neq j,
\]

\[
  p(\xi, u, x) = \frac{\tau(\xi, u)}{2 \delta^2} a_{ij}^-(\xi, u) \text{ for } i \neq j,
\]

\[
  p(\xi, u, \xi) = 0 \text{ otherwise.}
\]

Thanks to a contraction property due to the discount factor \( \gamma \), there exists a unique solution (the fixed-point) \( V^\delta \) to equation (4) for \( \xi \in \Sigma^d \) with the boundary condition \( V^\delta(\xi) = R(\xi) \) for \( \xi \in \partial \Sigma^d \). The following theorem (see [Kus90] or [FS93]) insures that \( V^\delta \) is a convergent approximation scheme.
Theorem 2 (Convergence of the FD scheme) $V^\delta$ converges to $V$ as $\delta \downarrow 0$:
\[
\lim_{\delta \to 0} V^\delta(x) = V(x) \quad \text{uniformly on } \overline{\mathcal{O}}
\]

Remark 2 Condition (3) insures that the $p(\xi, u, \zeta)$ are positive. If this condition does not hold, several possibilities to overcome this are described in [Kus90].

3 The reinforcement learning algorithm

Here we assume that $f$ is bounded from below. As the state dynamics ($f$ and $a$) is unknown from the system, we approximate it by building a model $\tilde{f}$ and $\tilde{a}$ from samples of trajectories $x_k(t)$: we consider series of successive states $x_k = x_k(t_k)$ and $y_k = x_k(t_k + \tau_k)$ such that:
- $\forall t \in [t_k, t_k + \tau_k], \quad x(t) \in N(\xi)$ neighbourhood of $\xi$ whose diameter is inferior to $k_N \cdot \delta$ for some positive constant $k_N$,
- the control $u$ is constant for $t \in [t_k, t_k + \tau_k]$,
- $\tau_k$ satisfies for some positive $k_1$ and $k_2$, 
  \[
  k_1 \delta \leq \tau_k \leq k_2 \delta.
  \]

Then incrementally update the model:
\[
\tilde{f}_n(\xi, u) = \frac{1}{n} \sum_{k=1}^{n} \frac{y_k - x_k}{\tau_k}
\]
\[
\tilde{a}_n(\xi, u) = \frac{1}{n} \sum_{k=1}^{n} \frac{(y_k - x_k - \tau_k \cdot \tilde{f}_n(\xi, u)) (y_k - x_k - \tau_k \cdot \tilde{f}_n(\xi, u))}{\tau_k}
\]
and compute the approximated time $\tilde{\tau}(x, u)$ and the approximated probabilities of transition $\tilde{p}(\xi, u, \zeta)$ by replacing $f$ and $a$ by $\tilde{f}$ and $\tilde{a}$ in (5) and (6).

We obtain the following updating rule of the $V^\delta$-value of state $\xi$:
\[
V_{n+1}^\delta(\xi) = \sup_{u \in U^\delta} \left[ \gamma \tilde{\tau}(x, u) \sum_\zeta \tilde{p}(\xi, u, \zeta) V_n^\delta(\zeta) + \tilde{\tau}(x, u) \tilde{r}(\xi, u) \right]
\]
which can be used as an off-line (synchronous, Gauss-Seidel, asynchronous) or online (for example by updating $V_n^\delta(\xi)$ as soon as a trajectory exits from the neighbourhood of $\xi$) DP algorithm (see [BBS95]).

Besides, when a trajectory hits the boundary $\partial \mathcal{O}$ at some exit point $x_k(\tau)$ then update the closest state $\xi \in \partial \mathcal{O}^\delta$ with:
\[
V_n^\delta(\xi) = R(x_k(\tau))
\]

Theorem 3 (Convergence of the algorithm) Suppose that the model as well as the $V^\delta$-value of every state $\xi \in \mathcal{O}^\delta$ and control $u \in U^\delta$ are regularly updated (respectively with (8) and (9)) and that every state $\xi \in \partial \mathcal{O}^\delta$ are updated with (10) at least once. Then $\forall \varepsilon > 0$, $\exists \Delta$ such that $\forall \delta \leq \Delta$, $\exists N, \forall n \geq N$,
\[
\sup_{\xi \in \mathcal{O}^\delta} |V_n^\delta(\xi) - V(\xi)| \leq \varepsilon \quad \text{with probability 1}
\]
4 Conclusion

This paper presents a model-based RL algorithm for continuous stochastic control problems. A model of the dynamics is approximated by the mean and the covariance of successive states. Then, a RL updating rule based on a convergent FD scheme is deduced and in the hypothesis of an adequate exploration, the convergence to the optimal solution is proved as the discretization step $\delta$ tends to 0 and the number of iteration tends to infinity. This result is to be compared to the model-free RL algorithm for the deterministic case in [Mun97]. An interesting possible future work should be to consider model-free algorithms in the stochastic case for which a Q-learning rule (see [Wat89]) could be relevant.

A Appendix: proof of the convergence

Let $M_f, M_a, M_{f_x}$ and $M_{\sigma_x}$ be the upper bounds of $f, a, f_x$ and $\sigma_x$ and $m_f$ the lower bound of $f$. Let $E^\delta = \sup_{\xi \in \Sigma^n} |V^\delta(\xi) - V(\xi)|$ and $E_n^\delta = \sup_{\xi \in \Sigma^n} |V_n^\delta(\xi) - V^\delta(\xi)|$.

A.1 Estimation error of the model $\widetilde{f}_n$ and $\widetilde{a}_n$ and the probabilities $\tilde{p}_n$

Suppose that the trajectory $x_k(t)$ occurred for some occurrence $w_k(t)$ of the brownian motion: $x_k(t) = x_k + \int_{t_k}^t f(x_k(t), u) dt + \int_{t_k}^t \sigma(x_k(t), u) dw_k$. Then we consider a trajectory $z_k(t)$ starting from $\xi$ at $t_k$ and following the same brownian motion:

$z_k(t) = \xi + \int_{t_k}^t f(z_k(t), u) dt + \int_{t_k}^t \sigma(z_k(t), u) dw_k$.

Let $z_k = z_k(t_k + \tau_k)$. Then $(y_k - x_k) - (z_k - \xi) = \int_{t_k}^t [f(x_k(t), u) - f(z_k(t), u)] dt + \int_{t_k}^{t_k + \tau_k} [\sigma(x_k(t), u) - \sigma(z_k(t), u)] dw_k$. Thus, from the $C^1$ property of $f$ and $\sigma$,

$$\|(y_k - x_k) - (z_k - \xi)\| \leq (M_{f_x} + M_{\sigma_x}) \cdot kN \cdot \tau_k \cdot \delta.$$  \hfill (11)

The diffusion processes has the following property (see for example the Itô-Taylor majoration in [KP95]): $E_x[\xi + \tau_k f(\xi, u) + O(\tau_k^2)]$ which, from (7), is equivalent to: $E_x[\frac{\xi - \xi_k}{\tau_k}] = f(\xi, u) + O(\delta)$. Thus from the law of large numbers and (11):

$$\lim_{n \to \infty} \sup_{\xi \in \Sigma^n} \left\| \bar{f}_n(\xi, u) - f(\xi, u) \right\| = \lim_{n \to \infty} \sup_{\xi \in \Sigma^n} \left\| \frac{1}{n} \sum_{k=1}^{n} \left[ \frac{\xi_k - \xi}{\tau_k} - \frac{\xi_k - \xi}{\tau_k} \right] \right\| + O(\delta) = (M_{f_x} + M_{\sigma_x}) \cdot kN \cdot \delta + O(\delta) \text{ w.p. 1} \hfill (12)$$

Besides, diffusion processes have the following property (again see [KP95]): $E_x[(z_k - \xi)(z_k - \xi)] = a(\xi, u) \tau_k + O(\tau_k^2)$ which, from (7), is equivalent to: $E_x\left[ \frac{(z_k - \xi)(z_k - \xi)}{\tau_k} \right] = a(\xi, u) + O(\delta^2)$. Let $\tau_k = z_k - \xi - \tau_k f(\xi, u)$ and $\tilde{r}_k = y_k - x_k - \tau_k f_n(\xi, u)$ which satisfy (from (11) and (12)):

$$\|\tau_k - \tilde{r}_k\| = (M_{f_x} + M_{\sigma_x}) \cdot kN \cdot \delta + O(\delta) \hfill (13)$$

From the definition of $\tilde{a}_n(\xi, u)$, we have: $\tilde{a}_n(\xi, u) - a(\xi, u) = \frac{1}{n} \sum_{k=1}^{n} \frac{\tilde{r}_k - \tilde{r}_k'}{\tau_k}$ and from the law of large numbers, (12) and (13), we have:

$$\lim_{n \to \infty} \sup_{\xi \in \Sigma^n} \left\| \tilde{a}_n(\xi, u) - a(\xi, u) \right\| = \lim_{n \to \infty} \sup_{\xi \in \Sigma^n} \left\| \frac{1}{n} \sum_{k=1}^{n} \frac{\tilde{r}_k - \tilde{r}_k'}{\tau_k} \right\| + O(\delta^2) = \|\tilde{r}_k - \tau_k\| \lim_{n \to \infty} \sup_{\xi \in \Sigma^n} \left( \left\| \frac{\tilde{r}_k}{\tau_k} \right\| \right) + O(\delta^2) = O(\delta^2) = O(\delta^2)$$
with probability 1. Thus there exists $k_f$ and $k_a$ s.t. $\exists \Delta_1, V \delta \leq \Delta_1, \exists N_1, n \geq N_1$, 
\begin{align}
\|f_n(\xi, u) - f(\xi, u)\| &\leq k_f . \delta \text{ w.p. } 1 \\
\|a_n(\xi, u) - a(\xi, u)\| &\leq k_a . \delta^2 \text{ w.p. } 1
\end{align}
(14)

Besides, from (5) and (14), we have:
\begin{align}
|\tau(\xi, u) - \tilde{\tau}_n(\xi, u)| \leq \frac{d (k_f \delta^2 + d . k_a \delta)}{(d . m_f \delta)^2} \delta^2 \leq k_r . \delta^2 \tag{15}
\end{align}

and from a property of exponential function,
\begin{align}
\left| \gamma^{\tau(\xi, u)} - \gamma^{\tilde{\tau}_n(\xi, u)} \right| = k_r . \ln \frac{1}{\gamma} . \delta^2. \tag{16}
\end{align}

We can deduce from (14) that:
\begin{align}
\limsup_{n \to \infty} |p(\xi, u, \zeta) - \tilde{p}_n(\xi, u, \zeta)| \leq \frac{(2 . \delta . M_f + d . M_a) (2 . k_f + d . k_a) \delta}{d m_f - (2 . k_f + d . k_a) \delta} \leq k_p \delta \text{ w.p. } 1 \tag{17}
\end{align}

with $k_p = 4 (d . M_a) (2 . k_f + d . k_a)$ for $\delta \leq \Delta_2 = \min \left\{ \frac{m_f}{2 . k_f + d . k_a}, \frac{d . M_a}{2 . k_f + d . k_a} \right\}$.

### A.2 Estimation of $|V_{n+1}^\delta(\xi) - V^\delta(\xi)|$

After having updated $V_{\xi}^\delta$ with rule (9), let $\Lambda$ denote the difference $|V_{n+1}^\delta(\xi) - V^\delta(\xi)|$. From (4), (9) and (8),
\begin{align}
\Lambda \leq \gamma^{\tau(\xi, u)} \sum_\zeta \left[ p(\xi, u, \zeta) - \tilde{p}(\xi, u, \zeta) \right] V^\delta(\zeta) + \left( \gamma^{\tau(\xi, u)} - \gamma^{\tilde{\tau}(\xi, u)} \right) \sum_\zeta \tilde{p}(\xi, u, \zeta) V^\delta(\zeta) \\
+ \gamma^{\tilde{\tau}(\xi, u)} \sum_\zeta \tilde{p}(\xi, u, \zeta) \left[ V^\delta(\zeta) - V_n^\delta(\zeta) \right] + \sum_\zeta \tilde{p}(\xi, u, \zeta) \tilde{\tau}(\xi, u) \left[ r(\xi, u) - \tilde{\tau}(\xi, u) \right] \\
+ \sum_\zeta \tilde{p}(\xi, u, \zeta) \left[ \tau(\xi, u) - \tilde{\tau}(\xi, u) \right] r(\xi, u) \text{ for all } u \in U^\delta
\end{align}

As $V$ is differentiable we have: $V(\zeta) = V(\xi) + V_\xi . (\zeta - \xi) + o(||\zeta - \xi||)$. Let us define a linear function $\tilde{V}$ such that: $\tilde{V}(x) = V(\xi) + V_\xi . (x - \xi)$. Then we have: $[p(\xi, u, \zeta) - \tilde{p}(\xi, u, \zeta)] V^\delta(\zeta) = [p(\xi, u, \zeta) - \tilde{p}(\xi, u, \zeta)] . [V^\delta(\zeta) - V_n^\delta(\zeta)] + [p(\xi, u, \zeta) - \tilde{p}(\xi, u, \zeta)] V^\delta(\zeta)$, thus: $\sum_\zeta [p(\xi, u, \zeta) - \tilde{p}(\xi, u, \zeta)] V^\delta(\zeta) = k_p . E^\delta . \delta + \sum_\zeta [p(\xi, u, \zeta) - \tilde{p}(\xi, u, \zeta)] \left[ \tilde{V}(\zeta) + o(\delta) \right] = \left[ \tilde{V}(\eta) - \tilde{V}(\tilde{\eta}) \right] + o(\delta)$ with: $\eta = \sum_\zeta p(\xi, u, \zeta) (\zeta - \xi)$ and $\tilde{\eta} = \sum_\zeta \tilde{p}(\xi, u, \zeta) (\zeta - \xi)$. Besides, from the convergence of the scheme (theorem 2), we have $E^\delta . \delta = o(\delta)$. From the linearity of $\tilde{V}$, $\left| \tilde{V}(\zeta) - \tilde{V}(\tilde{\zeta}) \right| \leq \left\| \zeta - \tilde{\zeta} \right\| M_V \leq 2 k_p \delta^2$. Thus
\begin{align}
\sum_\zeta [p(\xi, u, \zeta) - \tilde{p}(\xi, u, \zeta)] V^\delta(\zeta) = o(\delta) \text{ and from (15), (16) and the Lipschitz property of } r,
\end{align}
\begin{align}
\Lambda = \left| \gamma^{\tilde{\tau}(\xi, u)} \sum_\zeta \tilde{p}(\xi, u, \zeta) \left[ V^\delta(\zeta) - V_n^\delta(\zeta) \right] \right| + o(\delta).
\end{align}

As $\gamma^{\tilde{\tau}(\xi, u)} \leq 1 - \frac{\tau(\xi, u) - \tau(\tilde{\xi}, u)}{2} \ln \frac{1}{\gamma} \leq 1 - \frac{\tau(\xi, u) - k_f \delta^2}{2} \ln \frac{1}{\gamma} \leq 1 - \frac{(2d(M_f + d . M_a) - k_p \delta^2) \ln \frac{1}{\gamma}},$
we have:
\begin{align}
\Lambda = (1 - k . \delta) E_n^\delta + o(\delta) \tag{18}
\end{align}

with $k = \frac{1}{2d(M_f + d . M_a)}$. 

A.3 A sufficient condition for $\sup_{\xi \in \Sigma^d} |V_\delta^\delta(\xi) - V_\delta^\delta(\xi)| \leq \epsilon_2$

Let us suppose that for all $\xi \in \Sigma^d$, the following conditions hold for some $\alpha > 0$

$$E^\delta_n > \epsilon_2 \Rightarrow |V_{n+1}^\delta(\xi) - V_\delta^\delta(\xi)| \leq E^\delta_n - \alpha \quad (19)$$

$$E^\delta_n \leq \epsilon_2 \Rightarrow |V_{n+1}^\delta(\xi) - V_\delta^\delta(\xi)| \leq \epsilon_2 \quad (20)$$

From the hypothesis that all states $\xi \in \Sigma^d$ are regularly updated, there exists an integer $m$ such that at stage $n + m$ all the $\xi \in \Sigma^d$ have been updated at least once since stage $n$. Besides, since all $\xi \in \partial G^d$ are updated at least once with rule (10), $\forall \xi \in \partial G^d$, $|V_n^\delta(\xi) - V_\delta^\delta(\xi)| = |R(x_k(\tau)) - R(\xi)| \leq 2L_R \delta \leq \epsilon_2$ for any $\delta \leq \Delta_3 = \frac{\epsilon_2^2}{2L_R}$. Thus, from (19) and (20) we have:

$$E^\delta_n > \epsilon_2 \Rightarrow E^\delta_{n+m} \leq E^\delta_n - \alpha$$

$$E^\delta_n \leq \epsilon_2 \Rightarrow E^\delta_{n+m} \leq \epsilon_2$$

Thus there exists $N$ such that: $\forall n \geq N$, $E^\delta_n \leq \epsilon_2$.

A.4 Convergence of the algorithm

Let us prove theorem 3. For any $\epsilon > 0$, let us consider $\epsilon_1 > 0$ and $\epsilon_2 > 0$ such that $\epsilon_1 + \epsilon_2 = \epsilon$. Assume $E^\delta_n > \epsilon_2$, then from (18), $A = E^\delta_n - k.\delta.\epsilon_2 + o(\delta) \leq E^\delta_n - k.\delta.\frac{\epsilon_2^2}{2}$ for $\delta \leq \Delta_3$. Thus (19) holds for $\alpha = k.\delta.\frac{\epsilon_2^2}{2}$. Suppose now that $E^\delta_n \leq \epsilon_2$. From (18), $A \leq (1 - k.\delta)\epsilon_2 + o(\delta) \leq \epsilon_2$ for $\delta \leq \Delta_3$ and condition (20) is true.

Thus for $\delta \leq \min\{\Delta_1, \Delta_2, \Delta_3\}$, the sufficient conditions (19) and (20) are satisfied. So there exists $N$, for all $n \geq N$, $E^\delta_n \leq \epsilon_2$. Besides, from the convergence of the scheme (theorem 2), there exists $\Delta_0$ st. $\forall \delta \leq \Delta_0$, $\sup_{\xi \in \Sigma^d} |V_\delta^\delta(\xi) - V(\xi)| \leq \epsilon_1$.

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

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

References


