Voronoi model learning for batch mode reinforcement learning

Raphael Fonteneau Damien Ernst

Department of Electrical Engineering and Computer Science,
University of Liège, BELGIUM

2010

Abstract

We consider deterministic optimal control problems with continuous state spaces where the information on the system dynamics and the reward function is constrained to a set of system transitions. Each system transition gathers a state, the action taken while being in this state, the immediate reward observed and the next state reached. In such a context, we propose a new model learning–type reinforcement learning (RL) algorithm in batch mode, finite-time and deterministic setting. The algorithm, named Voronoi reinforcement learning (VRL), approximates from a sample of system transitions the system dynamics and the reward function of the optimal control problem using piecewise constant functions on a Voronoi–like partition of the state-action space.

1 Problem statement

We consider a discrete-time system whose dynamics over T stages is described by a time-invariant equation

$$x_{t+1} = f(x_t, u_t) \quad t = 0, 1, \dots, T-1, \tag{1}$$

where for all $t \in \{0, \dots, T-1\}$, the state x_t is an element of the bounded normed state space $\mathcal{X} \subset \mathbb{R}^{d_{\mathcal{X}}}$ and u_t is an element of a finite action space $\mathcal{U} = \{a^1, \dots, a^m\}$ with $m \in \mathbb{N}_0$. $x_0 \in \mathcal{X}$ is the initial state of the system. $T \in \mathbb{N}_0$ denotes the finite optimization horizon. An instantaneous reward

$$r_t = \rho(x_t, u_t) \in \mathbb{R} \tag{2}$$

is associated with the action $u_t \in \mathcal{U}$ taken while being in state $x_t \in \mathcal{X}$. We assume that the initial state of the system $x_0 \in \mathcal{X}$ is fixed. For a given open-loop sequence of actions $\mathbf{u} = (u_0, \dots, u_{T-1}) \in \mathcal{U}^T$, we denote by $J^{\mathbf{u}}(x_0)$ the $T-$stage return of the sequence of actions $\mathbf{u}$ when starting from x_0, defined as follows:

Definition 1.1 ($T-$stage return)
$\forall \mathbf{u} \in \mathcal{U}^T, \forall x_0 \in \mathcal{X}$,

$$J^{\mathbf{u}}(x_0) = \sum_{t=0}^{T-1} \rho(x_t, u_t) \tag{3}$$

with

$$x_{t+1} = f(x_t, u_t), \forall t \in \{0, \ldots, T-1\} \; . \quad (4)$$

We denote by $J^*(x_0)$ the maximal value:

Definition 1.2 (Maximal return)
$\forall x_0 \in \mathcal{X}$,

$$J^*(x_0) = \max_{\mathbf{u} \in \mathcal{U}^T} J^{\mathbf{u}}(x_0) \; . \quad (5)$$

Considering the fixed initial state x_0, an optimal sequence of actions $\mathbf{u}^*(x_0)$ is a sequence for which

$$J^{\mathbf{u}^*(x_0)}(x_0) = J^*(x_0) \; . \quad (6)$$

In this report, we assume that the functions f and ρ are unknown. Instead, we know a sample of n system transitions

$$\mathcal{F}_n = \left\{ \left(x^l, u^l, r^l, y^l \right) \right\}_{l=1}^{n} \quad (7)$$

where for all $l \in \{1, \ldots, n\}$

$$r^l = \rho(x^l, u^l) \quad (8)$$

and

$$y^l = f(x^l, u^l) \; . \quad (9)$$

The problem addressed in this report is to compute from the sample $\mathcal{F}_n$, an open-loop sequence of actions $\tilde{\mathbf{u}}^*_{\mathcal{F}_\mathbf{n}}(x_0)$ such that $\tilde{J}^{\tilde{\mathbf{u}}^*_{\mathcal{F}_\mathbf{n}}(\mathbf{x_0})}_{\mathcal{F}_n}(x_0)$ is as close as possible to $\tilde{J}^*_{\mathcal{F}_n}(x_0)$.

2 Model learning–type RL

Model learning–type reinforcement learning aims at solving optimal control problems by approximating the unknown functions f and ρ and solving the so approximated optimal control problem instead of the unknown actual optimal control problem. The values y^l (resp. r^l) of the function f (resp. ρ) in the state-action points (x^l, u^l) $\quad l = 1 \ldots n$ are used to learn a function $\tilde{f}_{\mathcal{F}_n}$ (resp. $\tilde{\rho}_{\mathcal{F}_n}$) over the whole space $\mathcal{X} \times \mathcal{U}$. The approximated optimal control problem defined by the functions $\tilde{f}_{\mathcal{F}_n}$ and $\tilde{\rho}_{\mathcal{F}_n}$ is solved and its solution is kept as an approximation of the solution of the optimal control problem defined by the actual functions f and ρ.

Given a sequence of actions $\mathbf{u} \in \mathcal{U}^T$ and a model learning–type reinforcement learning algorithm, we denote by $\tilde{J}^{\mathbf{u}}_{\mathcal{F}_n}(x_0)$ the approximated $T-$stage return of the sequence of actions $\mathbf{u}$, i.e. the $T-$stage return when considering the approximations $\tilde{f}_{\mathcal{F}_n}$ and $\tilde{\rho}_{\mathcal{F}_n}$:

Definition 2.1 (Approximated $T-$stage return)
$\forall \mathbf{u} \in \mathcal{U}^T, \forall x_0 \in \mathcal{X}$

$$\tilde{J}^{\mathbf{u}}_{\mathcal{F}_n}(x_0) = \sum_{t=0}^{T-1} \tilde{\rho}_{\mathcal{F}_n}\left(\tilde{x}_t, u_t\right) \quad (10)$$

with

$$\tilde{x}_{t+1} = \tilde{f}_{\mathcal{F}_n}(\tilde{x}_t, u_t), \; \forall t \in \{0, \ldots, T-1\} \tag{11}$$

and $\tilde{x}_0 = x_0$.

We denote by $\tilde{J}^*_{\mathcal{F}_n}(x_0)$ the maximal approximated $T-$stage return when starting from the initial state $x_0 \in \mathcal{X}$ according to the approximations $\tilde{f}_{\mathcal{F}_n}$ and $\tilde{\rho}_{\mathcal{F}_n}$:

Definition 2.2 (Maximal approximated $T-$stage return)
$\forall x_0 \in \mathcal{X}$,

$$\tilde{J}^*_{\mathcal{F}_n}(x_0) = \max_{\mathbf{u} \in \mathcal{U}^T} \tilde{J}^{\mathbf{u}}_{\mathcal{F}_n}(x_0) \; . \tag{12}$$

Using these notations, model learning–type RL algorithms aim at computing a sequence of actions $\tilde{\mathbf{u}}^*_{\mathcal{F}_\mathbf{n}}(x_0) \in \mathcal{U}^T$ such that $\tilde{J}^{\tilde{\mathbf{u}}^*_{\mathcal{F}_\mathbf{n}}(x_0)}_{\mathcal{F}_n}(x_0)$ is as close as possible (and ideally equal to) to $\tilde{J}^*_{\mathcal{F}_n}(x_0)$. These techniques implicitly assume that an optimal policy for the learned model also leads to high returns on the real problem.

3 The Voronoi Reinforcement Learning algorithm

This algorithm approximates the reward function ρ and the system dynamics f using piecewise constant approximations on a Voronoi–like [1] partition of the state-action space (which is equivalent to a nearest-neighbour approximation) and will be referred to by the VRL algorithm. Given an initial state $x_0 \in \mathcal{X}$, the VRL algorithm computes an open-loop sequence of actions which corresponds to an "optimal navigation" among the Voronoi cells.

Before fully describing this algorithm, we first assume that all the state-action pairs $\left\{(x^l, u^l)\right\}_{l=1}^n$ given by the sample of transitions $\mathcal{F}_n$ are unique, i.e.

$$\forall l, l' \in \{1, \ldots, n\}, (x^l, u^l) = (x^{l'}, u^{l'}) \implies l = l' \; . \tag{13}$$

We also assume that each action of the action space $\mathcal{U}$ has been tried at least once, i.e.,

$$\forall u \in \mathcal{U}, \exists l \in \{1, \ldots, n\}, u^l = u \; . \tag{14}$$

The model is based on the creation of n Voronoi cells $\left\{V^l\right\}_{l=1}^n$ which define a partition of size n of the state-action space. The Voronoi cell V^l associated to the element (x^l, u^l) of $\mathcal{F}_n$ is defined as the set of state-action pairs $(x, u) \in \mathcal{X} \times \mathcal{U}$ that satisfy:

$$(i) \quad u = u^l \;, \tag{15}$$

$$(ii) \quad l \in \arg\min_{l' : u^{l'} = u} \left\{ \|x - x^{l'}\|_{\mathcal{X}} \right\} \;, \tag{16}$$

$$(iii) \quad l = \min_{l'} \left\{ l' \in \arg\min_{l' : u^{l'} = u} \left\{ \|x - x^{l'}\|_{\mathcal{X}} \right\} \right\} . \tag{17}$$

One can verify that $\left\{V^l\right\}_{l=1}^n$ is indeed a partition of the state-action space $\mathcal{X} \times \mathcal{U}$ since every state-action $(x, u) \in \mathcal{X} \times \mathcal{U}$ belongs to one and only one Voronoi cell.

The function f (resp. ρ) is approximated by a piecewise constant function $\tilde{f}_{\mathcal{F}_n}$ (resp. $\tilde{\rho}_{\mathcal{F}_n}$) defined as follows:

$$\forall l \in \{1, \ldots, n\}, \forall (x, u) \in V^l, \quad \tilde{f}_{\mathcal{F}_n}(x, u) = y^l, \tag{18}$$

$$\tilde{\rho}_{\mathcal{F}_n}(x, u) = r^l \; . \tag{19}$$

3.1 Open-loop formulation

Using the approximations $\tilde{f}_{\mathcal{F}_n}$ and $\tilde{\rho}_{\mathcal{F}_n}$, we define a sequence of approximated optimal state-action value functions $\left(\tilde{Q}^*_{T-t}\right)_{t=0}^{T-1}$ as follows :

Definition 3.1 (Approximated optimal state-action value functions)
$\forall t \in \{0, \ldots, T-1\}, \forall (x, u) \in \mathcal{X} \times \mathcal{U}$,

$$\begin{aligned} \tilde{Q}^*_{T-t}(x,u) &= \tilde{\rho}_{\mathcal{F}_n}(x,u) \\ &+ \arg\max_{u' \in \mathcal{U}} \tilde{Q}^*_{T-t-1}\left(\tilde{f}_{\mathcal{F}_n}(x,u), u'\right) , \end{aligned} \tag{20}$$

with

$$Q^*_1(x,u) = \tilde{\rho}_{\mathcal{F}_n}(x,u), \quad \forall (x,u) \in \mathcal{X} \times \mathcal{U}. \tag{21}$$

Using the sequence of approximated optimal state-action value functions $\left(\tilde{Q}^*_{T-t}\right)_{t=0}^{T-1}$, one can infer an open-loop sequence of actions

$$\tilde{\mathbf{u}}^*_{\mathcal{F}_\mathbf{n}}(x_0) = (\tilde{u}^*_{\mathcal{F}_n,0}(x_0), \ldots, \tilde{u}^*_{\mathcal{F}_n,T-1}(x_0)) \in \mathcal{U}^T \tag{22}$$

which is an exact solution of the approximated optimal control problem, i.e. which is such that

$$\tilde{J}^{\tilde{\mathbf{u}}^*_{\mathcal{F}_\mathbf{n}}(\mathbf{x_0})}_{\mathcal{F}_n}(x_0) = \tilde{J}^*_{\mathcal{F}_n}(x_0) \tag{23}$$

as follows:

$$\tilde{u}^*_{\mathcal{F}_n,0}(x_0) \quad \in \quad \arg\max_{u' \in \mathcal{U}} \tilde{Q}^*_T(\tilde{x}^*_0, u') , \tag{24}$$

and, $\forall t \in \{0, \ldots, T-2\}$,

$$\tilde{u}^*_{\mathcal{F}_n,t+1}(x_0) \quad \in \quad \arg\max_{u' \in \mathcal{U}} \tilde{Q}^*_{T-(t+1)}\left(\tilde{f}_{\mathcal{F}_n}\left(\tilde{x}^*_t, \tilde{u}^*_{\mathcal{F}_n,t}(x_0)\right), u'\right) \tag{25}$$

where

$$\tilde{x}^*_{t+1} = \tilde{f}_{\mathcal{F}_n}(\tilde{x}^*_t, \tilde{u}^*_{\mathcal{F}_n,t}(x_0)), \forall t \in \{0, \ldots, T-1\}. \tag{26}$$

and $\tilde{x}^*_0 = x_0$.

All the approximated optimal state-action value functions $\left(\tilde{Q}^*_{T-t}\right)_{t=0}^{T-1}$ are piecewise constant over each Voronoi cell, a property that can be exploited for computing them easily as it is shown in Figure 1. The VRL algorithm has linear complexity with respect to the cardinality n of the sample of system transitions $\mathcal{F}_n$, the optimization horizon T and the cardinality m of the action space $\mathcal{U}$.

3.2 Closed-loop formulation

Using the sequence of approximated optimal state-action value functions $\left(\tilde{Q}^*_{T-t}\right)_{t=0}^{T-1}$, one can infer a closed-loop sequence of actions

$$\tilde{\mathbf{v}}^*_{\mathcal{F}_\mathbf{n}}(x_0) = (\tilde{v}^*_{\mathcal{F}_n,0}(x_0), \ldots, \tilde{v}^*_{\mathcal{F}_n,T-1}(x_0)) \in \mathcal{U}^T \tag{27}$$

Algorithm 1 The Voronoi Reinforcement Learning (VRL) algorithm. $Q_{T-t,l}$ is the value taken by the function $\tilde{Q}^*_{T-t}$ in the Voronoi cell V^l.

Inputs: an initial state $x_0 \in \mathcal{X}$, a sample of transitions $\mathcal{F}_n = \left\{\left(x^l, u^l, r^l, y^l\right)\right\}_{l=1}^n$;
Output: a sequence of actions $\tilde{\mathbf{u}}^*_{\mathcal{F}_\mathbf{n}}(x_0)$ and $\tilde{J}^*_{\mathcal{F}_n}(x_0)$;
Initialization:
Create a $n \times m$ matrix V such that $V(i,j)$ contains the index of the Voronoi cell (VC) where $\left(\tilde{f}_{\mathcal{F}_n}(x^i, u^i), a^j\right)$ lies ;
for $i = 1$ **to** n **do**
  $Q_{1,i} \leftarrow r^i$;
end for
Algorithm:
for $t = T-2$ **to** 0 **do**
  for $i = 1$ **to** n **do**
    $l \leftarrow \underset{l' \in \{1,\dots,m\}}{\arg\max} \left\{Q_{T-t-1,V(i,l')}\right\}$;
    $Q_{T-t,i} \leftarrow r^i + Q_{T-t-1,V(i,l)}$;
  end for
end for
$l \leftarrow \underset{l' \in \{1,\dots,m\}}{\arg\max} Q_{T,i'}$ where i' denotes the index of the VC where $(x_0, a^{l'})$ lies ;
$l_0^* \leftarrow$ index of the VC where (x_0, a^l) lies ;
$\tilde{J}^*_{\mathcal{F}_n}(x_0) \leftarrow Q_{T,l_0^*}$;
$i \leftarrow l_0^*$;
$\tilde{u}^*_{\mathcal{F}_n,0}(x_0) \leftarrow u^{l_0^*}$;
for $t = 0$ **to** $T-2$ **do**
  $l^*_{t+1} \leftarrow \underset{l' \in \{1,\dots,m\}}{\arg\max} \left\{Q_{T-t-1,V(i,l')}\right\}$;
  $\tilde{u}^*_{\mathcal{F}_n,t+1}(x_0) \leftarrow a^{l^*_{t+1}}$;
  $i \leftarrow V(i, l^*_{t+1})$;
end for
Return: $\tilde{\mathbf{u}}^*_{\mathcal{F}_\mathbf{n}}(x_0) = (\tilde{u}^*_{\mathcal{F}_n,0}(x_0), \dots, \tilde{u}^*_{\mathcal{F}_n,T-1}(x_0))$ and $\tilde{J}^*_{\mathcal{F}_n}(x_0)$.

by replacing the approximated system dynamics $\tilde{f}_{\mathcal{F}_n}$ with the true system dynamics in Equations (24), (25) and (26) as follows:

$$\tilde{v}^*_{\mathcal{F}_n,0}(x_0) \quad = \quad \arg\max_{v' \in \mathcal{U}} \tilde{Q}^*_T(\tilde{x}^*_0, v') \ ,$$

and, $\forall t \in \{0, \dots, T-2\}$,

$$\tilde{v}^*_{\mathcal{F}_n,t+1}(x_0) \quad = \quad \arg\max_{v' \in \mathcal{U}} \tilde{Q}^*_{T-(t+1)} \left(f \left(\tilde{x}^*_t, \tilde{v}^*_{\mathcal{F}_n,t}(x_0) \right), v' \right)$$

where

$$\tilde{x}^*_{t+1} = f(\tilde{x}^*_t, \tilde{v}^*_{t,\mathcal{F}_n}(x_0)), \forall t \in \{0, \dots, T-1\}. \tag{28}$$

and $\tilde{x}^*_0 = x_0$.

4 Theoretical analysis of the VRL algorithm

We propose to analyze the convergence of the Voronoi RL algorithm when the functions f and ρ are Lipschitz continuous and the sparsity of the sample of transitions decreases towards zero. We first assume the Lipschitz continuity of the functions f and ρ :

Assumption 4.1 (Lipschitz continuity of f and ρ)

$$\begin{aligned} &\exists L_f, L_\rho > 0 : \forall u \in \mathcal{U}, \forall x, x' \in \mathcal{X}, \\ &\qquad \|f(x,u) - f(x',u)\|_{\mathcal{X}} \quad \leq \quad L_f \|x - x'\|_{\mathcal{X}} \ , & (29) \\ &\qquad |\rho(x,u) - \rho(x',u)| \quad \leq \quad L_\rho \|x - x'\|_{\mathcal{X}} \ . & (30) \end{aligned}$$

For each action $u \in \mathcal{U}$, we denote by f_u (resp. ρ_u) the restrictions of the function f (resp. ρ) to the action u:

$$\begin{aligned} \forall u \in \mathcal{U}, \forall x \in \mathcal{X}, f_u(x) \quad &= \quad f(x,u) \ , & (31) \\ \rho_u(x) \quad &= \quad \rho(x,u) \ . & (32) \end{aligned}$$

All the functions $\{f_u\}_{u \in \mathcal{U}}$ and $\{\rho_u\}_{u \in \mathcal{U}}$ are thus also Lipschitz continuous. Given a sample of system transitions $\mathcal{F}_n$, and given an action $u \in \mathcal{U}$, we also introduce the restrictions of the function $\tilde{f}_{\mathcal{F}_n,u}$ and $\tilde{\rho}_{\mathcal{F}_n,u}$ as follows:

$$\begin{aligned} \forall u \in \mathcal{U}, \forall x \in \mathcal{X}, \tilde{f}_{\mathcal{F}_n,u}(x) \quad &= \quad \tilde{f}_{\mathcal{F}_n}(x,u) \ , & (33) \\ \tilde{\rho}_{\mathcal{F}_n,u}(x) \quad &= \quad \tilde{\rho}_{\mathcal{F}_n}(x,u) \ . & (34) \end{aligned}$$

Given a Voronoi cell $V^l \quad l \in \{1, \dots, n\}$, we denote by $\Delta^l_{\mathcal{F}_n}$ the radius of the Voronoi–like cell V^l defined as follows :

Definition 4.2 (Radius of Voronoi cells)
$\forall l \in \{1, \dots, n\}$,

$$\Delta^l_{\mathcal{F}_n} = \sup_{(x,u^l) \in V^l} \left\| x - x^l \right\|_{\mathcal{X}} \ . \tag{35}$$

We then introduce the sparsity of the sample of transitions $\mathcal{F}_n$, denoted by $\alpha_{\mathcal{F}_n}$:

Definition 4.3 (Sparsity of $\mathcal{F}_n$)

$$\alpha_{\mathcal{F}_n} = \max_{l \in \{1,\ldots,n\}} \quad \Delta^l_{\mathcal{F}_n} . \tag{36}$$

The sparsity of the sample of system transitions $\mathcal{F}_n$ can be seen, in a sense, as the "maximal radius" of all Voronoi cells. We suppose that a sequence of sample of transitions $(\mathcal{F}_n)_{n=n_0}^{\infty}$ (with $n_0 \geq m$) is known, and we assume that the corresponding sequence of sparsities $(\alpha_{\mathcal{F}_n})_{n=n_0}^{\infty}$ converges towards zero.

4.1 Consistency of the open-loop VRL algorithm

To each sample of transitions $\mathcal{F}_n$ are associated two piecewise constant approximated functions $\tilde{f}_{\mathcal{F}_n}$ and $\tilde{\rho}_{\mathcal{F}_n}$, and a sequence of actions $\tilde{\mathbf{u}}^*_{\mathcal{F}_\mathbf{n}}(x_0)$ computed using the VRL algorithm which is a solution of the approximated optimal control problem defined by the functions $\tilde{f}_{\mathcal{F}_n}$ and $\tilde{\rho}_{\mathcal{F}_n}$. We have the following theorem:

Theorem 4.4 (Consistency of the Voronoi RL algorithm)
$\forall x_0 \in \mathcal{X}$,

$$\lim_{n\to\infty} J^{\tilde{\mathbf{u}}^*_{\mathcal{F}_\mathbf{n}}(x_0)}(x_0) = J^*(x_0) . \tag{37}$$

Before giving the proof of Theorem 4.4, let us first introduce a few lemmas.

Lemma 4.5 (Uniform convergence of $\tilde{f}_{\mathcal{F}_n,u}$ and $\tilde{\rho}_{\mathcal{F}_n,u}$ towards f_u and ρ_u)

$$\forall u \in \mathcal{U}, \qquad \lim_{n\to\infty} \sup_{x\in\mathcal{X}} \left\| f_u(x) - \tilde{f}_{\mathcal{F}_n,u}(x) \right\|_{\mathcal{X}} = 0 , \tag{38}$$

$$\lim_{n\to\infty} \sup_{x\in\mathcal{X}} |\rho_u(x) - \tilde{\rho}_{\mathcal{F}_n,u}(x)| = 0 . \tag{39}$$

Proof. Let $u \in \mathcal{U}$, let $x \in \mathcal{X}$, and let V^l be the Voronoi cell where (x, u) lies (then, $u = u^l$). One has

$$\tilde{f}_{\mathcal{F}_n,u}(x) = y^l , \tag{40}$$

$$\tilde{\rho}_{\mathcal{F}_n,u}(x) = r^l . \tag{41}$$

which implies that

$$\left\| \tilde{f}_{\mathcal{F}_n,u}(x) - f_u(x^l) \right\|_{\mathcal{X}} = 0 , \tag{42}$$

$$\left| \tilde{\rho}_{\mathcal{F}_n,u}(x) - \rho_u(x^l) \right| = 0 . \tag{43}$$

Then,

$$\begin{aligned} \left\| f_u(x) - \tilde{f}_{\mathcal{F}_n,u}(x) \right\|_{\mathcal{X}} &\leq \left\| f_u(x) - f_u(x^l) \right\|_{\mathcal{X}} \\ &+ \left\| f_u(x^l) - \tilde{f}_{\mathcal{F}_n,u}(x) \right\|_{\mathcal{X}} & (44) \\ &\leq L_f \left\| x - x^l \right\|_{\mathcal{X}} + 0 & (45) \\ &\leq L_f \Delta^l_{\mathcal{F}_n} & (46) \\ &\leq L_f \alpha_{\mathcal{F}_n} , & (47) \end{aligned}$$

and similarly for the functions ρ_u and $\tilde{\rho}_{\mathcal{F}_n,u}$,

$$|\rho_u(x) - \tilde{\rho}_{\mathcal{F}_n,u}(x)| \leq L_\rho \alpha_{\mathcal{F}_n} . \tag{48}$$

This ends the proof since $\alpha_{\mathcal{F}_n} \to 0$. ■

Lemma 4.6 (Uniform convergence of the sum of functions)
Let $(h_n : \mathcal{X} \to \mathbb{R})_{n\in\mathbb{N}}$ (resp. $(h'_n : \mathcal{X} \to \mathbb{R})_{n\in\mathbb{N}}$) be a sequence of functions that uniformly converges towards $h : \mathcal{X} \to \mathbb{R}$ (resp. $h' : \mathcal{X} \to \mathbb{R}$). Then, the sequence of functions $((h_n + h'_n) : \mathcal{X} \to \mathbb{R})_{n\in\mathbb{N}}$ uniformly converges towards the function $(h+h')$.

Proof. Let $\epsilon > 0$. Since $(h_n)_{n\in\mathbb{N}}$ uniformly converges towards h, there exists $n_h \in \mathbb{N}$ such that

$$\forall n \geq n_h, \forall x \in \mathcal{X}, |h_n(x) - h(x)| \leq \frac{\epsilon}{2} \,. \tag{49}$$

Since $(h'_n)_{n\in\mathbb{N}}$ uniformly converges towards h', there exists $n_{h'} \in \mathbb{N}$ such that

$$\forall n \geq n_{h'}, \forall x \in \mathcal{X}, |h'_n(x) - h'(x)| \leq \frac{\epsilon}{2} \,. \tag{50}$$

We denote by $n_{\max} = \max(n_h, n_{h'})$. One has
$\forall n \geq n_{\max}, \forall x \in \mathcal{X}$,

$$\begin{aligned} |(h_n(x) - h'_n(x)) - (h(x) + h'(x))| &\leq |h_n(x) - h(x)| + |h'_n(x) - h'(x)| && (51)\\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} && (52)\\ &\leq \epsilon \,, && (53) \end{aligned}$$

which ends the proof. ■

Lemma 4.7 (Uniform convergence of composed functions)

- *Let $(g_n : \mathcal{X} \to \mathcal{X})_{n\in\mathbb{N}}$ be a sequence of functions that uniformly converges towards $g : \mathcal{X} \to \mathcal{X}$;*
- *Let $(g'_n : \mathcal{X} \to \mathcal{X})_{n\in\mathbb{N}}$ be a sequence of functions that uniformly converges towards $g' : \mathcal{X} \to \mathcal{X}$. Let us assume that g' is $L_{g'}-$Lipschitzian;*
- *Let $(h_n : \mathcal{X} \to \mathbb{R})_{n\in\mathbb{N}}$ be a sequence of functions that uniformly converges towards $h : \mathcal{X} \to \mathbb{R}$. Let us assume that h is L_h-Lipschitzian.*

Then,

- *The sequence of functions $(g'_n \circ g_n)_{n\in\mathbb{N}}$ uniformly converges towards the function $g' \circ g$.*
- *The sequence of functions $(h_n \circ g_n)_{n\in\mathbb{N}}$ uniformly converges towards the function $h \circ g$,*

where the notation $h_n \circ g_n$ (resp. $g'_n \circ g$, $h \circ g$ and $g' \circ g$) denotes the mapping $x \to h_n(g_n(x))$ (resp. $x \to g'_n(g_n(x))$, $x \to h(g(x))$ and $x \to g'(g(x))$).

Proof. Let us prove the second bullet. Let $\epsilon > 0$. Since $(g_n)_{n\in\mathbb{N}}$ uniformly converges towards g, there exists $n_g \in \mathbb{N}$ such that

$$\forall n \geq n_g, \forall x \in \mathcal{X}, \|g_n(x) - g(x)\|_{\mathcal{X}} \leq \frac{\epsilon}{2L_h} \,. \tag{54}$$

Since $(h_n)_{n\in\mathbb{N}}$ uniformly converges towards h, there exists $n_h \in \mathbb{N}$ such that

$$\forall n \geq n_h, \forall x \in \mathcal{X}, |h_n(x) - h(x)| \leq \frac{\epsilon}{2} \, . \tag{55}$$

We denote by $n_{h\circ g} = \max(n_h, n_g)$. One has
$\forall n \geq n_{h\circ g}, \forall x \in \mathcal{X}$,

$$\begin{aligned} |h_n(g_n(x)) - h(g(x))| &\leq |h_n(g_n(x)) - h(g_n(x))| + |h(g_n(x)) - h(g(x))| && (56) \\ &\leq \frac{\epsilon}{2} + L_h \|g_n(x) - g(x)\|_{\mathcal{X}} && (57) \\ &\leq \frac{\epsilon}{2} + L_h \frac{\epsilon}{2L_h} && (58) \\ &\leq \epsilon, && (59) \end{aligned}$$

which proves that the sequence of functions $(h_n \circ g_n)_n$ uniformly converges towards $h \circ g$. ■

Lemma 4.8 (Convergence of $\tilde{J}^{\mathbf{u}}_{\mathcal{F}_n}(x_0)$ towards $J^{\mathbf{u}}(x_0)$,$\forall \mathbf{u} \in \mathcal{U}^T$)
$\forall \mathbf{u} \in \mathcal{U}^T, \forall x_0 \in \mathcal{X}$,

$$\lim_{n\to\infty} \left| \tilde{J}^{\mathbf{u}}_{\mathcal{F}_n}(x_0) - J^{\mathbf{u}}(x_0) \right| = 0 \, . \tag{60}$$

Proof. Let $\mathbf{u} \in \mathcal{U}^T$ be a fixed sequence of actions. For all $n \in \mathbb{N}, n \geq n_0$ the function $\tilde{J}^{\mathbf{u}}_{\mathcal{F}_n} : \mathcal{X} \to \mathbb{R}$ can be written as follows :

$$\begin{aligned} \tilde{J}^{\mathbf{u}}_{\mathcal{F}_n} &= \tilde{\rho}_{\mathcal{F}_n,u_0} + \tilde{\rho}_{\mathcal{F}_n,u_1} \circ \tilde{f}_{\mathcal{F}_n,u_0} \\ &+ \ \dots \\ &+ \ \tilde{\rho}_{\mathcal{F}_n,T-1} \circ \tilde{f}_{\mathcal{F}_n,u_{T-2}} \circ \dots \circ \tilde{f}_{\mathcal{F}_n,u_0} \, . \end{aligned} \tag{61}$$

Since all the functions $\{\tilde{\rho}_{\mathcal{F}_n,u_t}\}_{0\leq t\leq T-1}$ and $\left\{\tilde{f}_{\mathcal{F}_n,u_t}\right\}_{0\leq t\leq T-1}$ uniformly converge towards the functions $\{f_{u_t}\}_{0\leq t\leq T-1}$ and $\{\rho_{u_t}\}_{0\leq t\leq T-1}$, respectively, and since all the functions $\{f_{u_t}\}_{0\leq t\leq T-1}$ and $\{\rho_{u_t}\}_{0\leq t\leq T-1}$ are Lipschitz continuous, Lemma 4.6 and Lemma 4.7 ensure that the function $x_0 \to \tilde{J}^{\mathbf{u}}_{\mathcal{F}_n}(x_0)$ uniformly converges to the function $x_0 \to J^{\mathbf{u}}(x_0)$. This implies the convergence of the sequence $\left(\tilde{J}^{\mathbf{u}}_{\mathcal{F}_n}(x_0)\right)_{n\in\mathbb{N}}$ towards $J^{\mathbf{u}}(x_0)$, for any sequence of actions $\mathbf{u} \in \mathcal{U}^T$, and for any initial state $x_0 \in \mathcal{X}$. ■

Proof of Theorem 4.4. Let us proof Equation 37. Let $\mathbf{u}^*(x_0)$ be an optimal sequence of actions, and $\left(\tilde{\mathbf{u}}^*_{\mathcal{F}_\mathbf{n}}(x_0)\right)_{n\in\mathbb{N}}$ be a sequence of sequence of actions computed by the Voronoi RL algorithm. Each sequence of actions $\tilde{\mathbf{u}}^*_{\mathcal{F}_\mathbf{n}}(x_0)$ is optimal with respect to the approximated model defined by the approximated functions $\tilde{f}_{\mathcal{F}_n}$ and $\tilde{\rho}_{\mathcal{F}_n}$. One then has

$$\forall n \geq m, \forall \mathbf{u} \in \mathcal{U}^T, \tilde{J}^{\tilde{\mathbf{u}}^*_{\mathcal{F}_\mathbf{n}}(x_0)}_{\mathcal{F}_n}(x_0) \geq \tilde{J}^{\mathbf{u}}_{\mathcal{F}_n}(x_0) \, . \tag{62}$$

The previous inequality is also valid for the sequence of actions $\mathbf{u}^*(x_0)$:

$$\forall n \geq m, \tilde{J}^{\tilde{\mathbf{u}}^*_{\mathcal{F}_\mathbf{n}}(x_0)}_{\mathcal{F}_n}(x_0) \geq \tilde{J}^{\mathbf{u}^*(x_0)}_{\mathcal{F}_n}(x_0) \, . \tag{63}$$

Then, $\forall n \geq m$,

$$\tilde{J}_{\mathcal{F}_n}^{\tilde{\mathbf{u}}^*_{\mathcal{F}_\mathbf{n}}(\mathbf{x_0})}(x_0) - J^{\tilde{\mathbf{u}}^*_{\mathcal{F}_\mathbf{n}}(\mathbf{x_0})}(x_0) + J^{\tilde{\mathbf{u}}^*_{\mathcal{F}_\mathbf{n}}(\mathbf{x_0})}(x_0) \geq \tilde{J}_{\mathcal{F}_n}^{\mathbf{u}^*(\mathbf{x_0})}(x_0) - J^{\mathbf{u}^*(\mathbf{x_0})}(x_0) + J^{\mathbf{u}^*(\mathbf{x_0})}(x_0) \; . \tag{64}$$

According to Lemma 4.8, one can write

$$\lim_{n\to\infty} \tilde{J}_{\mathcal{F}_n}^{\tilde{\mathbf{u}}^*_{\mathcal{F}_\mathbf{n}}(\mathbf{x_0})}(x_0) - J^{\tilde{\mathbf{u}}^*_{\mathcal{F}_\mathbf{n}}(\mathbf{x_0})}(x_0) = 0 \; , \tag{65}$$

$$\lim_{n\to\infty} \tilde{J}_{\mathcal{F}_n}^{\mathbf{u}^*(\mathbf{x_0})}(x_0) - J^{\mathbf{u}^*(\mathbf{x_0})}(x_0) = 0 \; . \tag{66}$$

which leads to

$$\lim_{n\to\infty} J^{\tilde{\mathbf{u}}^*_{\mathcal{F}_\mathbf{n}}(\mathbf{x_0})}(x_0) \geq \lim_{n\to\infty} J^{\mathbf{u}^*(\mathbf{x_0})}(x_0) = J^*(x_0) \; . \tag{67}$$

On the other hand, since $\mathbf{u}^*(\mathbf{x_0})$ is an optimal sequence of actions, one has

$$\forall n \in \mathbb{N}_0, J^{\tilde{\mathbf{u}}^*_{\mathcal{F}_\mathbf{n}}(\mathbf{x_0})}(x_0) \leq J^{\mathbf{u}^*(x_0)}(x_0) = J^*(x_0) \; , \tag{68}$$

which leads to

$$\lim_{n\to\infty} J^{\tilde{\mathbf{u}}^*_{\mathcal{F}_\mathbf{n}}(\mathbf{x_0})}(x_0) \leq J^*(x_0) \; . \tag{69}$$

Equations 67 and 69 allow to conclude the proof:

$$\lim_{n\to\infty} J^{\tilde{\mathbf{u}}^*_{\mathcal{F}_\mathbf{n}}(\mathbf{x_0})}(x_0) = J^*(x_0) \; . \tag{70}$$

■

Acknowledgements

Raphael Fonteneau acknowledges the financial support of the FRIA (Fund for Research in Industry and Agriculture). Damien Ernst is a research associate of the FRS-FNRS. This report presents research results of the Belgian Network BIOMAGNET (Bioinformatics and Modeling: from Genomes to Networks), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office.We also acknowledge financial support from NIH grants P50 DA10075 and R01 MH080015. The scientific responsibility rests with its authors.

References

[1] F. Aurenhammer. Voronoi diagrams – a survey of a fundamental geometric data structure. *ACM Computing Surveys (CSUR)*, 23(3):345–405, 1991.