Inferring bounds on the performance of a control policy from a sample of trajectories

Raphael Fonteneau, Susan Murphy, Louis Wehenkel and Damien Ernst

***Abstract*— We propose an approach for inferring bounds on the finite-horizon return of a control policy from an off-policy sample of trajectories collecting state transitions, rewards, and control actions. In this paper, the dynamics, control policy, and reward function are supposed to be deterministic and Lipschitz continuous. Under these assumptions, a polynomial algorithm, in terms of the sample size and length of the optimization horizon, is derived to compute these bounds, and their tightness is characterized in terms of the sample density.**

I. Introduction

In financial [6], medical [9] and engineering sciences [1], as well as in artificial intelligence [13], variants (or generalizations) of the following discrete-time optimal control problem arise quite frequently: a system, characterized by its state-transition function $x_{t+1} = f(x_t, u_t)$, should be controlled by using a policy $u_t = h(t, x_t)$ so as to maximize a cumulated reward $\sum_{t=0}^{T-1} \rho(x_t, u_t)$ over a finite horizon T.

Among the solution approaches that have been proposed for this class of problems we have, on the one hand, dynamic programming [1] and model predictive control [3] which compute optimal solutions from an analytical or computational model of the real system, and, on the other hand, reinforcement learning approaches [13], [8], [5], [11] which compute approximations of optimal control policies based only on data gathered from the real system. In between, we have approximate dynamic programming approaches which use datasets generated by using a model (e.g. by Monte-Carlo simulation) so as to derive approximate solutions while complying with computational requirements [2].

Whatever the approach (model based, data based, Monte-Carlo based, (or even finger based)) used to derive a control policy for a given problem, one major question that remains open today is to ascertain the *actual* performance of the derived control policy [7], [12] when applied to the *real* system behind the model or the dataset (or the finger). Indeed, for many applications, even if it is perhaps not paramount to have a policy h which is very close to the optimal one, it is however crucial to be able to guarantee that the considered policy h leads for some initial states x_0 to high-enough cumulated rewards on the real system that is considered.

In this paper, we thus focus on the evaluation of control policies on the sole basis of the actual behaviour of the concerned real system. We use to this end a sample of trajectories $(x_0, u_0, r_0, x_1, \ldots, r_{T-1}, x_T)$ gathered from interactions with the real system, where states $x_t \in X$, actions $u_t \in U$ and instantaneous rewards $r_t = \rho(x_t, u_t) \in \mathbb{R}$ at successive discrete instants $t = 0, 1, \ldots, T-1$ will be exploited so as to evaluate bounds on the performance of a given control policy $h(t, x) : \{0, 1, \ldots, T-1\} \times X \to U$ when applied to a given initial state x_0 of the real system.

Actually, our proposed approach does not require full-length trajectories since it relies only on a set of one-step system transitions $\mathcal{F} = \{(x^l, u^l, r^l, y^l)\}_{l=1}^{|\mathcal{F}|}$, each one providing the knowledge of a sample of information (x, u, r, y), named four-tuple, where y is the state reached after taking action u in state x and r the instantaneous reward associated with the transition. We however assume that the state and action spaces are normed and that the system dynamics ($y = f(x, u)$) and the reward function ($r = \rho(x, u)$) and control policy ($u = h(t, x)$) are deterministic and Lipschitz continuous.

In a few words, the approach works by identifying in $\mathcal{F}$ a sequence of T four-tuples $[(x^{l_0}, u^{l_0}, r^{l_0}, y^{l_0}), (x^{l_1}, u^{l_1}, r^{l_1}, y^{l_1}), \ldots, (x^{l_{T-1}}, u^{l_{T-1}}, r^{l_{T-1}}, y^{l_{T-1}})]$ ($l_t \in \{1, \ldots, |\mathcal{F}|\}$), which maximizes a specific numerical criterion. This criterion is made of the sum of the T rewards corresponding to these four-tuples ($\sum_{t=0}^{T-1} r^{l_t}$) and T negative terms. The negative term corresponding to the four-tuple $(x^{l_t}, u^{l_t}, r^{l_t}, y^{l_t})$ of the sequence represents an upper bound variation of the cumulated rewards over the remaining time steps that can occur by simulating the system from a state x^{l_t} rather than $y^{l_{t-1}}$ (with $y^{l_{-1}} = x_0$) and by using at time t the action u^{l_t} rather than $h(t, y^{l_{t-1}})$. We provide a polynomial algorithm to compute this optimal sequence of tuples and derive a tightness characterization of the corresponding performance bound in terms of the density of the sample $\mathcal{F}$.

The rest of this paper is organized as follows. In Section II, we formalize the problem considered in this paper. In Section III, we show that the state-action value function of a policy over the N last steps of an episode is Lipschitz continuous. Section IV uses this result to compute from a sequence of four-tuples a lower bound on the cumulated reward obtained by a policy h when starting from a given $x_0 \in X$, while Section V proposes a polynomial algorithm for identifying the sequence of four-tuples which leads to the best bound. Section VI studies the tightness of this bound and shows that it can be characterized by $C\alpha^*$ where C is a positive constant and α^* is the maximum distance between any element of the state-action space $X \times U$ and its closest state-action pair $(x^l, u^l) \in \mathcal{F}$. Finally, Section VII concludes and outlines

Raphael Fonteneau, Louis Wehenkel and Damien Ernst are with the Department of Electrical Engineering and Computer Science of the University of Liège. Susan Murphy is with the University of Michigan. Emails: samurphy@umich.edu, {fonteneau, lwh, ernst}@montefiore.ulg.ac.be.

directions for future research.

II. Formulation of the problem

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

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

where for all t, the state x_t is an element of the state space X and the action u_t is an element of the action space U (both X and U are assumed to be normed vector spaces). $T \in \mathbb{N}_0$ is referred to as the *optimization horizon*.

The transition from t to $t+1$ is associated with an instantaneous reward $r_t = \rho(x_t, u_t) \in \mathbb{R}$.

For every initial state x_0 and for every sequence of actions the cumulated reward over T stages (also named return over T stages) is defined as

$$J_T^{(u_0, u_1, \ldots, u_{T-1})}(x_0) = \sum_{t=0}^{T-1} \rho(x_t, u_t). \tag{2}$$

We consider in this paper deterministic time-varying T-stage policies $h : \{0, 1, \ldots, T-1\} \times X \to U$ which select at time t the action u_t based on the current time and the current state ($u_t = h(t, x_t)$). The return over T stages of a policy h from a state x_0 is denoted by

$$J_T^h(x_0) = \sum_{t=0}^{T-1} \rho(x_t, h(t, x_t)). \tag{3}$$

We also assume that the dynamics f, the reward function ρ and the policy h are Lipschitz continuous, i.e., that there exist finite constants $L_f, L_\rho, L_h \in \mathbb{R}$ such that:

$$\begin{aligned} \|f(x,u) - f(x',u')\| &\leq L_f\big(\|x - x'\| + \|u - u'\|\big), &(4)\\ |\rho(x,u) - \rho(x',u')| &\leq L_\rho\big(\|x - x'\| + \|u - u'\|\big), &(5)\\ \|h(t,x) - h(t,x')\| &\leq L_h\|x - x'\|, &(6) \end{aligned}$$

$\forall x, x' \in X, \forall u, u' \in U, \forall t \in \{0, \ldots, T-1\}$. The smallest constants satisfying those inequalities are named the Lipschitz constants.

We further suppose that:

1) the system dynamics f and the reward function ρ are unknown,
2) an arbitrary set of one-step system transitions (also named four-tuples) $\mathcal{F} = \{(x^l, u^l, r^l, y^l)\}_{l=1}^{|\mathcal{F}|}$ is known. Each four-tuple is such that $y^l = f(x^l, u^l)$ and $r^l = \rho(x^l, u^l)$,
3) three constants L_f, L_ρ, L_h satisfying the above-written inequalities are known.[1]

Under these assumptions, we want to find for an arbitrary initial state x_0 of the system a *lower bound* on the return over T stages of any given policy h.

[1]These constants do not necessarily have to be the smallest ones satisfying these inequalities (i.e., the Lispchitz constants), even if, the smaller they are, the tighter the bound will be.

III. Lipschitz continuity of the state-action value function

For $N = 1, \ldots, T$, let us define the family of functions $Q_N^h : X \times U \to \mathbb{R}$ as follows:

$$Q_N^h(x,u) = \rho(x,u) + \sum_{t=T-N+1}^{T-1} \rho(x_t, h(t, x_t)), \tag{7}$$

where $x_{T-N+1} = f(x,u)$. $Q_N^h(x,u)$ gives the sum of rewards from instant $t = T-N$ to instant $T-1$ when (i) the system is in state x at instant $T-N$, (ii) the action chosen at instant $T-N$ is u and (iii) the actions are selected at subsequent instants according to the policy h ($u_t = h(t, x_t), \forall t > T-N$).

The function J_T^h can be deduced from Q_N^h as follows:

$$\forall x \in X,\ J_T^h(x) = Q_T^h(x, h(0, x)). \tag{8}$$

We also have $\forall x \in X, \forall u \in U$,

$$Q_{N+1}^h(x,u) = \rho(x,u) + Q_N^h(f(x,u), h(T-N, f(x,u)). \tag{9}$$

We prove hereafter the Lipschitz continuity of $Q_N^h, \forall N \in \{1, \ldots, T\}$.

Lemma 3.1 (Lipschitz continuity of Q_N^h): $\forall N \in \{1, \ldots, T\}$, there exists a finite constant $L_{Q_N} \in \mathbb{R}^+$ such that $\forall x, x' \in X, \forall u, u' \in U$,

$$|Q_N^h(x,u) - Q_N^h(x',u')| \leq L_{Q_N}\big(\|x - x'\| + \|u - u'\|\big). \tag{10}$$

Proof: We consider the statement $\mathcal{H}(N)$:
There exists a finite constant $L_{Q_N} \in \mathbb{R}^+$ such that $\forall x, x' \in X, \forall u, u' \in U$,

$$|Q_N^h(x,u) - Q_N^h(x',u')| \leq L_{Q_N}\big(\|x - x'\| + \|u - u'\|\big).$$

We prove by mathematical induction that $\mathcal{H}(N)$ is true $\forall N \in \{1, \ldots, T\}$. For the sake of clarity, we denote $|Q_N^h(x,u) - Q_N^h(x',u')|$ by Δ_N.

- Basis: $N = 1$

We have $\Delta_N = |\rho(x,u) - \rho(x',u')|$, and the Lipschitz continuity of ρ allows to write

$$\Delta_N \leq L_{Q_1}\big(\|x - x'\| + \|u - u'\|\big),$$

with $L_{Q_1} \doteq L_\rho$. This proves $\mathcal{H}(1)$.

- Induction step: we suppose that $\mathcal{H}(N)$ is true, $1 \leq N \leq T-1$.

Using Equation (9), we can write
$\Delta_{N+1} = \big|Q_{N+1}^h(x,u) - Q_{N+1}^h(x',u')\big| = \big|\rho(x,u) - \rho(x',u') + Q_N^h(f(x,u), h(T-N, f(x,u))) - Q_N^h(f(x',u'), h(T-N, f(x',u')))\big|$

and, from there,
$\Delta_{N+1} \leq \big|\rho(x,u) - \rho(x',u')\big| + \big|Q_N^h(f(x,u), h(T-N, f(x,u))) - Q_N^h(f(x',u'), h(T-N, f(x',u')))\big|$.

$\mathcal{H}(N)$ and the Lipschitz continuity of ρ give

$$\Delta_{N+1} \leq L_\rho\big(\|x-x'\|+\|u-u'\|\big) + L_{Q_N}\Big(\|f(x,u) - f(x',u')\| + \|h(T-N, f(x,u)) - h(T-N, f(x',u'))\|\Big).$$

Using the Lipschitz continuity of f and h, we have

$$\Delta_{N+1} \leq L_\rho\big(\|x-x'\|+\|u-u'\|\big) + L_{Q_N}\Big(L_f\big(\|x-x'\|+\|u-u'\|\big) + L_h L_f\big(\|x-x'\|+\|u-u'\|\big)\Big),$$

and, from there,

$$\Delta_{N+1} \leq L_{Q_{N+1}}\big(\|x-x'\|+\|u-u'\|\big),$$

with $L_{Q_{N+1}} \doteq L_\rho + L_{Q_N} L_f(1+L_h)$. This proves that $\mathcal{H}\ (N+1)$ is true, and ends the proof.

■

Let $L^*_{Q_N}$ be the Lipschitz constant of the function Q^h_N, that is the smallest value of L_{Q_N} that satisfies inequality (10). We have the following result:

Lemma 3.2 (Upper bound on $L^*_{Q_N}$):

$$L^*_{Q_N} \leq L_\rho\Big(\sum_{t=0}^{N-1}[L_f(1+L_h)]^t\Big) \tag{11}$$

Proof: A sequence of positive constants $L_{Q_1}, \ldots, L_{Q_N}$ is defined in the proof of Lemma 3.1. Each constant L_{Q_N} of this sequence is an upper-bound on the Lipschitz constant related to the function Q^h_N. These L_{Q_N} constants satisfy the relationship

$$L_{Q_{N+1}} = L_\rho + L_{Q_N} L_f(1+L_h) \tag{12}$$

(with $L_{Q_1} = L_\rho$) from which the lemma can be proved in a straightforward way.

■

The value of the constant L_{Q_N} will influence the lower bound on the return of the policy h that will be established later in this paper. The larger this constant, the looser the bounds. When using these bounds, L_{Q_N} should therefore preferably be chosen as small as possible while still ensuring that inequality (10) is satisfied. Later in this paper, we will use the upper bound (11) to select a value for L_{Q_N}. More specifically, we will choose

$$L_{Q_N} = L_\rho\Big(\sum_{t=0}^{N-1}[L_f(1+L_h)]^t\Big). \tag{13}$$

IV. Computing a lower bound on $J^h_T(x_0)$ from a sequence of four-tuples

The algorithm described in Table I provides a way of computing from any T-length sequence of four-tuples $\tau = [(x^{l_t}, u^{l_t}, r^{l_t}, y^{l_t})]_{t=0}^{T-1}$ a lower bound on $J^h_T(x_0)$, provided that the initial state x_0, the policy h and three constants L_f, L_ρ and L_h satisfying inequalities (4-6) are given. The algorithm is a direct consequence of Theorem 4.1 below.

Inputs: An initial state x_0, a policy h, a sequence of four-tuples $\tau = [(x^{l_t}, u^{l_t}, r^{l_t}, y^{l_t})]_{t=0,\ldots,T-1}$, and three constants L_f, L_ρ, L_h which satisfy inequalities (4-6).
Output: A lower bound on $J^h_T(x_0)$.
Algorithm:
Set $lb = 0$
Set $y^{l-1} = x_0$
For $t = 0$ to $T-1$ **do**
Set $L_{Q_{T-t}} = L_\rho\Big(\sum_{k=0}^{T-t-1}[L_f(1+L_h)]^k\Big)$
Set $lb = lb + r^{l_t} - L_{Q_{T-t}}\big(\|x^{l_t} - y^{l_{t-1}}\| + \|u^{l_t} - h(t, y^{l_{t-1}})\|\big)$
end for
Return lb

TABLE I
AN ALGORITHM FOR COMPUTING FROM A SEQUENCE OF FOUR-TUPLES τ A LOWER BOUND ON $J^h_T(x_0)$.

The lower bound on $J^h_T(x_0)$ derived in Theorem 4.1 can be interpreted as follows. The sum of the rewards of the "broken" trajectory formed by the sequence of four-tuples τ can never be greater than $J^h_T(x_0)$, provided that every reward r^{l_t} is penalized by a factor $L_{Q_{T-t}}\big(\|x^{l_t} - y^{l_{t-1}}\| + \|u^{l_t} - h(t, y^{l_{t-1}})\|\big)$. This factor is in fact an upper bound on the variation of the function Q^h_{T-t} that can occur when "jumping" from $(y^{l_t}, h(t, y^{l_t}))$ to $(x^{l_{t+1}}, u^{l_{t+1}})$. An illustration of this interpretation is given in Figure 1.

Theorem 4.1 (Lower bound on $J^h_T(x_0)$): Let x_0 be an initial state of the system, h a policy, $\tau = [(x^{l_t}, u^{l_t}, r^{l_t}, y^{l_t})]_{t=0}^{T-1}$ a sequence of tuples. Then we have the following lower bound on $J^h_T(x_0)$

$$\sum_{t=0}^{T-1}(r^{l_t} - L_{Q_{T-t}}\delta_t) \leq J^h_T(x_0), \tag{14}$$

where

$$\delta_t = \|x^{l_t} - y^{l_{t-1}}\| + \|u^{l_t} - h(t, y^{l_{t-1}})\| \ \forall t \in \{0, 1, \ldots, T-1\},$$

with $y^{l-1} = x_0$.

Proof: Using Equation (8) and the Lipschitz continuity of Q^h_T, we can write

$$|Q^h_T(x_0, u_0) - Q^h_T(x^{l_0}, u^{l_0})| \leq L_{Q_T}\big(\|x_0 - x^{l_0}\| + \|u_0 - u^{l_0}\|\big),$$

and with $u_0 = h(0, x_0)$,

$$\begin{aligned} |J^h_T(x_0) - Q^h_T(x^{l_0}, u^{l_0})| = \\ |Q^h_T(x_0, h(0,x_0)) - Q^h_T(x^{l_0}, u^{l_0})| \leq \\ L_{Q_T}\big(\|x_0 - x^{l_0}\| + \|h(0,x_0) - u^{l_0}\|\big). \end{aligned}$$

It follows that

$$Q^h_T(x^{l_0}, u^{l_0}) - L_{Q_T}\delta_0 \leq J^h_T(x_0).$$

By definition of the state-action evaluation function Q^h_T, we have

$$Q^h_T(x^{l_0}, u^{l_0}) = \rho(x^{l_0}, u^{l_0}) + Q^h_{T-1}(f(x^{l_0}, u^{l_0}), h(1, f(x^{l_0}, u^{l_0})))$$

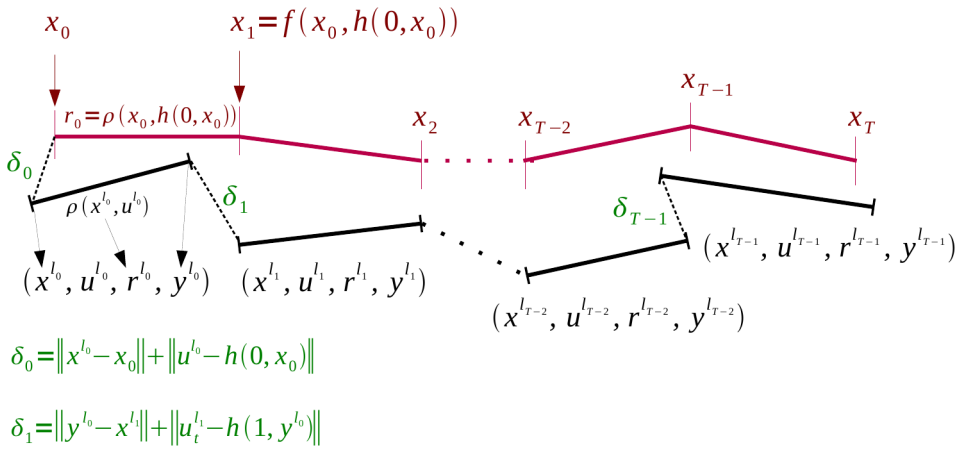

Fig. 1. A graphical interpretation of the different terms composing the bound on $J_T^h(x_0)$ inferred from a sequence of four-tuples (see Equation (14)). The bound is equal to the sum of all the rewards corresponding to this sequence of four-tuples (the terms r^{l_t} $t = 0, 1, \ldots, T-1$ on the figure) minus the sum of all the terms $L_{Q_{T-t}}\delta_t$.

and from there

$$Q_T^h(x^{l_0}, u^{l_0}) = r^{l_0} + Q_{T-1}^h(y^{l_0}, h(1, y^{l_0})).$$

Thus,

$$Q_{T-1}^h(y^{l_0}, h(1, y^{l_0})) + r^{l_0} - L_{Q_T}\delta_0 \leq J_T^h(x_0).$$

By using the Lipschitz property of the function Q_{T-1}^h, we can write
$|Q_{T-1}^h(y^{l_0}, h(1, y^{l_0})) - Q_{T-1}^h(x^{l_1}, u^{l_1})| \leq L_{Q_{T-1}}\big(\|y^{l_0} - x^{l_1}\| + \|h(1, y^{l_0}) - u^{l_1}\|\big) = L_{Q_{T-1}}\delta_1$,
which implies that $Q_{T-1}^h(x^{l_1}, u^{l_1}) - L_{Q_{T-1}}\delta_1 \leq Q_{T-1}^h(y^{l_0}, h(1, y^{l_0}))$. We have therefore

$$Q_{T-1}^h(x^{l_1}, u^{l_1}) + r^{l_0} - L_{Q_T}\delta_0 - L_{Q_{T-1}}\delta_1 \leq J_T^h(x_0).$$

By iterating this derivation, we obtain inequality (14) which completes the proof.

■

V. Finding the highest lower bound

Let

$$B^h(\tau, x_0) = \sum_{t=0}^{T-1}[r^{l_t} - L_{Q_{T-t}}\delta_t], \tag{15}$$

with $\delta_t = \|x^{l_t} - y^{l_{t-1}}\| + \|u^{l_t} - h(t, y^{l_{t-1}})\|$, be the function that maps a T-length sequence of four-tuples τ and the initial state of the system x_0 into the lower bound on $J_T^h(x_0)$ proved by Theorem 4.1.

Let $\mathcal{F}^T$ denote the set of all possible T-length sequences of four-tuples built from the elements of $\mathcal{F}$, and let $B_{\mathcal{F}^T}^*(x_0) = \max_{\tau \in \mathcal{F}^T} B^h(\tau, x_0)$.

In this section, we provide an algorithm for computing in an efficient way the value of $B_{\mathcal{F}^T}^*(x_0)$. A naive approach for computing this value would consist in doing an exhaustive search over all the elements of $\mathcal{F}^T$. However, as soon as the optimization horizon T grows, this approach becomes computationally impractical even if $\mathcal{F}$ has only a handful of elements.

Our algorithm for computing $B_{\mathcal{F}^T}^*(x_0)$ is summarized in Table II. It is in essence identical to the Viterbi algorithm [14], and we observe that its complexity is linear with respect to the optimization horizon T and quadratic with respect to the size $|\mathcal{F}|$ of the sample of four-tuples.

The rationale behind this algorithm is the following. Let us first introduce some notations. Let $\tau(i)$ denote the index of the ith four-tuple of the sequence τ ($\tau(i) = l_i$), let $B^h(\tau, x_0)(j) = \sum_{t=0}^{j}(r^{l_t} - L_{Q_{T-t}}\delta_t)$ and let τ^* be a sequence of tuples such that $\tau^* \in \arg\max_{\tau \in \mathcal{F}^T} B^h(\tau, x_0)$.

We have that

$$B_{\mathcal{F}^T}^*(x_0) = B^h(\tau^*, x_0)(T-2) + V_1(\tau^*(T-1))$$

where V_1 is a $|\mathcal{F}|$-dimensional vector whose ith component is:

$$\max_{i'}\big(r^{i'} - L_{Q_1}\big(\|x^{i'} - y^i\| + \|u^{i'} - h(T-1, y^i)\|\big)\big).$$

Now let use observe that:

$$B_{\mathcal{F}^T}^*(x_0) = B^h(\tau^*, x_0)(T-3) + V_2(\tau^*(T-2))$$

where V_2 is a $|\mathcal{F}|$-dimensional vector whose ith component is:

$$\max_{i'}\big(r^{i'} - L_{Q_2}\big(\|x^{i'} - y^i\| + \|u^{i'} - h(T-2, y^i)\|\big) + V_1(i')\big).$$

By proceeding recursively, it is therefore possible to determine the value of $B^h(\tau^*, x_0) = B_{\mathcal{F}^T}^*(x_0)$ without having to screen all the elements of $\mathcal{F}^T$.

Inputs: An initial state x_0, a policy h, a set of four-tuples $\mathcal{F} = \{(x^l, u^l, r^l, y^l)\}_{l=1}^{|\mathcal{F}|}$ and three constants L_f, L_ρ, L_h which satisfy inequalities (4-6).
Output: A lower bound on $J_T^h(x_0)$ equal to $B^*_{\mathcal{F}^T}(x_0)$.
Algorithm:
Create two $|\mathcal{F}|$-dimensional vectors V_A and V_B
Set $V_A(i) = 0$ and $V_B(i) = 0$, $\forall i = \{1, \ldots, |\mathcal{F}|\}$
For $t = T-1$ to 1 **do**
  For $i = 1, \ldots, |\mathcal{F}|$ **do**, *(update the value of V_A)*
    Set $L_{Q_{T-t}} = L_\rho\Big(\sum_{k=0}^{T-t-1}[L_f(1+L_h)]^k\Big)$
    Set $u = h(t, y^i)$
    Set $V_A(i) = \max_{i'}(r^{i'} - L_{Q_{T-t}}(\|x^{i'} - y^i\| + \|u^{i'} - u\|) + V_B(i'))$
  end for
  Set $V_B = V_A$
end for
Set $u_0 = h(0, x_0)$
Set $lb^* = \max_{i'}\Big(r^{i'} - L_{Q_T}(\|x^{i'} - x_0\| + \|u^{i'} - u_0\|) + V_B(i')\Big)$
Return lb^*

TABLE II
A VITERBI-LIKE ALGORITHM FOR COMPUTING THE HIGHEST LOWER BOUND $B^*_{\mathcal{F}^T}(x_0)$ (SEE EQN (15)) OVER ALL THE SEQUENCES OF FOUR-TUPLES τ MADE FROM ELEMENTS OF $\mathcal{F}$.

Although this is rather evident, we want to stress the fact that $B^*_{\mathcal{F}^T}(x_0)$ can not decrease when new elements are added to $\mathcal{F}$. In other words, the quality of this lower bound is monotonically increasing when new samples are collected. To quantify this behavior, we characterize in the next section the tightness of this lower bound as a function of the density of the sample of four-tuples.

VI. TIGHTNESS OF THE LOWER BOUND $B^*_{\mathcal{F}^T}(x_0)$

In this section we study the relation of the tightness of $B^*_{\mathcal{F}^T}(x_0)$ with respect to the distance between the elements $(x, u) \in X \times U$ and the pairs (x^l, u^l) formed by the two first elements of the four-tuples composing $\mathcal{F}$. We prove in Theorem 6.1 that if $X \times U$ is bounded, then

$$J_T^h(x_0) - B^*_{\mathcal{F}^T}(x_0) \leq C\alpha^*,$$

where C is a constant depending only on the control problem and where α^* is the maximum distance from any $(x, u) \in X \times U$ to its closest neighbor in $\{(x^l, u^l)\}_{l=1}^{|\mathcal{F}|}$.

The main philosophy behind the proof is the following. First, a sequence of four-tuples whose state-action pairs (x^{l_t}, u^{l_t}) stand close to the different state-action pairs (x_t, u_t) visited when the system is controlled by h is built. Then, it is shown that the lower bound B computed when considering this particular sequence is such that $J_T^h(x_0) - B \leq C\alpha^*$. From there, the proof follows immediately.

Theorem 6.1: Let x_0 be an initial state, h a policy, and $\mathcal{F} = \{(x^l, u^l, r^l, y^l)\}_{l=1}^{|\mathcal{F}|}$ a set of four-tuples. We suppose that

$$\exists\ \alpha \in \mathbb{R}^+ :$$

$$\sup_{(x,u)\in X\times U}\Big\{\min_{l\in\{1,\ldots,|\mathcal{F}|\}}\{\|x^l - x\| + \|u^l - u\|\}\Big\} \leq \alpha, \quad (16)$$

and we note α^* the smallest constant which satisfies (16). Then

$$\exists\ C \in \mathbb{R}^+ : J_T^h(x_0) - B^*_{\mathcal{F}^T}(x_0) \leq C\alpha^*. \quad (17)$$

Proof:
Let $(x_0, u_0, r_0, x_1, u_1, \ldots, x_{T-1}, u_{T-1}, r_{T-1}, x_T)$ be the trajectory of the system starting from x_0 when the actions are selected $\forall t \in \{0, 1, \ldots, T-1\}$ according to the policy h.

Let $\tau = [(x^{l_t}, u^{l_t}, r^{l_t}, y^{l_t})]_{t=0}^{T-1}$ be a sequence of four-tuples that satisfies $\forall t \in \{0, 1, \ldots, T-1\}$

$$\|x^{l_t} - x_t\| + \|u^{l_t} - u_t\| = \min_{l\in\{1,\ldots,|\mathcal{F}|\}}\|x^l - x_t\| + \|u^l - u_t\|.$$

We have

$$B^h(\tau, x_0) = \sum_{t=0}^{T-1}[r^{l_t} - L_{Q_{T-t}}\delta_t]$$

where

$$\delta_t = \|x^{l_t} - y^{l_{t-1}}\| + \|u^{l_t} - h(t, y^{l_{t-1}})\|\ \ \forall t \in \{0, 1, \ldots, T-1\}.$$

Let us focus on δ_t. We have that

$$\delta_t = \|x^{l_t} - x_t + x_t - y^{l_{t-1}}\| + \|u^{l_t} - u_t + u_t - h(t, y^{l_{t-1}})\|,$$

and hence

$$\delta_t \leq \|x^{l_t} - x_t\| + \|x_t - y^{l_{t-1}}\| + \|u^{l_t} - u_t\| + \|u_t - h(t, y^{l_{t-1}})\|.$$

Using inequality (16), we can write

$$\|x^{l_t} - x_t\| + \|u^{l_t} - u_t\| \leq \alpha^*,$$

and so we have

$$\delta_t \leq \alpha^* + \|x_t - y^{l_{t-1}}\| + \|u_t - h(t, y^{l_{t-1}})\|. \quad (18)$$

- On the one hand, we have

$$\|x_t - y^{l_{t-1}}\| = \|f(x_{t-1}, u_{t-1}) - f(x^{l_{t-1}}, u^{l_{t-1}})\|$$

and the Lipschitz continuity of f implies that

$$\|x_t - y^{l_{t-1}}\| \leq L_f\big(\|x_{t-1} - x^{l_{t-1}}\| + \|u_{t-1} - u^{l_{t-1}}\|\big),$$

so, as $\|x_{t-1} - x^{l_{t-1}}\| + \|u_{t-1} - u^{l_{t-1}}\| \leq \alpha^*$, we have

$$\|x_t - y^{l_{t-1}}\| \leq L_f\alpha^*. \quad (19)$$

- On the other hand, we have

$$\|u_t - h(t, y^{l_{t-1}})\| = \|h(t, x_t) - h(t, y^{l_{t-1}})\|$$

and the Lipschitz continuity of h implies that

$$\|u_t - h(t, y^{l_{t-1}})\| \leq L_h\|x_t - y^{l_{t-1}}\|.$$

Since $\|x_t - y^{l_{t-1}}\| \leq L_f \alpha^*$ (see (19)) we obtain

$$\|u_t - h(t, y^{l_{t-1}})\| \leq L_h L_f \alpha^*. \quad (20)$$

Furthermore, (18), (19) and (20) imply that

$$\delta_t \leq \alpha^* + L_f \alpha^* + L_h L_f \alpha^* = \alpha^*(1 + L_f(1 + L_h))$$

and

$$B^h(\tau, x_0) \geq \sum_{t=0}^{T-1} [r^{l_t} - L_{Q_{T-t}} \alpha^*(1 + L_f(1 + L_h))] \doteq B.$$

We also have, by definition of $B^*_{\mathcal{F}^T}(x_0)$

$$J^h_T(x_0) \geq B^*_{\mathcal{F}^T}(x_0) \geq B^h(\tau, x_0) \geq B.$$

Thus

$$|J^h_T(x_0) - B^*_{\mathcal{F}^T}(x_0)| \leq |J^h_T(x_0) - B| = J^h_T(x_0) - B,$$

and we have
$J^h_T(x_0) - B = |\sum_{t=0}^{T-1} [(r_t - r^{l_t}) + L_{Q_{T-t}} \alpha^*(1 + L_f(1 + L_h))]|,$

$$J^h_T(x_0) - B \leq \sum_{t=0}^{T-1} [|r_t - r^{l_t}| + L_{Q_{T-t}} \alpha^*(1 + L_f(1 + L_h))].$$

The Lipschitz continuity of ρ allows to write
$|r_t - r^{l_t}| = |\rho(x_t, u_t) - \rho(x^{l_t}, u^{l_t})| \leq L_\rho(\|x_t - x^{l_t}\| + \|u_t - u^{l_t}\|),$

and using inequality (16), we have

$$|r_t - r^{l_t}| \leq L_\rho \alpha^*.$$

Finally, we obtain
$J^h_T(x_0) - B \leq \sum_{t=0}^{T-1} [L_\rho \alpha^* + L_{Q_{T-t}} \alpha^*(1 + L_f(1 + L_h))],$

$$J^h_T(x_0) - B \leq T L_\rho \alpha^* + \sum_{t=0}^{T-1} L_{Q_{T-t}} \alpha^*(1 + L_f(1 + L_h)),$$

$$J^h_T(x_0) - B \leq \alpha^* \left(T L_\rho + \sum_{t=0}^{T-1} L_{Q_{T-t}} \big(1 + L_f(1 + L_h)\big) \right).$$

Thus
$J^h_T(x_0) - B^*(x_0) \leq$

$$\alpha^* \left(T L_\rho + \sum_{t=0}^{T-1} L_{Q_{T-t}} \big(1 + L_f(1 + L_h)\big) \right),$$

which completes the proof. ∎

VII. Conclusions and future research

We have introduced in this paper an approach for deriving from a sample of trajectories a *lower* bound on the finite-horizon return of any policy from any given initial state. We also have proposed a dynamic programming (Viterbi-like) algorithm for computing this lower bound whose complexity is linear in the optimization horizon and quadratic in the total number of state transitions of the sample of trajectories. This approach and algorithm may directly be transposed in order to compute an *upper* bound, so as to bracket the performance of the given policy, when applied to a given initial state. We also have derived a characterization of these bounds, in terms of the density of the coverage of the state-action space by the sample of trajectories used to compute them. This analysis shows that the lower (and upper) bound converges at least linearly towards the true value of the return with the density of the sample (measured by the maximal distance of any state-action pair to this sample).

The Lipschitz continuity assumptions upon which the results have been built may seem restrictive, and they indeed are. Indeed, when facing a real-life problem, it may be difficult to establish whether its systems dynamics and reward function are indeed Lipschitz continuous. Secondly, even if one can guarantee that the Lipschitz assumptions are satisfied, it is still important to be able to establish some not too-conservative approximations of the Lipschitz constants. Indeed, the larger they are, the looser the bounds will be. In the same order of ideas, the choice of the norms on the state space and the action space might influence the value of the bounds and should thus also be chosen carefully.

While the approach has been designed for computing some lower bounds on the cumulated reward obtained by a given policy, it could also serve as the base for designing new reinforcement learning algorithms which would output policies that lead to the maximization of these lower bounds.

The proposed approach could also be used in combination with batch-mode reinforcement learning algorithms for identifying the pieces of trajectories that influence the most the lower bounds of the RL policy and, from there, for selecting a concise set of four-tuples from which it is possible to extract a good policy. This problem is particularly important when batch-mode RL algorithms are used to design autonomous intelligent agents. Indeed, after a certain time of interaction with their environment, the sample of information these agents collect may become so numerous that batch-mode RL techniques may become computationally impractical [4].

Since there exist in this context many non-deterministic problems for which it would be interesting to be able to have a lower bound on the performances of a policy (e.g., those related to the inference from clinical data of decision rules for treating chronic-like diseases [10]), extending our approach to stochastic systems would certainly be relevant. Future research on this topic could follow several paths: the study of lower bounds on the expected cumulated rewards, the design of worst-case lower bounds, a study of the case where the disturbances are part of the trajectories, etc.

Acknowledgements

This paper 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. Damien Ernst acknowledges the financial support of the Belgian National Fund of Scientific Research (FNRS) of which he is a Research Associate. The authors are also very grateful to Florence Belmudes, Bertrand Cornélusse, Jing Dai, Boris Defourny and Renaud Detry for

their helpful suggestions for improving the quality of the manuscript.

REFERENCES

[1] D.P. Bertsekas. *Dynamic Programming and Optimal Control*, volume III. Athena Scientific, Belmont, MA, 2nd edition, 2005.

[2] D.P. Bertsekas and J.N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[3] E.F. Camacho and C. Bordons. *Model Predictive Control*. Springer, 2004.

[4] D. Ernst. Selecting concise sets of samples for a reinforcement learning agent. In *Proceedings of the Third International Conference on Computational Intelligence, Robotics and Autonomous Systems (CIRAS 2005)*, page 6, 2005.

[5] D. Ernst, P. Geurts, and L. Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.

[6] J.E. Ingersoll. *Theory of Financial Decision Making*. Rowman and Littlefield Publishers, Inc., 1987.

[7] M. Kearns and S. Singh. Finite-sample convergence rates for Q-learning and indirect algorithms. In *In Neural Information Processing Systems 12*, pages 996–1002. MIT Press, 1999.

[8] M. Lagoudakis and R. Parr. Least-squares policy iteration. *Jounal of Machine Learning Research*, 4:1107–1149, 2003.

[9] S.A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society, Series B*, 65(2):331–366, 2003.

[10] S.A. Murphy. An experimental design for the development of adaptative treatment strategies. *Statistics in Medicine*, 24:1455–1481, 2005.

[11] D. Ormoneit and S. Sen. Kernel-based reinforcement learning. *Machine Learning*, 49(2-3):161–178, 2002.

[12] R.E. Schapire. On the worst-case analysis of temporal-difference learning algorithms. *Machine Learning*, 22(1/2/3), 1996.

[13] R.S. Sutton and A.G. Barto. *Reinforcement Learning, an Introduction*. MIT Press, 1998.

[14] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260– 269, 1967.