Towards Min Max Generalization in Reinforcement Learning

Raphael Fonteneau[1], Susan A. Murphy[2], Louis Wehenkel[1], and Damien Ernst[1]

[1] University of Liège, Belgium
{raphael.fonteneau,l.wehenkel,dernst}@ulg.ac.be
[2] University of Michigan, USA
samurphy@umich.edu

Abstract. In this paper, we introduce a min max approach for addressing the generalization problem in Reinforcement Learning. The min max approach works by determining a sequence of actions that maximizes the worst return that could possibly be obtained considering any dynamics and reward function compatible with the sample of trajectories and some prior knowledge on the environment. We consider the particular case of deterministic Lipschitz continuous environments over continuous state spaces, finite action spaces, and a finite optimization horizon. We discuss the non-triviality of computing an exact solution of the min max problem even after reformulating it so as to avoid search in function spaces. For addressing this problem, we propose to replace, inside this min max problem, the search for the worst environment given a sequence of actions by an expression that lower bounds the worst return that can be obtained for a given sequence of actions. This lower bound has a tightness that depends on the sample sparsity. From there, we propose an algorithm of polynomial complexity that returns a sequence of actions leading to the maximization of this lower bound. We give a condition on the sample sparsity ensuring that, for a given initial state, the proposed algorithm produces an optimal sequence of actions in open-loop. Our experiments show that this algorithm can lead to more cautious policies than algorithms combining dynamic programming with function approximators.

1 Introduction

Since the late sixties, the field of Reinforcement Learning (RL) [27] has studied the problem of inferring from the sole knowledge of observed system trajectories, near-optimal solutions to optimal control problems. The original motivation was to design computational agents able to learn by themselves how to interact in a rational way with their environment. The techniques developed in this field have appealed researchers trying to solve sequential decision making problems in many fields such as Finance [15], Medicine [19, 20] or Engineering [23].

RL algorithms are challenged when dealing with large or continuous state spaces. Indeed, in such cases they have to generalize the information contained in a generally sparse sample of trajectories. The dominating approach for generalizing this information is to combine RL algorithms with function approximators [2, 16, 9]. Usually, these approximators generalize the information contained in the sample to areas poorly covered by the sample by implicitly assuming that the properties of the system in those areas are similar to the properties of the system in the nearby areas well covered by the sample. This in turn often leads to low performance guarantees on the inferred policy when large state space areas are poorly covered by the sample. This can be explained by the fact that when computing the performance guarantees of these policies, one needs to take into account that they may actually drive the system into the poorly visited areas to which the generalization strategy associates a favorable environment behavior, while the environment may actually be particularly adversarial in those areas. This is corroborated by theoretical results which show that the performance guarantees of the policies inferred by these algorithms degrade with the sample sparsity where, loosely speaking, the sparsity can be seen as the radius of the largest non-visited state space area.[3]

As in our previous work [12] from which this paper is an extended version, we assume a deterministic Lipschitz continuous environment over continuous state spaces, finite action spaces, and a finite time-horizon. In this context, we introduce a min max approach to address the generalization problem. The min max approach works by determining a sequence of actions that maximizes the worst return that could possibly be obtained considering any dynamics and reward functions compatible with the sample of trajectories, and a weak prior knowledge given in the form of upper bounds on the Lipschitz constants of the environment. However, we show that finding an exact solution of the min max problem is far from trivial, even after reformulating the problem so as to avoid the search in the space of all compatible functions. To circumvent these difficulties, we propose to replace, inside this min max problem, the search for the worst environment given a sequence of actions by an expression that lower bounds the worst return that can be obtained for a given sequence of actions. This lower bound is derived from [11] and has a tightness that depends on the sample sparsity. From there, we propose a Viterbi–like algorithm [28] for computing an open-loop sequence of actions to be used from a given initial state to maximize that lower bound. This algorithm is of polynomial computational complexity in the size of the dataset and the optimization horizon. It is named CGRL for Cautious Generalization (oriented) Reinforcement Learning since it essentially shows a cautious behaviour in the sense that it computes decisions that avoid driving the system into areas of the state space that are not well enough covered by the available dataset, according to the prior information about the dynamics and reward function.

[3] Usually, these theoretical results do not give lower bounds per se but a distance between the actual return of the inferred policy and the optimal return. However, by adapting in a straightforward way the proofs behind these results, it is often possible to get a bound on the distance between the estimate of the return of the inferred policy computed by the RL algorithm and its actual return and, from there, a lower bound on the return of the inferred policy.

Besides, the CGRL algorithm does not rely on function approximators and it computes, as a byproduct, a lower bound on the return of its open-loop sequence of decisions. We also provide a condition on the sample sparsity ensuring that, for a given initial state, the CGRL algorithm produces an optimal sequence of actions in open-loop, and we suggest directions for leveraging our approach to a larger class of problems in RL.

The rest of the paper is organized as follows. Section 2 briefly discusses related work. In Section 3, we formalize the min max approach to generalization, and we discuss its non trivial nature in Section 4. In Section 5, we exploit the results of [11] for lower bounding the worst return that can be obtained for a given sequence of actions. Section 6 proposes a polynomial algorithm for inferring a sequence of actions maximizing this lower bound and states a condition on the sample sparsity for its optimality. Section 7 illustrates the features of the proposed algorithm and Section 8 discusses its interest, while Section 9 concludes.

2 Related work

The min max approach to generalization followed by the CGRL algorithm results in the output of policies that are likely to drive the agent only towards areas well enough covered by the sample. Heuristic strategies have already been proposed in the RL literature to infer policies that exhibit such a conservative behavior. As a way of example, some of these strategies associate high negative rewards to trajectories falling outside of the well covered areas. Other works in RL have already developped min max strategies when the environment behavior is partially unknown [17, 4, 24]. However, these strategies usually consider problems with finite state spaces where the uncertainities come from the lack of knowledge of the transition probabilities [7, 5]. In model predictive control (MPC) where the environment is supposed to be fully known [10], min max approaches have been used to determine the optimal sequence of actions with respect to the "worst case" disturbance sequence occuring [1]. The CGRL algorithm relies on a methodology for computing a lower bound on the worst possible return (considering any compatible environment) in a deterministic setting with a mostly unknown actual environment. In this, it is related to works in the field of RL which try to get from a sample of trajectories lower bounds on the returns of inferred policies [18, 22].

3 Problem Statement

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

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

where for all t, the state x_t is an element of the compact state space $\mathcal{X} \subset \mathbb{R}^{d_\mathcal{X}}$ where $\mathbb{R}^{d_\mathcal{X}}$ denotes the $d_\mathcal{X}-$dimensional Euclidian space and u_t is an element of the finite (discrete) action space $\mathcal{U}$. $T \in \mathbb{N}_0$ is referred to as the optimization horizon. An instantaneous reward $r_t = \rho(x_t, u_t) \in \mathbb{R}$ is associated with the action u_t taken while being in state x_t. For every initial state $x \in \mathcal{X}$ and for every sequence of actions $(u_0, \ldots, u_{T-1}) \in \mathcal{U}^T$, the cumulated reward over T stages (also named return over T stages) is defined as

$$J^{u_0,\ldots,u_{T-1}}(x) = \sum_{t=0}^{T-1} \rho(x_t, u_t) \ ,$$

where $x_{t+1} = f(x_t, u_t)$,$\forall t \in \{0, \ldots, T-1\}$ and $x_0 = x$. We assume that the system dynamics f and the reward function ρ are Lipschitz continuous, i.e. that there exist finite constants $L_f, L_\rho \in \mathbb{R}$ such that: $\forall x', x'' \in \mathcal{X}, \forall u \in \mathcal{U}$,

$$\begin{aligned} \|f(x', u) - f(x'', u)\|_{\mathcal{X}} &\leq L_f \|x' - x''\|_{\mathcal{X}} \ , \\ |\rho(x', u) - \rho(x'', u)| &\leq L_\rho \|x' - x''\|_{\mathcal{X}} \ , \end{aligned}$$

where $\|.\|_{\mathcal{X}}$ denotes the Euclidian norm over the space $\mathcal{X}$. We further suppose that: (i) the system dynamics f and the reward function ρ are unknown, (ii) a set of one-step transitions $\mathcal{F}_n = \{(x^l, u^l, r^l, y^l)\}_{l=1}^n$ is known where each one-step transition is such that $y^l = f(x^l, u^l)$ and $r^l = \rho(x^l, u^l)$, (iii) $\forall a \in \mathcal{U}$, $\exists (x, u, r, y) \in \mathcal{F}_n : u = a$ (each action $a \in \mathcal{U}$ appears at least once in $\mathcal{F}_n$) and (iv) two constants L_f and L_ρ satisfying the above-written inequalities are known.[4] We define the set of functions $\mathcal{L}^f_{\mathcal{F}_n}$ (resp. $\mathcal{L}^\rho_{\mathcal{F}_n}$) from $\mathcal{X} \times \mathcal{U}$ into $\mathcal{X}$ (resp. into $\mathbb{R}$) as follows :

$$\mathcal{L}^f_{\mathcal{F}_n} = \left\{ f' : \mathcal{X} \times \mathcal{U} \to \mathcal{X} \,\middle|\, \left\{ \begin{array}{l} \forall x', x'' \in \mathcal{X}, \forall u \in \mathcal{U} \ , \\ \|f'(x', u) - f'(x'', u)\|_{\mathcal{X}} \leq L_f \|x' - x''\|_{\mathcal{X}} \ , \\ \forall l \in \{1, \ldots, n\}, f'(x^l, u^l) = f(x^l, u^l) = y^l \end{array} \right. \right\} \ ,$$

$$\mathcal{L}^\rho_{\mathcal{F}_n} = \left\{ \rho' : \mathcal{X} \times \mathcal{U} \to \mathbb{R} \,\middle|\, \left\{ \begin{array}{l} \forall x', x'' \in \mathcal{X}, \forall u \in \mathcal{U} \ , \\ |\rho'(x', u) - \rho'(x'', u)| \leq L_\rho \|x' - x''\|_{\mathcal{X}} \ , \\ \forall l \in \{1, \ldots, n\}, \rho'(x^l, u^l) = \rho(x^l, u^l) = r^l \end{array} \right. \right\} \ .$$

In the following, we call a "compatible environment" any pair $(f', \rho') \in \mathcal{L}^f_{\mathcal{F}_n} \times \mathcal{L}^\rho_{\mathcal{F}_n}$. Given a compatible environment (f', ρ'), a sequence of actions $(u_0, \ldots, u_{T-1}) \in \mathcal{U}^T$ and an initial state $x \in \mathcal{X}$, we introduce the $(f', \rho')-$return over T stages when starting from $x \in \mathcal{X}$:

$$J^{u_0,\ldots,u_{T-1}}_{(f',\rho')}(x) = \sum_{t=0}^{T-1} \rho'(x'_t, u_t) \ ,$$

where $x'_0 = x$ and $x'_{t+1} = f'(x'_t, u_t)$, $\forall t \in \{0, \ldots, T-1\}$. We introduce $L^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x)$ such that

$$L^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x) = \min_{(f',\rho') \in \mathcal{L}^f_{\mathcal{F}_n} \times \mathcal{L}^\rho_{\mathcal{F}_n}} \left\{ J^{u_0,\ldots,u_{T-1}}_{(f',\rho')}(x) \right\} \ .$$

[4] These constants do not necessarily have to be the smallest ones satisfying these inequalities (i.e., the Lispchitz constants).

The existence of $L^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x)$ is ensured by the following arguments: (i) the space $\mathcal{X}$ is compact, (ii) the set $\mathcal{L}^f_{\mathcal{F}_n} \times \mathcal{L}^\rho_{\mathcal{F}_n}$ is closed and bounded considering the $\|.\|_\infty$ norm ($\|(f',\rho')\|_\infty = \sup_{(x,u)\in\mathcal{X}\times\mathcal{U}} \|(f'(x,u),\rho'(x,u))\|_{\mathbb{R}^{d_\mathcal{X}+1}}$ where $\|.\|_{\mathbb{R}^{d_\mathcal{X}+1}}$ is the Euclidian norm over $\mathbb{R}^{d_\mathcal{X}+1}$) and (iii) one can show that the mapping $\mathcal{M}^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n,x} : \mathcal{L}^f_{\mathcal{F}_n} \times \mathcal{L}^\rho_{\mathcal{F}_n} \to \mathbb{R}$ such that $\mathcal{M}^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n,x}(f',\rho') = J^{u_0,\ldots,u_{T-1}}_{(f',\rho')}(x)$ is a continuous mapping. Furthermore, this also proves that

$$\forall(u_0,\ldots,u_{T-1}) \in \mathcal{U}^T, \forall x \in \mathcal{X}, \exists (f^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n,x}, \rho^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n,x}) \in \mathcal{L}^f_{\mathcal{F}_n} \times \mathcal{L}^\rho_{\mathcal{F}_n} : \\ J^{u_0,\ldots,u_{T-1}}_{(f^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n,x},\rho^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n,x})}(x) = L^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x). \tag{1}$$

Our goal is to compute, given an initial state $x \in \mathcal{X}$, an open-loop sequence of actions $(\dot{u}_0(x),\ldots,\dot{u}_{T-1}(x)) \in \mathcal{U}^T$ that gives the highest return in the least favorable compatible environment. This problem can be formalized as the min max problem:

$$(\dot{u}_0(x),\ldots,\dot{u}_{T-1}(x)) \in \operatorname*{arg\,max}_{(u_0,\ldots,u_{T-1})\in\mathcal{U}^T} \left\{ L^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x) \right\}.$$

4 Reformulation of the min max problem

Since $\mathcal{U}$ is finite, one could solve the min max problem by computing for each $(u_0,\ldots,u_{T-1}) \in \mathcal{U}^T$ the value of $L^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x)$. As the latter computation is posed as an infinite-dimensional minimization problem over the function space $\mathcal{L}^f_{\mathcal{F}_n} \times \mathcal{L}^\rho_{\mathcal{F}_n}$, we first show that it can be reformulated as a finite-dimensional problem over $\mathcal{X}^{T-1} \times \mathbb{R}^T$. This is based on the observation that $L^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x)$ is actually equal to the lowest sum of rewards that could be collected along a trajectory compatible with an environment from $\mathcal{L}^f_{\mathcal{F}_n} \times \mathcal{L}^\rho_{\mathcal{F}_n}$, and is precisely stated by the following Theorem (see Appendix A for the proof).

Theorem 1 (Equivalence). *Let $(u_0,\ldots,u_{T-1}) \in \mathcal{U}^T$ and $x \in \mathcal{X}$. Let $K^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x)$ be the solution of the following optimization problem:*

$$K^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x) = \min_{\substack{\hat{r}_0\ \ldots\ \hat{r}_{T-1} \in \mathbb{R} \\ \hat{x}_0\ \ldots\ \hat{x}_{T-1} \in \mathcal{X}}} \left\{ \sum_{t=0}^{T-1} \hat{r}_t \right\},$$

where the variables $\hat{x}_0,\ldots,\hat{x}_{T-1}$ and $\hat{r}_0,\ldots,\hat{r}_{T-1}$ satisfy the constraints

$$\left.\begin{array}{l} |\hat{r}_t - r^{l_t}| \le L_\rho \|\hat{x}_t - x^{l_t}\|_\mathcal{X} \ , \\ \|\hat{x}_{t+1} - y^{l_t}\|_\mathcal{X} \le L_f \|\hat{x}_t - x^{l_t}\|_\mathcal{X} \end{array}\right\} \forall l_t \in \{1,\ldots,n | u^{l_t} = u_t\} \ ,$$

$$\left.\begin{array}{l} |\hat{r}_t - \hat{r}_{t'}| \le L_\rho \|\hat{x}_t - \hat{x}_{t'}\|_\mathcal{X} \ , \\ \|\hat{x}_{t+1} - \hat{x}_{t'+1}\|_\mathcal{X} \le L_f \|\hat{x}_t - \hat{x}_{t'}\|_\mathcal{X} \end{array}\right\} \forall t,t' \in \{0,\ldots,T-1 | u_t = u_{t'}\} \ ,$$

$$\hat{x}_0 = x \ .$$

Then,

$$K^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x) = L^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x) \ .$$

Unfortunately, this latter minimization problem turns out to be non-convex in its generic form and, hence "off the shelf" algorithms will only be able to provide upper bounds on its value. Furthermore, the overall complexity of an algorithm that would be based on the enumeration of $\mathcal{U}^T$, combined with a local optimizer for the inner loop, may be intractable as soon as the cardinality of the action space $\mathcal{U}$ and/or the optimization horizon T become large.

We leave the exploration of the above formulation for future research. Instead, in the following subsections, we use the results from [11] to define a maximal lower bound $B^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x) \leq L^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x)$ for a given initial state $x \in \mathcal{X}$ and a sequence $(u_0, \ldots, u_{T-1}) \in \mathcal{U}^T$. Furthermore, we show that the maximization of this lower bound $B^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x)$ with respect to the choice of a sequence of actions lends itself to a dynamic programming type of decomposition. In the end, this yields a polynomial algorithm for the computation of a sequence of actions $(\hat{u}^*_{\mathcal{F}_n,0}(x), \ldots, \hat{u}^*_{\mathcal{F}_n,T-1}(x))$ maximizing a lower bound of the original min max problem, i.e. $(\hat{u}^*_{\mathcal{F}_n,0}(x), \ldots, \hat{u}^*_{\mathcal{F}_n,T-1}(x)) \in \underset{(u_0,\ldots,u_{T-1})\in\mathcal{U}^T}{\arg\max} \left\{B^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x)\right\}$.

5 Lower bound on the return of a given sequence of actions

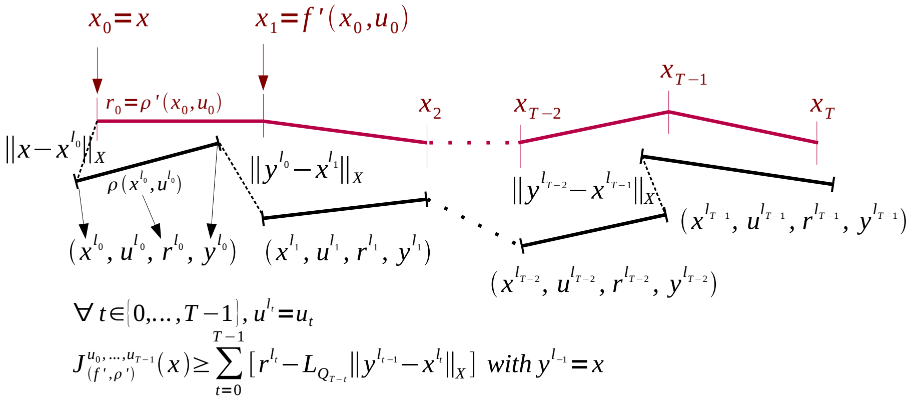
$x_0 = x$
$x_1 = f'(x_0, u_0)$
$r_0 = \rho'(x_0, u_0)$
x_2
x_{T-2}
x_{T-1}
x_T
$\|x - x^{l_0}\|_X$
$\rho(x^{l_0}, u^{l_0})$
$\|y^{l_0} - x^{l_1}\|_X$
$\|y^{l_{T-2}} - x^{l_{T-1}}\|_X$
$(x^{l_0}, u^{l_0}, r^{l_0}, y^{l_0})$
$(x^{l_1}, u^{l_1}, r^{l_1}, y^{l_1})$
$(x^{l_{T-2}}, u^{l_{T-2}}, r^{l_{T-2}}, y^{l_{T-2}})$
$(x^{l_{T-1}}, u^{l_{T-1}}, r^{l_{T-1}}, y^{l_{T-1}})$
$\forall t \in \{0, \ldots, T-1\}, u^{l_t} = u_t$
$J^{u_0,\ldots,u_{T-1}}_{(f',\rho')}(x) \geq \sum_{t=0}^{T-1} [r^{l_t} - L_{Q_{T-t}} \|y^{l_{t-1}} - x^{l_t}\|_X]$ with $y^{l_{-1}} = x$

Fig. 1. A graphical interpretation of the different terms composing the bound on $J^{u_0,\ldots,u_{T-1}}_{(f',\rho')}(x)$ computed from a sequence of one-step transitions.

In this section, we present a method for computing, from a given initial state $x \in \mathcal{X}$, a sequence of actions $(u_0, \ldots, u_{T-1}) \in \mathcal{U}^T$, a dataset of transitions, and weak prior knowledge about the environment, a lower bound on $L^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x)$. The method is adapted from [11]. In the following, we denote by $\mathcal{F}^T_{n,(u_0,\ldots,u_{T-1})}$

the set of all sequences of one-step system transitions $[(x^{l_0}, u^{l_0}, r^{l_0}, y^{l_0}), \ldots, (x^{l_{T-1}}, u^{l_{T-1}}, r^{l_{T-1}}, y^{l_{T-1}})]$ that may be built from elements of $\mathcal{F}_n$ and that are compatible with $u_0, \ldots, u_{T-1}$, i.e. for which $u^{l_t} = u_t$, $\forall t \in \{0, \ldots, T-1\}$. First, we compute a lower bound on $L^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x)$ from any given element τ from $\mathcal{F}^T_{n,(u_0,\ldots,u_{T-1})}$. This lower bound $B(\tau, x)$ is made of the sum of the T rewards corresponding to τ ($\sum_{t=0}^{T-1} r^{l_t}$) and T negative terms. Every negative term is associated with a one-step transition. More specifically, the negative term corresponding to the transition $(x^{l_t}, u^{l_t}, r^{l_t}, y^{l_t})$ of τ represents an upper bound on the 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 $y^{l_{-1}} = x$) and considering any compatible environment (f', ρ') from $\mathcal{L}^f_{\mathcal{F}_n} \times \mathcal{L}^\rho_{\mathcal{F}_n}$. By maximizing $B(\tau, x)$ over $\mathcal{F}^T_{n,(u_0,\ldots,u_{T-1})}$, we obtain a maximal lower bound on $L^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x)$. Furthermore, we prove that the distance from the maximal lower bound to the actual return $J^{u_0,\ldots,u_{T-1}}(x)$ can be characterized in terms of the sample sparsity.

5.1 Computing a bound from a given sequence of one-step transitions

We have the following lemma.

Lemma 1. *Let $(u_0, \ldots, u_{T-1}) \in \mathcal{U}^T$ be a sequence of actions and $x \in \mathcal{X}$ an initial state. Let $\tau = [(x^{l_t}, u^{l_t}, r^{l_t}, y^{l_t})]_{t=0}^{T-1} \in \mathcal{F}^T_{n,(u_0,\ldots,u_{T-1})}$. Then,*

$$B(\tau, x) \leq L^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x) \leq J^{u_0,\ldots,u_{T-1}}(x) \ ,$$

with

$$B(\tau, x) \doteq \sum_{t=0}^{T-1} \left[r^{l_t} - L_{Q_{T-t}} \| y^{l_{t-1}} - x^{l_t} \|_{\mathcal{X}} \right] \ , y^{l_{-1}} = x \ , L_{Q_{T-t}} = L_\rho \sum_{i=0}^{T-t-1} (L_f)^i \ .$$

The proof is given in Appendix B. The lower bound on $L^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x)$ derived in this lemma can be interpreted as follows. Given any compatible environment $(f', \rho') \in \mathcal{L}^f_{\mathcal{F}_n} \times \mathcal{L}^\rho_{\mathcal{F}_n}$, the sum of the rewards of the "broken" trajectory formed by the sequence of one-step system transitions τ can never be greater than $J^{u_0,\ldots,u_{T-1}}_{(f',\rho')}(x)$, provided that every reward r^{l_t} is penalized by a factor $L_{Q_{T-t}} \| y^{l_{t-1}} - x^{l_t} \|_{\mathcal{X}}$. This factor is in fact an upper bound on the variation of the $(T-t)$-state-action value function given any compatible environment (f', ρ') (see Appendix B) that can occur when "jumping" from $(y^{l_{t-1}}, u_t)$ to (x^{l_t}, u_t). An illustration of this is given in Figure 1.

5.2 Tightness of highest lower bound over all compatible sequences of one-step transitions

We define

$$B^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x) = \max_{\tau \in \mathcal{F}^T_{n,(u_0,\ldots,u_{T-1})}} B(\tau, x)$$

and we analyze in this subsection the distance from the lower bound $B^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x)$ to the actual return $J^{u_0,\ldots,u_{T-1}}(x)$ as a function of the sample sparsity. The sample sparsity is defined as follows: let $\mathcal{F}_{n,a} = \{(x^l, u^l, r^l, y^l) \in \mathcal{F}_n | u^l = a\}$ ($\forall a,\ \mathcal{F}_{n,a} \neq \emptyset$ according to assumption (iii) given in Section 3). Since $\mathcal{X}$ is a compact subset of $\mathbb{R}^{d_{\mathcal{X}}}$, it is bounded and there exists $\alpha \in \mathbb{R}^+$:

$$\forall a \in \mathcal{U}\ ,\ \sup_{x' \in X} \left\{ \min_{(x^l,u^l,r^l,y^l) \in \mathcal{F}_{n,a}} \left\{ \|x^l - x'\|_{\mathcal{X}} \right\} \right\} \leq \alpha\ . \tag{2}$$

The smallest α which satisfies equation (2) is named the sample sparsity and is denoted by $\alpha^*_{\mathcal{F}_n}$. We have the following theorem.

Theorem 2 (Tightness of highest lower bound).

$$\exists\ C > 0 :\ \forall (u_0, \ldots, u_{T-1}) \in \mathcal{U}^T, J^{u_0,\ldots,u_{T-1}}(x) - B^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x) \leq C \alpha^*_{\mathcal{F}_n}.$$

The proof of Theorem 2 is given in the Appendix of [12]. The lower bound $B^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x)$ thus converges to the $T-$stage return of the sequence of actions $(u_0, \ldots, u_{T-1}) \in \mathcal{U}^T$ when the sample sparsity $\alpha^*_{\mathcal{F}_n}$ decreases to zero.

6 Computing a sequence of actions maximizing the highest lower bound

Let $\mathfrak{B}^*_{\mathcal{F}_n}(x) = \underset{(u_0,\ldots,u_{T-1}) \in \mathcal{U}^T}{\arg\max} \left\{ B^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x) \right\}$. The CGRL algorithm computes for each initial state $x \in \mathcal{X}$ a sequence of actions $(\hat{u}^*_{\mathcal{F}_n,0}(x), \ldots, \hat{u}^*_{\mathcal{F}_n,T-1}(x))$ that belongs to $\mathfrak{B}^*_{\mathcal{F}_n}(x)$. From what precedes, it follows that the actual return $J^{\hat{u}^*_{\mathcal{F}_n,0}(x),\ldots,\hat{u}^*_{\mathcal{F}_n,T-1}(x)}(x)$ of this sequence is lower-bounded by the quantity $\max_{(u_0,\ldots,u_{T-1}) \in \mathcal{U}^T} B^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x)$. Due to the tightness of the lower bound $B^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x)$, the value of the return which is guaranteed will converge to the true return of the sequence of actions when $\alpha^*_{\mathcal{F}_n}$ decreases to zero. Additionaly, we prove in Section 6.1 that when the sample sparsity $\alpha^*_{\mathcal{F}_n}$ decreases below a particular threshold, the sequence $(\hat{u}^*_{\mathcal{F}_n,0}(x), \ldots, \hat{u}^*_{\mathcal{F}_n,T-1}(x))$ is optimal. To identify a sequence of actions that belongs to $\mathfrak{B}^*_{\mathcal{F}_n}(x)$ without computing for all sequences $(u_0, \ldots, u_{T-1}) \in \mathcal{U}^T$ the value $B^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x)$, the CGRL algorithm exploits the fact that the problem of finding an element of $\mathfrak{B}^*_{\mathcal{F}_n}(x)$ can be reformulated as a shortest path problem.

6.1 Convergence of $(\hat{u}^*_{\mathcal{F}_n,0}(x), \ldots, \hat{u}^*_{\mathcal{F}_n,T-1}(x))$ towards an optimal sequence of actions

We prove hereafter that when $\alpha^*_{\mathcal{F}_n}$ gets lower than a particular threshold, the CGRL algorithm can only output optimal policies.

Theorem 3 (Convergence of the CGRL algorithm).
Let

$$\mathfrak{J}^*(x) = \left\{ (u_0, \ldots, u_{T-1}) \in \mathcal{U}^T | J^{u_0,\ldots,u_{T-1}}(x) = J^*(x) \right\}\ ,$$

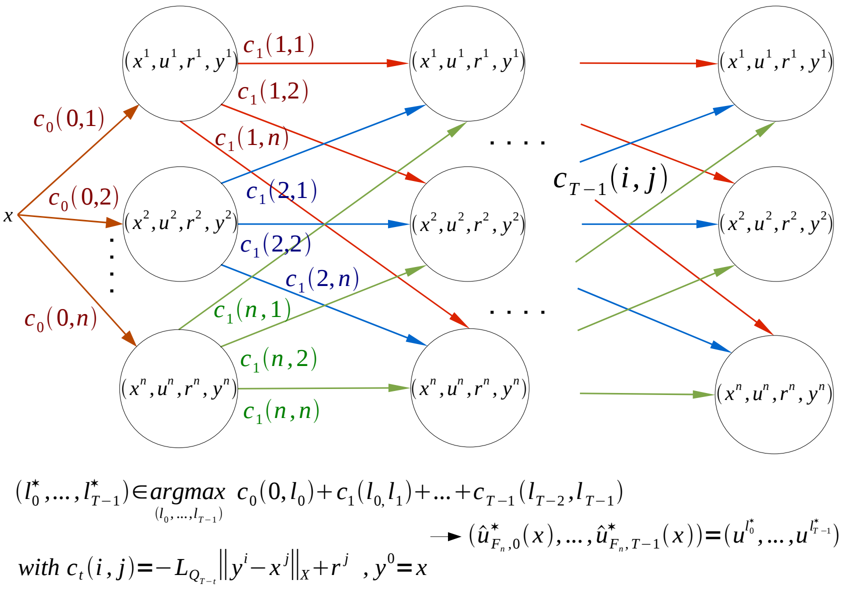
x
$c_0(0,1)$
$c_0(0,2)$
$c_0(0,n)$
(x^1,u^1,r^1,y^1)
(x^2,u^2,r^2,y^2)
(x^n,u^n,r^n,y^n)
$c_1(1,1)$
$c_1(1,2)$
$c_1(1,n)$
$c_1(2,1)$
$c_1(2,2)$
$c_1(2,n)$
$c_1(n,1)$
$c_1(n,2)$
$c_1(n,n)$
$c_{T-1}(i,j)$
$(l_0^*,\dots,l_{T-1}^*)\in \underset{(l_0,\dots,l_{T-1})}{argmax}\ c_0(0,l_0)+c_1(l_0,l_1)+\dots+c_{T-1}(l_{T-2},l_{T-1})$
$\rightarrow (\hat{u}^*_{F_n,0}(x),\dots,\hat{u}^*_{F_n,T-1}(x))=(u^{l_0^*},\dots,u^{l_{T-1}^*})$
with $c_t(i,j)=-L_{Q_{T-t}}\|y^i-x^j\|_X+r^j\ \ ,y^0=x$

Fig. 2. A graphical interpretation of the CGRL algorithm.

and let us suppose that $\mathfrak{J}^*(x) \neq \mathcal{U}^T$ *(if* $\mathfrak{J}^*(x) = \mathcal{U}^T$*, the search for an optimal sequence of actions is indeed trivial). We define*

$$\epsilon(x) = \min_{(u_0,\dots,u_{T-1})\in\mathcal{U}^T\setminus\mathfrak{J}^*(x)} \left\{J^*(x) - J^{u_0,\dots,u_{T-1}}(x)\right\} .$$

Then

$$C\alpha^*_{\mathcal{F}_n} < \epsilon(x) \implies (\hat{u}^*_{\mathcal{F}_n,0}(x),\dots,\hat{u}^*_{\mathcal{F}_n,T-1}(x)) \in \mathfrak{J}^*(x) .$$

The proof of Theorem 3 is given in the Appendix of [12].

6.2 Cautious Generalization Reinforcement Learning algorithm

The CGRL algorithm computes an element of the set $\mathfrak{B}^*_{\mathcal{F}_n}(x)$ defined previously. Let $\mathcal{D} : \mathcal{F}_n^T \to \mathcal{U}^T$ be the operator that maps a sequence of one-step system transitions $\tau = [(x^{l_t}, u^{l_t}, r^{l_t}, y^{l_t})]_{t=0}^{T-1}$ into the sequence of actions $(u^{l_0},\dots,u^{l_{T-1}})$. Using this operator, we can write

$$\mathfrak{B}^*_{\mathcal{F}_n}(x) = \left\{ (u_0,\dots,u_{T-1}) \in \mathcal{U}^T \middle| \begin{array}{l} \exists\tau \in \underset{\tau\in\mathcal{F}_n^T}{\arg\max}\left\{B(\tau,x)\right\}\ , \\ \mathcal{D}(\tau) = (u_0,\dots,u_{T-1}) \end{array} \right\} .$$

Or, equivalently

$$\mathfrak{B}^*_{\mathcal{F}_n}(x) = \left\{ (u_0, \ldots, u_{T-1}) \in \mathcal{U}^T \,\middle|\, \begin{array}{l} \exists \tau \in \underset{\tau \in \mathcal{F}_n^T}{\arg\max} \sum_{t=0}^{T-1} \left[r^{l_t} - L_{Q_{T-t}} \| y^{l_{t-1}} - x^{l_t} \|_{\mathcal{X}} \right] \ , \\ \mathcal{D}(\tau) = (u_0, \ldots, u_{T-1}) \end{array} \right\} .$$

From this expression, we can notice that a sequence of one-step transitions τ such that $\mathcal{D}(\tau)$ belongs to $\mathfrak{B}^*_{\mathcal{F}_n}(x)$ can be obtained by solving a shortest path problem on the graph given in Figure 2. The CGRL algorithm works by solving this problem using the Viterbi algorithm and by applying the operator $\mathcal{D}$ to the sequence of one-step transitions τ corresponding to its solution. Its complexity is quadratic with respect to the cardinality n of the input sample $\mathcal{F}_n$ and linear with respect to the optimization horizon T.

7 Illustration

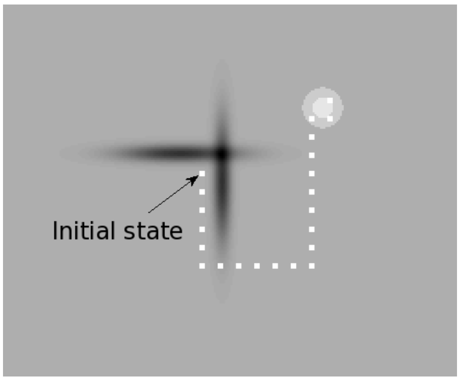

Fig. 3. CGRL with $\mathcal{F}$.

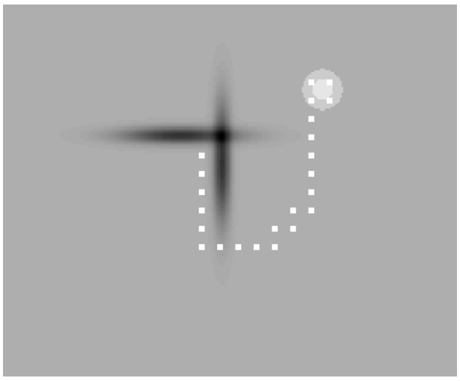

Fig. 4. FQI with $\mathcal{F}$.

In this section, we illustrate the CGRL algorithm on a variant of the puddle world benchmark introduced in [26]. In this benchmark, a robot whose goal is to collect high cumulated rewards navigates on a plane. A puddle stands in between the initial position of the robot and the high reward area. If the robot is in the puddle, it gets highly negative rewards. An optimal navigation strategy drives the robot around the puddle to reach the high reward area. Two datasets of one-step transitions have been used in our example. The first set $\mathcal{F}$ contains elements that uniformly cover the area of the state space that can be reached within T steps. The set $\mathcal{F}'$ has been obtained by removing from $\mathcal{F}$ the elements corresponding to the highly negative rewards.[5] The full specification of the benchmark and

[5] Although this problem might be treated by on-line learning methods, in some settings - for whatever reason - on-line learning may be impractical and all one will have is a batch of trajectories

the exact procedure for generating $\mathcal{F}$ and $\mathcal{F}'$ are given in [12]. On Figure 3, we

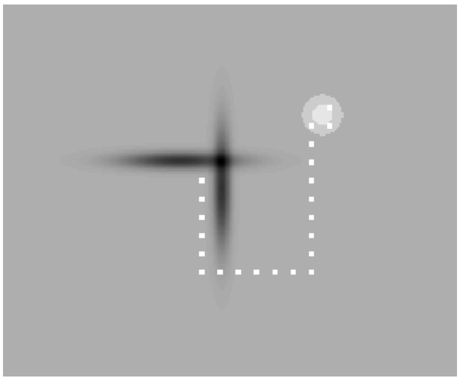

Fig. 5. CGRL with $\mathcal{F}'$.

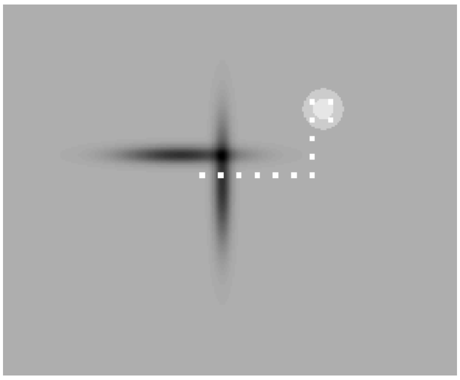

Fig. 6. FQI with $\mathcal{F}'$.

have drawn the trajectory of the robot when following the sequence of actions computed by the CGRL algorithm. Every state encountered is represented by a white square. The plane upon which the robot navigates has been colored such that the darker the area, the smaller the corresponding rewards are. In particular, the puddle area is colored in dark grey/black. We see that the CGRL policy drives the robot around the puddle to reach the high-reward area − which is represented by the light-grey circles. The CGRL algorithm also computes a lower bound on the cumulated rewards obtained by this action sequence. Here, we found out that this lower bound was rather conservative.

Figure 4 represents the policy inferred from $\mathcal{F}$ by using the (finite-time version of the) Fitted Q Iteration algorithm (FQI) combined with extremely randomized trees as function approximators [9]. The trajectories computed by the CGRL and FQI algorithms are very similar and so are the sums of rewards obtained by following these two trajectories. However, by using $\mathcal{F}'$ rather that $\mathcal{F}$, the CGRL and FQI algorithms do not lead to similar trajectories, as it is shown on Figures 5 and 6. Indeed, while the CGRL policy still drives the robot around the puddle to reach the high reward area, the FQI policy makes the robot cross the puddle. In terms of optimality, this latter navigation strategy is much worse. The difference between both navigation strategies can be explained as follows. The FQI algorithm behaves as if it were associating to areas of the state space that are not covered by the input sample, the properties of the elements of this sample that are located in the neighborhood of these areas. This in turn explains why it computes a policy that makes the robot cross the puddle. The same behavior could probably be observed by using other algorithms that combine dynamic programming strategies with kernel-based approximators or averagers [3, 14, 21]. The CGRL algorithm generalizes the information contained in the dataset, by assuming, given the intial state, the most adverse behavior for the environment according to its weak prior knowledge about the environment. This results in the fact that the CGRL algorithm penalizes sequences of decisions that could drive the robot in areas not well covered by the sample, and this explains why the CGRL algorithm drives the robot around the puddle when run with $\mathcal{F}'$.

8 Discussion

The CGRL algorithm outputs a sequence of actions as well as a lower bound on its return. When $L_f > 1$ (e.g. when the system is unstable), this lower bound will decrease exponentially with T. This may lead to very low performance guarantees when the optimization horizon T is large. However, one can also observe that the terms $L_{Q_{T-t}}$ − which are responsible for the exponential decrease of the lower bound with the optimization horizon − are multiplied by the distance between the end state of a one-step transition and the beginning state of the next one-step transition of the sequence τ ($\|y^{l^*_{t-1}} - x^{l^*_t}\|_{\mathcal{X}}$) solution of the shortest path problem of Figure 2. Therefore, if these states $y^{l^*_{t-1}}$ and $x^{l^*_t}$ are close to each other, the CGRL algorithm can lead to good performance guarantees even for large values of T. It is also important to notice that this lower bound does not depend explicitly on the sample sparsity $\alpha^*_{\mathcal{F}_n}$, but depends rather on the initial state for which the sequence of actions is computed. Therefore, this may lead to cases where the CGRL algorithm provides good performance guarantees for some specific initial states, even if the sample does not cover every area of the state space well enough.

Other RL algorithms working in a similar setting as the CGRL algorithm, while not exploiting the weak prior knowledge about the environment, do not output a lower bound on the return of the policy h they infer from the sample of trajectories $\mathcal{F}_n$. However, some lower bounds on the return of h can still be computed. For instance, this can be done by exploiting the results of [11] upon which the CGRL algorithm is based. However, one can show that following the strategy described in [11] would necessarily lead to a bound lower than the lower bound associated to the sequence of actions computed by the CGRL algorithm. Another strategy would be to design global lower bounds on their policy by adapting proofs used to establish the consistency of these algorithms. As a way of example, by proceeding like this, we can design a lower bound on the return of the policy given by the FQI algorithm when combined with some specific approximators which have, among others, Lipschitz continuity properties. These algorithms compute a sequence of state-action value functions $\hat{Q}_1, \hat{Q}_2, \dots, \hat{Q}_T$ and compute the policy $h : \{0, 1, \dots, T-1\} \times X$ defined as follows : $h(t, x_t) \in \arg\max_{u \in \mathcal{U}} \hat{Q}_{T-t}(x_t, u)$. For instance when using kernel-based approximators [21], we have as result that the return of h when starting from a state x is larger than $\hat{Q}_T(x, h(0, x)) - (C_1 T + C_2 T^2) \cdot \alpha^*_{\mathcal{F}_n}$ where C_1 and C_2 depends on L_f, L_ρ, the Lipschtiz constants of the class of approximation and an upper bound on ρ (the proof of this result can be found in [13]). The explicit dependence of this lower bound on $\alpha^*_{\mathcal{F}_n}$ as well as the large values of C_1 and C_2 tend to lead to a very conservative lower bound, especially when $\mathcal{F}_n$ is sparse.

9 Conclusions

In this paper, we have considered min max based approaches for addressing the generalization problem in RL. In particular, we have proposed and studied an algorithm that outputs a policy that maximizes a lower bound on the worst return that may be obtained with an environment compatible with some observed system transitions. The proposed algorithm is of polynomial complexity and avoids regions of the state space where the sample density is too low according to the prior information. A simple example has illustrated that this strategy can lead to cautious policies where other batch-mode RL algorithms fail because they unsafely generalize the information contained in the dataset.

From the results given in [11], it is also possible to derive in a similar way tight upper bounds on the return of a policy. In this respect, it would also be possible to adopt a “max max” generalization strategy by inferring policies that maximize these tight upper bounds. We believe that exploiting together the policy based on a min max generalization strategy and the one based on a max max generalization strategy could offer interesting possibilities for addressing the exploitation-exploration tradeoff faced when designing intelligent agents. For example, if the policies coincide, it could be an indication that further exploration is not needed.

When using batch mode reinforcement learning algorithms to design autonomous intelligent agents, a problem arises. After a long enough time of interaction with their environment, the sample the agents collect may become so large that batch mode RL-techniques may become computationally impractical, even with small degree polynomial algorithms. As suggested by [8], a solution for addressing this problem would be to retain only the most “informative samples”. In the context of the proposed algorithm, the complexity for computing the optimal sequence of decisions is quadratic in the size of the dataset. We believe that it would be interesting to design lower complexity algorithms based on subsampling the dataset based on the initial state information.

The work reported in this paper has been carried out in the particular context of deterministic Lipschtiz continuous environments. We believe that extending this work to environments which satisfy other types of properties (for instance, Hölder continuity assumptions or properties that are not related with continuity) or which are possibly also stochastic is a natural direction for further research.

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 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. We also acknowledge financial support from NIH grants P50 DA10075 and R01 MH080015. We also thank Bertrand Cornélusse and Quentin Louveaux for their insightful comments on the min max formulation. The scientific responsibility rests with its authors.

References

1. Bemporad, A., Morari, M.: Robust model predictive control: A survey. Robustness in Identification and Control 245, 207–226 (1999)
2. Bertsekas, D., Tsitsiklis, J.: Neuro-Dynamic Programming. Athena Scientific (1996)
3. Boyan, J., Moore, A.: Generalization in reinforcement learning: Safely approximating the value function. In: Advances in Neural Information Processing Systems 7 (NIPS 1995). pp. 369–376. MIT Press, Denver, CO, USA (1995)
4. Chakratovorty, S., Hyland, D.: Minimax reinforcement learning. In: Proceedings of AIAA Guidance, Navigation, and Control Conference and Exhibit. San Francisco, CA, USA (2003)
5. Csáji, B.C., Monostori, L.: Value function based reinforcement learning in changing Markovian environments. Journal of Machine Learning Research 9, 1679–1709 (2008)
6. De Berg, M., Cheong, O., Van Kreveld, M., Overmars, M.: Computational Geometry: Algorithms and Applications. Springer-Verlag (2008)
7. Delage, E., Mannor, S.: Percentile optimization for Markov decision processes with parameter uncertainty. Operations Research (2006)
8. Ernst, D.: 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). Singapore (2005)
9. Ernst, D., Geurts, P., Wehenkel, L.: Tree-based batch mode reinforcement learning. Journal of Machine Learning Research 6, 503–556 (2005)
10. Ernst, D., Glavic, M., Capitanescu, F., Wehenkel, L.: Reinforcement learning versus model predictive control: a comparison on a power system problem. IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics 39, 517–529 (2009)
11. Fonteneau, R., Murphy, S., Wehenkel, L., Ernst, D.: Inferring bounds on the performance of a control policy from a sample of trajectories. In: Proceedings of the 2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (IEEE ADPRL 09). Nashville, TN, USA (2009)
12. Fonteneau, R., Murphy, S., Wehenkel, L., Ernst, D.: A cautious approach to generalization in reinforcement learning. In: Proceedings of the Second International Conference on Agents and Artificial Intelligence (ICAART 2010). Valencia, Spain (2010)
13. Fonteneau, R., Murphy, S.A., Wehenkel, L., Ernst, D.: Computing bounds for kernel-based policy evaluation in reinforcement learning. Tech. rep., Arxiv (2010)
14. Gordon, G.: Approximate Solutions to Markov Decision Processes. Ph.D. thesis, Carnegie Mellon University (1999)
15. Ingersoll, J.: Theory of Financial Decision Making. Rowman and Littlefield Publishers, Inc. (1987)

16. Lagoudakis, M., Parr, R.: Least-squares policy iteration. Jounal of Machine Learning Research 4, 1107–1149 (2003)
17. Littman, M.L.: Markov games as a framework for multi-agent reinforcement learning. In: Proceedings of the Eleventh International Conference on Machine Learning (ICML 1994). New Brunswick, NJ, USA (1994)
18. Mannor, S., Simester, D., Sun, P., Tsitsiklis, J.: Bias and variance in value function estimation. In: Proceedings of the Twenty-first International Conference on Machine Learning (ICML 2004). Banff, Alberta, Canada (2004)
19. Murphy, S.: Optimal dynamic treatment regimes. Journal of the Royal Statistical Society, Series B 65(2), 331–366 (2003)
20. Murphy, S.: An experimental design for the development of adaptive treatment strategies. Statistics in Medicine 24, 1455–1481 (2005)
21. Ormoneit, D., Sen, S.: Kernel-based reinforcement learning. Machine Learning 49(2-3), 161–178 (2002)
22. Qian, M., Murphy, S.: Performance guarantees for individualized treatment rules. Tech. Rep. 498, Department of Statistics, University of Michigan (2009)
23. Riedmiller, M.: Neural fitted Q iteration - first experiences with a data efficient neural reinforcement learning method. In: Proceedings of the Sixteenth European Conference on Machine Learning (ECML 2005). pp. 317–328. Porto, Portugal (2005)
24. Rovatous, M., Lagoudakis, M.: Minimax search and reinforcement learning for adversarial tetris. In: Proceedings of the 6th Hellenic Conference on Artificial Intelligence (SETN'10). Athens, Greece (2010)
25. Rudin, W.: Real and Complex Analysis. McGraw-Hill (1987)
26. Sutton, R.: Generalization in reinforcement learning: Successful examples using sparse coding. In: Advances in Neural Information Processing Systems 8 (NIPS 1996). pp. 1038–1044. MIT Press, Denver, CO, USA (1996)
27. Sutton, R., Barto, A.: Reinforcement Learning. MIT Press (1998)
28. Viterbi, A.: Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory 13(2), 260– 269 (1967)

A Proof of Theorem 1

- Let us first prove that $L^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x) \leq K^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x)$. Let us assume that we know a set of variables $\hat{x}_0, \ldots, \hat{x}_{T-1}$ and $\hat{r}_0, \ldots, \hat{r}_{T-1}$ that are solution of the optimization problem. To each action $u \in \mathcal{U}$, we associate the sets $A_u = \{x^l \in \{x^1, \ldots, x^n\} | u^l = u\}$ and $B_u = \{\hat{x}_t \in \{\hat{x}_0, \ldots, \hat{x}_{T-1}\} | u_t = u\}$. Let $S_u = A_u \cup B_u$. For simplicity in the proof, we assume that the points of S_u are in *general position*, i.e., no $(d_{\mathcal{X}} + 1)$ points from S_u lie in a $(d_{\mathcal{X}}-1)-$dimensional plane (the points are affinely independent). This allows to compute a $d_{\mathcal{X}}-$dimensional triangulation $\{\Delta^1, \ldots, \Delta^p\}$ of the convex hull $H(S_u)$ defined by the set of points S_u [6]. We introduce for every value of $u \in \mathcal{U}$ two Lipschitz continuous functions $\tilde{f}_u : \mathcal{X} \to \mathcal{X}$ and $\tilde{\rho}_u : \mathcal{X} \to \mathbb{R}$ defined as follows:
  - *Inside the convex hull $H(S_u)$*
    Let $g^f_u : S_u \to \mathcal{X}$ and $g^\rho_u : S_u \to \mathbb{R}$ be such that:

$$\forall x^l \in A_u \ , \begin{cases} g^f_u(x^l) = f(x^l, u) \\ g^\rho_u(x^l) = \rho(x^l, u) \end{cases} \text{ and } \forall \hat{x}_t \in B_u \backslash A_u \ , \begin{cases} g^f_u(\hat{x}_t) = \hat{x}_{t+1} \\ g^\rho_u(\hat{x}_t) = \hat{r}_t \end{cases} .$$

Then, we define the functions $\tilde{f}_u$ and $\tilde{\rho}_u$ inside $H(S_u)$ as follows:

$$\forall k \in \{1, \ldots, p\}, \forall x' \in \varDelta^k \ , \tilde{f}_u(x') = \sum_{i=1}^{d_{\mathcal{X}}+1} \lambda_i^k(x') g_u^f(s_i^k) \ ,$$
$$\tilde{\rho}_u(x') = \sum_{i=1}^{d_{\mathcal{X}}+1} \lambda_i^k(x') g_u^\rho(s_i^k) \ ,$$

where $s_i^k \quad i = 1 \ldots (d_{\mathcal{X}} + 1)$ are the vertices of $\varDelta^k$ and $\lambda_i^k(x)$ are such that $x' = \sum_{i=1}^{d_{\mathcal{X}}+1} \lambda_i^k(x') s_i^k$ with $\sum_{i=1}^{d_{\mathcal{X}}+1} \lambda_i^k(x') = 1$ and $\lambda_i^k(x') \geq 0, \ \forall i$.

- *Outside the convex hull $H(S_u)$*
  According the Hilbert Projection Theorem [25], for every point $x'' \in \mathcal{X}$, there exists a unique point $y'' \in H(S_u)$ such that $\|x'' - y''\|_{\mathcal{X}}$ is minimized over $H(S_u)$. This defines a mapping $t_u : \mathcal{X} \to H(S_u)$ which is $1-$Lipschitzian. Using the mapping t_u, we define the functions $\tilde{f}_u$ and $\tilde{\rho}_u$ outside $H(S_u)$ as follows:

$$\forall x'' \in \mathcal{X} \backslash H(S_u), \tilde{f}_u(x'') = \tilde{f}_u(t_u(x'')) \text{ and } \tilde{\rho}_u(x'') = \tilde{\rho}_u(t_u(x'')) \ .$$

We finally introduce the functions $\tilde{f}$ and $\tilde{\rho}$ over the space $\mathcal{X} \times \mathcal{U}$ as follows:

$$\forall (x', u) \in \mathcal{X} \times \mathcal{U}, \tilde{f}(x', u) = \tilde{f}_u(x') \text{ and } \tilde{\rho}(x', u) = \tilde{\rho}_u(x') \ .$$

One can easily show that the pair $(\tilde{f}, \tilde{\rho})$ belongs to $\mathcal{L}^f_{\mathcal{F}_n} \times \mathcal{L}^\rho_{\mathcal{F}_n}$ and satisfies

$$J^{u_0, \ldots, u_{T-1}}_{(\tilde{f}, \tilde{\rho})}(x) = \sum_{t=0}^{T-1} \tilde{\rho}(\hat{x}_t, u_t) = \sum_{t=0}^{T-1} \hat{r}_t$$

with $\hat{x}_{t+1} = \tilde{f}(\hat{x}_t, u_t)$ and $\hat{x}_0 = x$. This proves that

$$L^{u_0, \ldots, u_{T-1}}_{\mathcal{F}_n}(x) \leq K^{u_0, \ldots, u_{T-1}}_{\mathcal{F}_n}(x) \ .$$

(Note that one could still build two functions $(\tilde{f}, \tilde{\rho}) \in \mathcal{L}^f_{\mathcal{F}_n} \times \mathcal{L}^\rho_{\mathcal{F}_n}$ even if the sets of points $(S_u)_{u \in \mathcal{U}}$ are not in general position)

– Then, let us prove that $K^{u_0, \ldots, u_{T-1}}_{\mathcal{F}_n}(x) \leq L^{u_0, \ldots, u_{T-1}}_{\mathcal{F}_n}(x)$. We consider the environment $(f^{u_0, \ldots, u_{T-1}}_{\mathcal{F}_n, x}, \rho^{u_0, \ldots, u_{T-1}}_{\mathcal{F}_n, x})$ introduced in Equation (1) at the end of Section 3. One has

$$L^{u_0, \ldots, u_{T-1}}_{\mathcal{F}_n}(x) = J^{u_0, \ldots, u_{T-1}}_{(f^{u_0, \ldots, u_{T-1}}_{\mathcal{F}_n, x}, \rho^{u_0, \ldots, u_{T-1}}_{\mathcal{F}_n, x})}(x) = \sum_{t=0}^{T-1} \tilde{r}_t \ ,$$

with $\forall t \in \{0, \ldots, T-1\}$,

$$\tilde{r}_t = \rho^{u_0, \ldots, u_{T-1}}_{\mathcal{F}_n, x}(\tilde{x}_t, u_t) \ , \tilde{x}_{t+1} = f^{u_0, \ldots, u_{T-1}}_{\mathcal{F}_n, x}(\tilde{x}_t, u_t) \ , \tilde{x}_0 = x \ .$$

The variables $\tilde{x}_0, \ldots, \tilde{x}_{T-1}$ and $\tilde{r}_0, \ldots, \tilde{r}_{T-1}$ satisfy the constraints introduced in Theorem (1). This proves that

$$K^{u_0, \ldots, u_{T-1}}_{\mathcal{F}_n}(x) \leq L^{u_0, \ldots, u_{T-1}}_{\mathcal{F}_n}(x)$$

and completes the proof.

B Proof of Lemma 1

Before proving Lemma 1, we prove a preliminary result related to the Lipschitz continuity of state-action value functions. Let $(f', \rho') \in \mathcal{L}^f_{\mathcal{F}_n} \times \mathcal{L}^\rho_{\mathcal{F}_n}$ be a compatible environment. For $N = 1, \ldots, T$, let us define the family of $(f', \rho')-$state-action value functions $Q^{u_0,\ldots,u_{T-1}}_{N,(f',\rho')} : \mathcal{X}' \times \mathcal{U} \to \mathbb{R}$ as follows:

$$Q^{u_0,\ldots,u_{T-1}}_{N,(f'\rho')}(x', u) = \rho'(x', u) + \sum_{t=T-N+1}^{T-1} \rho'(x'_t, u_t),$$

where $x'_{T-N+1} = f'(x', u)$ and $x'_{t+1} = f'(x'_t, u_t),\ \forall t \in \{T - N + 1, \ldots, T - 1\}$. $Q^{u_0,\ldots,u_{T-1}}_{N,(f',\rho')}(x', u)$ gives the sum of rewards from instant $t = T - N$ to instant $T - 1$ given the compatible environment (f', ρ') when (i) the system is in state $x' \in \mathcal{X}$ at instant $T - N$, (ii) the action chosen at instant $T - N$ is u and (iii) the actions chosen at instants $t > T - N$ are u_t. The value $J^{u_0,\ldots,u_{T-1}}_{(f',\rho')}(x)$ can be deduced from the value of $Q^{u_0,\ldots,u_{T-1}}_{T,(f',\rho')}(x, u_0)$ as follows:

$$\forall x \in \mathcal{X},\ J^{u_0,\ldots,u_{T-1}}_{(f',\rho')}(x) = Q^{u_0,\ldots,u_{T-1}}_{T,(f',\rho')}(x, u_0). \tag{3}$$

We also have $\forall x' \in \mathcal{X}, \forall u \in \mathcal{U}, \forall N \in \{1, \ldots, T - 1\}$

$$Q^{u_0,\ldots,u_{T-1}}_{N+1,(f',\rho')}(x', u) = \rho'(x', u) + Q^{u_0,\ldots,u_{T-1}}_{N,(f',\rho')}(f'(x', u), u_{T-N}) \tag{4}$$

Lemma 2 (Lipschitz continuity of $Q^{u_0,\ldots,u_{T-1}}_{N,(f',\rho')}$).
$\forall N \in \{1, \ldots, T\}$, $\forall (x', x'') \in \mathcal{X}^2, \forall u \in \mathcal{U}$,

$$|Q^{u_0,\ldots,u_{T-1}}_{N,(f',\rho')}(x', u) - Q^{u_0,\ldots,u_{T-1}}_{N,(f',\rho')}(x'', u)| \leq L_{Q_N} \|x' - x''\|_{\mathcal{X}} \ ,$$

with $L_{Q_N} = L_\rho \sum_{i=0}^{N-1} (L_f)^i$.

Proof. We consider the statement $\mathcal{H}(N)$: $\forall (x', x'') \in \mathcal{X}^2, \forall u \in \mathcal{U}$,

$$|Q^{u_0,\ldots,u_{T-1}}_{N,(f',\rho')}(x', u) - Q^{u_0,\ldots,u_{T-1}}_{N,(f',\rho')}(x'', u)| \leq L_{Q_N} \|x' - x''\|_{\mathcal{X}}.$$

We prove by induction that $\mathcal{H}(N)$ is true $\forall N \in \{1, \ldots, T\}$. For the sake of clarity, we denote $|Q^{u_0,\ldots,u_{T-1}}_{N,(f',\rho')}(x', u) - Q^{u_0,\ldots,u_{T-1}}_{N,(f',\rho')}(x'', u)|$ by Δ_N.

- *Basis ($N = 1$) :* We have $\Delta_N = |\rho'(x', u) - \rho'(x'', u)|$, and since $\rho' \in \mathcal{L}^\rho_{\mathcal{F}_n}$, we can write $\Delta_N \leq L_\rho \|x' - x''\|_{\mathcal{X}}$. This proves $\mathcal{H}(1)$.
- *Induction step:* We suppose that $\mathcal{H}(N)$ is true, $1 \leq N \leq T - 1$. Using equation (4), we can write $\Delta_{N+1} = \left|Q^{u_0,\ldots,u_{T-1}}_{N+1,(f',\rho')}(x', u) - Q^{u_0,\ldots,u_{T-1}}_{N+1,(f',\rho')}(x'', u)\right|$ $= \left|\rho'(x', u) - \rho'(x'', u) + Q^{u_0,\ldots,u_{T-1}}_{N,(f',\rho')}(f'(x', u), u_{T-N}) - Q^{u_0,\ldots,u_{T-1}}_{N,(f',\rho')}(f'(x'', u), u_{T-N})\right|$ and, from there, $\Delta_{N+1} \leq \left|\rho'(x', u) - \rho'(x'', u)\right| + \left|Q^{u_0,\ldots,u_{T-1}}_{N,(f',\rho')}(f'(x', u), u_{T-N}) - Q^{u_0,\ldots,u_{T-1}}_{N,(f',\rho')}(f'(x'', u), u_{T-N})\right|$. $\mathcal{H}(N)$ and the Lipschitz continuity of ρ' give

  $$\Delta_{N+1} \leq L_\rho \|x' - x''\|_{\mathcal{X}} + L_{Q_N} \|f'(x', u) - f'(x'', u)\|_{\mathcal{X}}.$$

  Since $f' \in \mathcal{L}^f_{\mathcal{F}_n}$, the Lipschitz continuity of f' gives $\Delta_{N+1} \leq L_\rho \|x' - x''\|_{\mathcal{X}} + L_{Q_N} L_f \|x' - x''\|_{\mathcal{X}}$, then $\Delta_{N+1} \leq L_{Q_{N+1}} \|x' - x''\|_{\mathcal{X}}$ since $L_{Q_{N+1}} = L_\rho + L_{Q_N} L_f$. This proves $\mathcal{H}(N + 1)$ and ends the proof.

Proof of Lemma 1

- The inequality $L^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x) \leq J^{u_0,\ldots,u_{T-1}}(x)$ is trivial since (f,ρ) belongs to $\mathcal{L}^f_{\mathcal{F}_n} \times \mathcal{L}^{\rho}_{\mathcal{F}_n}$.
- Let $(f',\rho') \in \mathcal{L}^f_{\mathcal{F}_n} \times \mathcal{L}^{\rho}_{\mathcal{F}_n}$ be a compatible environment. By assumption we have $u^{l_0} = u_0$, then we use equation (3) and the Lipschitz continuity of $Q^{u_0,\ldots,u_{T-1}}_{T,(f',\rho')}$ to write

  $$|J^{u_0,\ldots,u_{T-1}}_{(f',\rho')}(x) - Q^{u_0,\ldots,u_{T-1}}_{T,(f',\rho')}(x^{l_0},u_0)| \leq L_{Q_T}\|x - x^{l_0}\|_{\mathcal{X}}.$$

  It follows that $Q^{u_0,\ldots,u_{T-1}}_{T,(f',\rho')}(x^{l_0},u_0) - L_{Q_T}\|x - x^{l_0}\|_{\mathcal{X}} \leq J^{u_0,\ldots,u_{T-1}}_{(f',\rho')}(x)$. According to equation (4), we have $Q^{u_0,\ldots,u_{T-1}}_{T,(f',\rho')}(x^{l_0},u_0) = \rho'(x^{l_0},u_0) + Q^{u_0,\ldots,u_{T-1}}_{T-1,(f'\rho')}(f'(x^{l_0},u_0),u_1)$ and from there

  $$Q^{u_0,\ldots,u_{T-1}}_{T,(f',\rho')}(x^{l_0},u_0) = r^{l_0} + Q^h_{T-1,(f',\rho')}(y^{l_0},u_1).$$

  Thus, $Q^{u_0,\ldots,u_{T-1}}_{T-1,(f',\rho')}(y^{l_0},u_1) + r^{l_0} - L_{Q_T}\|x - x^{l_0}\|_{\mathcal{X}} \leq J^{u_0,\ldots,u_{T-1}}_{(f',\rho')}(x)$. The Lipschitz continuity of $Q^{u_0,\ldots,u_{T-1}}_{T-1,(f',\rho')}$ with $u_1 = u^{l_1}$ gives

  $$|Q^{u_0,\ldots,u_{T-1}}_{T-1,(f',\rho')}(y^{l_0},u_1) - Q^{u_0,\ldots,u_{T-1}}_{T-1,(f',\rho')}(x^{l_1},u^{l_1})| \leq L_{Q_{T-1}}\|y^{l_0} - x^{l_1}\|_{\mathcal{X}} \ .$$

  This implies that

  $$Q^{u_0,\ldots,u_{T-1}}_{T-1,(f',\rho')}(x^{l_1},u_1) - L_{Q_{T-1}}\|y^{l_0} - x^{l_1}\|_{\mathcal{X}} \leq Q^{u_0,\ldots,u_{T-1}}_{T-1,(f',\rho')}(y^{l_0},u_1).$$

  We have therefore

  $$\begin{aligned}&Q^{u_0,\ldots,u_{T-1}}_{T-1,(f',\rho')}(x^{l_1},u_1) + r^{l_0} - L_{Q_T}\|x - x^{l_0}\|_{\mathcal{X}} - L_{Q_{T-1}}\|y^{l_0} - x^{l_1}\|_{\mathcal{X}}\\ &\leq J^{u_0,\ldots,u_{T-1}}_{(f',\rho')}(x).\end{aligned}$$

  By developing this iteration, we obtain

  $$J^{u_0,\ldots,u_{T-1}}_{(f',\rho')}(x) \geq \sum_{t=0}^{T-1}\left[r^{l_t} - L_{Q_{T-t}}\|y^{l_{t-1}} - x^{l_t}\|_{\mathcal{X}}\right] \ . \tag{5}$$

  The right side of Equation (5) does not depend on the choice of $(f',\rho') \in \mathcal{L}^f_{\mathcal{F}_n} \times \mathcal{L}^{\rho}_{\mathcal{F}_n}$; Equation (5) is thus true for $(f',\rho') = (f^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n,x}, \rho^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n,x})$ (cf. Equation (1) in Section 3). This finally gives

  $$L^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x) \geq \sum_{t=0}^{T-1}\left[r^{l_t} - L_{Q_{T-t}}\|y^{l_{t-1}} - x^{l_t}\|_{\mathcal{X}}\right]$$

  since $L^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n}(x) = J^{u_0,\ldots,u_{T-1}}_{(f^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n,x}, \rho^{u_0,\ldots,u_{T-1}}_{\mathcal{F}_n,x})}(x)$.