University of Liège – University of Michigan

A Cautious Approach to Generalization in Reinforcement Learning

Raphaël Fonteneau, Susan Murphy, Louis Wehenkel, Damien Ernst

January, 22nd 2010

Valencia

Outline

Introduction

Problem statement

Contributions

Illustration

Discussion

Conclusions and future work

Introduction

- *Reinforcement Learning* (*RL*) *algorithms are challenged when dealing with large or continuous spaces*
- *In those cases, the dominating approach is to use function approximators*
- *This, in turn, leads to low performance guarantees when spaces are poorly covered by the sample*
- *We propose an algorithm which exploits weak prior knowledge about its environment for computing a sequence of actions which tend to avoid regions where performance is uncertain.*

Problem statement

- *We consider a discrete-time system whose dynamics over T stages is given by* $x_{t+1} = f(x_t, u_t)$
- x_t *lies in a normed state space* X, u_t *lies in a finite action space* U
- *An instantaneous reward* $r_t = \rho(x_t, u_t)$ *is associated with the action* u_t *while being in state* x_t
- *The performance of a given sequence of actions* $(u_0, \ldots, u_{T-1})$ *when starting from an initial state* $x_0 = x$ *(also called* T*-stage return) is given by*

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

Problem statement

- *We define* $J^{\star}(x) = \max_{(u_0,\ldots,u_{T-1}) \in U^T} J^{u_0,\ldots,u_{T-1}}(x)$

- *The goal is to find a sequence of actions* $\hat{u}_0(x),\ldots,\hat{u}_{T-1}(x)$ *such that* $J^{\hat{u}_0(x),\ldots,\hat{u}_{T-1}(x)}(x)$ *is as close as possible to* $J^{\star}(x)$.

Problem statement

- *The system dynamics f and reward function ρ are **unknown**, replaced by a set of n one-step system transitions*

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

that all satisfy $r^l = \rho(x^l, u^l)$ *and* $y^l = f(x^l, u^l)$

- *Weak prior knowledge: we know two constants* L_f *and* L_ρ *such that*

$$L_f, L_\rho > 0, \forall (x, x') \in X^2, u \in U,$$
$$\| f(x,u) - f(x',u) \| \leq L_f \| x - x' \|,$$
$$| \rho(x,u) - \rho(x',u) | \leq L_\rho \| x - x' \|.$$

Contributions

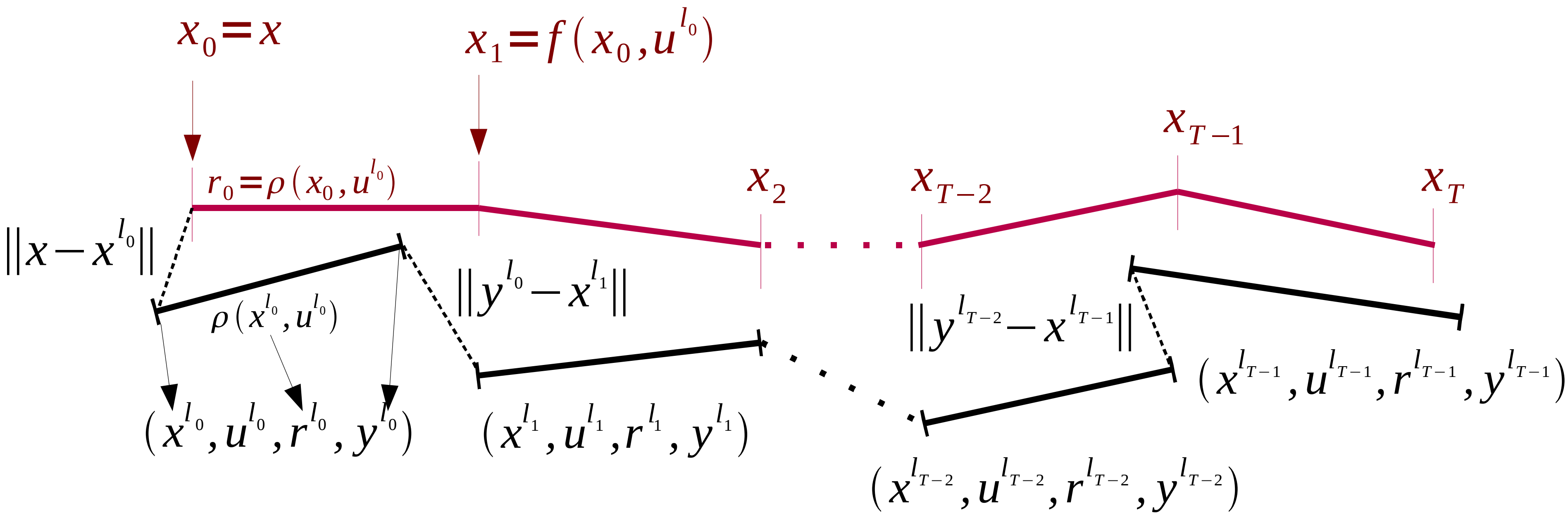
- *Computation of a lower bound on the return of a given sequence of actions*

- *Computation of a sequence of action that maximizes this lower bound : the CGRL algorithm* (*Cautious approach to Generalization in RL*)

- *Consistency properties.*

Contributions

- *Computing a lower bound from a sequence of one-step system transitions*



Contributions

Lemma *Let* $\tau = [(x^{l_t}, u^{l_t}, r^{l_t}, y^{l_t})]_{t=0}^{T-1}$ *be a sequence of one-step system transitions.*

Then,

$$J^{u^{l_0},\ldots,u^{l_{T-1}}}(x) \geq B(\tau, x),$$
$$B(\tau, x) = \sum_{t=0}^{T-1} [r^{l_t} - L_{Q_{T-t}} \| x^{l_t} - y^{l_{t-1}} \|]$$

with

$$y^{l_{-1}} = x,$$
$$L_{Q_{T-t}} = L_\rho \sum_{i=0}^{T-t-1} [L_f]^i.$$

Contributions

- *Given a sequence of actions* $(u_0, \ldots, u_{T-1})$*, we denote by* $F^T_{u_0,\ldots,u_{T-1}}$ *the set of all sequence of one-step system transitions* $\tau = [(x^{l_t}, u^{l_t}, r^{l_t}, y^{l_t})]_{t=0}^{T-1}$ *that satisfy the condition*

$$\forall t \in \{0, \ldots, T-1\}, u^{l_t} = u_t$$

- *Among those sequences, one can determine a sequence that leads to the highest lower bound, denoted by*

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

- *The tightness of* $B^{u_0,\ldots,u_{T-1}}(x)$ *can be expressed as a function of the sparsity of the sample of one-step transitions.*

Contributions

Definition *Given an action* a *, let* F_a *be the set of all one-step transitions* (x^l, u^l, r^l, y^l) *such that* $u^l = a$ *. Let us assume that all* F_a *are non-empty, and let us suppose that there exists* $\alpha > 0$ *such that*

$$\forall a \in U, \sup_{x' \in X} \left\{ \min_{(x^l, u^l, r^l, y^l) \in F_a} \| x' - x^l \| \right\} \leq \alpha$$

The smallest α *which satisfies the previous condition is named the sample sparsity and is denoted by* α^*

''The sparsity can be seen as the radius of the largest non-visited state space area''.

Contributions

Theorem

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

The lower bound $B^{u_0, \ldots, u_{T-1}}(x)$ *thus converges to the* T*-stage return of the sequence of actions* $(u_0, \ldots, u_{T-1})$ *when the sample sparsity decreases to zero*

- *We thus propose to use* $B^{u_0, \ldots, u_{T-1}}(x)$ *as an inference criterion, and we propose an algorithm that computes a sequence that leads to the maximization of* $B^{u_0, \ldots, u_{T-1}}(x)$.

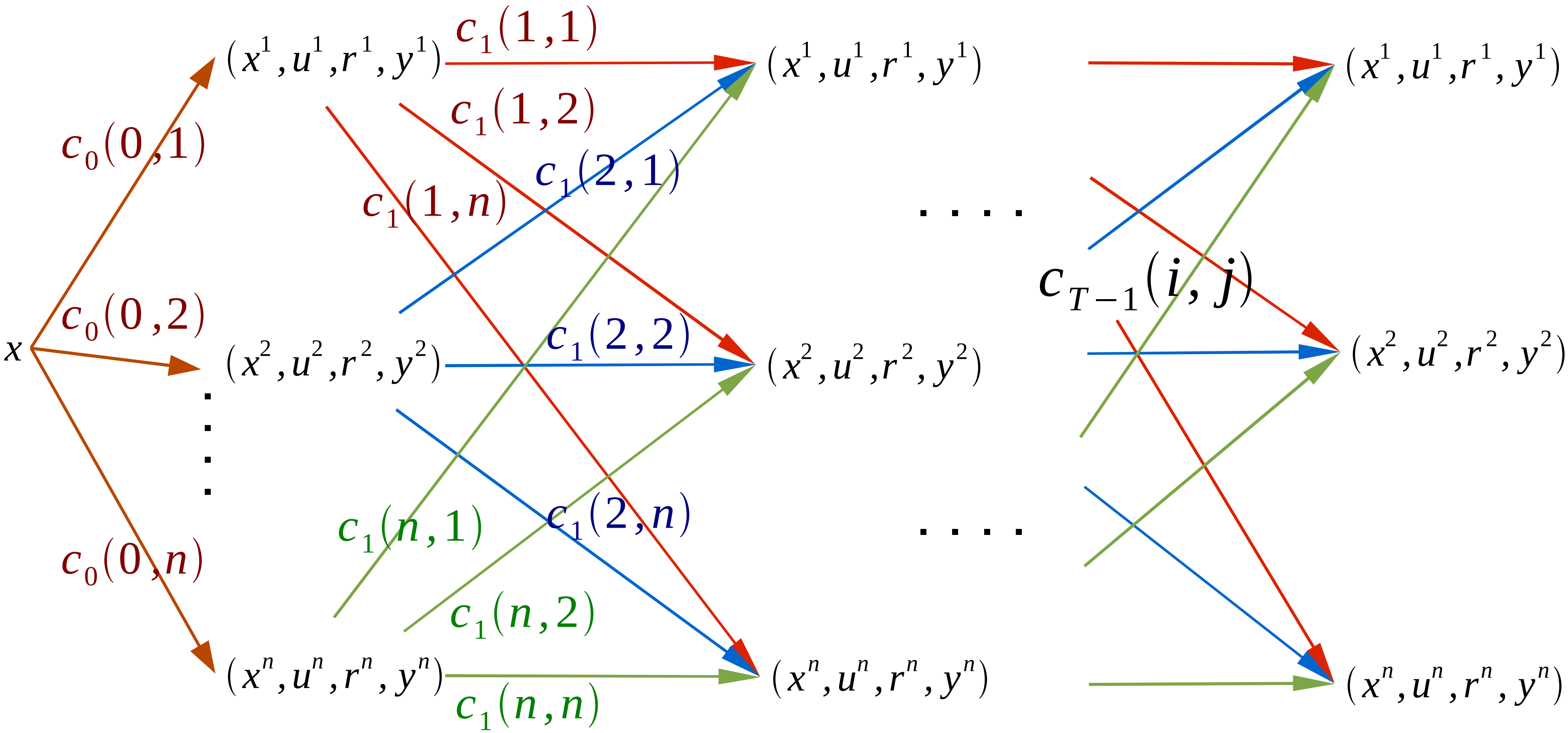
Contributions

- *We introduce the CGRL algorithm that computes, given an initial state* x *, a sequence of actions* $\hat{u}_0(x), \ldots, \hat{u}_{T-1}(x)$ *that belongs to the set*

$$\mathbf{B}_x^\star = \underset{u_0,\ldots,u_{T-1} \in U^T}{argmax} \; B^{u_0,\ldots,u_{T-1}}(x)$$

- *To identify such a sequence without computing for all sequences* $(u_0, \ldots, u_{T-1})$ *the value* $B^{u_0,\ldots,u_{T-1}}(x)$ *, the CGRL algorithm reformulates the problem of finding an element of* $\mathbf{B}_x^\star$ *into a shortest path problem*

- *The complexity is* $O(T\, n^2)$ *.*



$$l_0^\star,\dots,l_{T-1}^\star \in \underset{l_0,\dots,l_{T-1}}{argmax} \;\; c_0(0,l_0)+c_1(l_0,l_1)+\dots+c_{T-1}(l_{T-2},l_{T-1})$$

$$with \;\; c_t(i,j) = -L_{Q_{T-t}}\left\|y^i - x^j\right\| + r^j,\; y^0 = x \longrightarrow \hat{u}_0^\star(x),\dots,\hat{u}_{T-1}^\star(x) = u^{l_0^\star},\dots,u^{l_{T-1}^\star}$$

Contributions

Theorem [Consistency of the CGRL algorithm]

Let $\boldsymbol{J}_x^\star = \underset{(u_0,\ldots,u_{T-1})\in U^T}{\operatorname{argmax}} J^{u_0,\ldots,u_{T-1}}(x)$.

Let us suppose that $\boldsymbol{J}_x^\star \neq U^T$ *(otherwise, the problem is trivial).*

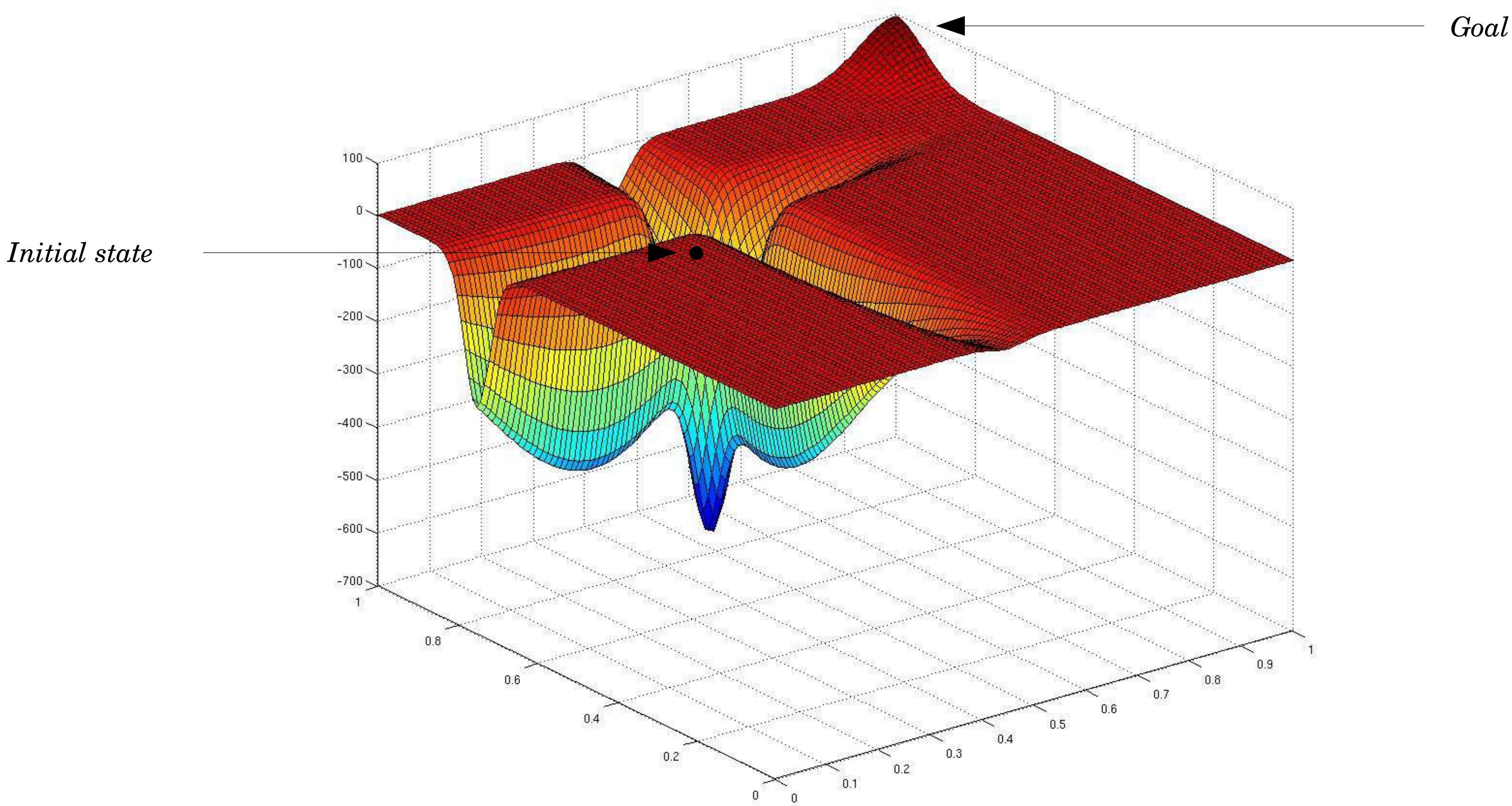
We define $\epsilon(x) = \underset{(u_0,\ldots,u_{T-1})\in U^T \setminus \boldsymbol{J}_x^\star}{\min} J^\star(x) - J^{u_0,\ldots,u_{T-1}}(x)$.

Then,

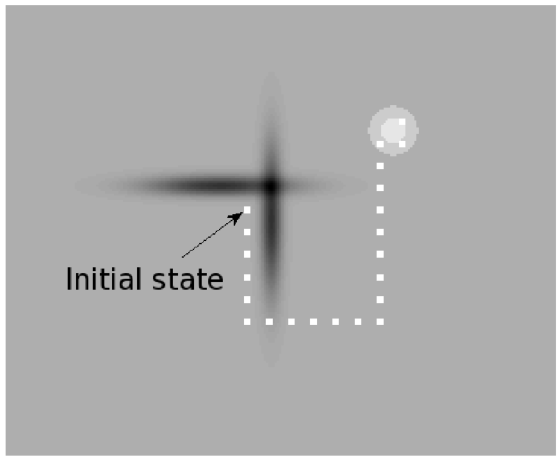
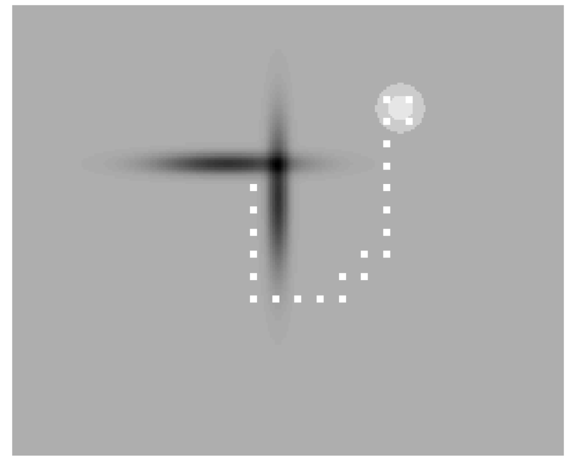
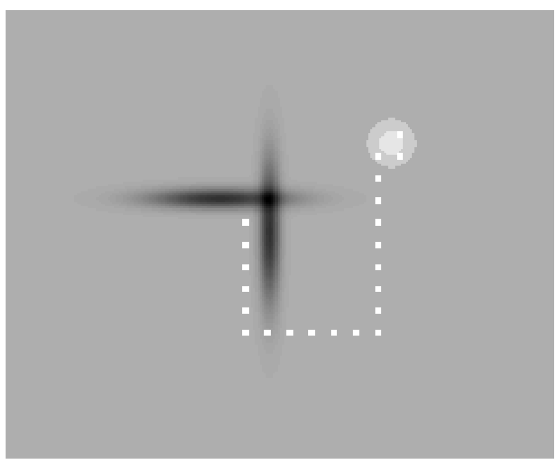
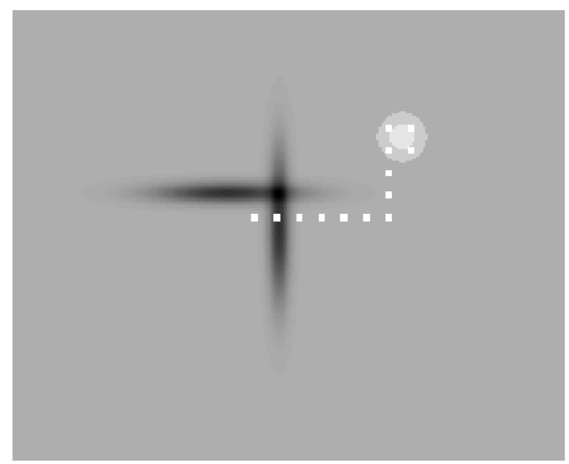
$$C\alpha^\star < \epsilon(x) \Rightarrow (\hat{u}_0(x),\ldots,\hat{u}_{T-1}(x)) \in \boldsymbol{J}_x^\star.$$

Illustration

- *The puddle world benchmark*



Illustration

	CGRL	*FQI (Fitted Q Iteration)*
The state space is uniformly covered by the sample		
Information about the Puddle area is removed		

Discussion

- *The CGRL algorithm outputs a sequence of actions and a lower bound on its return*

- *The tightness of the lower bound depends on the sparsity, but not explicitly*

- *One can obtain good performance guarantees even if the state space is not well covered everywhere.*

Conclusions and future work

- *We have proposed a new strategy for RL using a batch of one-step system transitions. The proposed CGRL algorithm is polynomial complexity and avoids regions of the state space where the sample sparsity is too big according to prior information*

- *Illustrations show that this strategy can lead to cautious policies when other RL algorithms fail because of unsafely generalization*

- *One could similarly compute upper bounds, and derive an ''optimistic'' generalization RL algorithm which could be combined with CGRL in order to address the exploration/exploitation task*

- *Closed-loop strategies* (*considering a receding horizon*) *could be used in a stochastic framework.*

Appendix

- *Another illustration : enhancement of an HIV infected patient's treatment (simulated data)*
- *Model:*

$$\frac{dT_1}{dt} = \lambda_1 - d_1 T_1 - (1-\epsilon_1) k_1 V T_1$$

$$\frac{dT_2}{dt} = \lambda_2 - d_2 T_2 - (1-f\epsilon_1) k_2 V T_2$$

$$\frac{dT_1^*}{dt} = (1-\epsilon_1) k_1 V T_1 - \delta T_1^* - m_1 E T_1^*$$

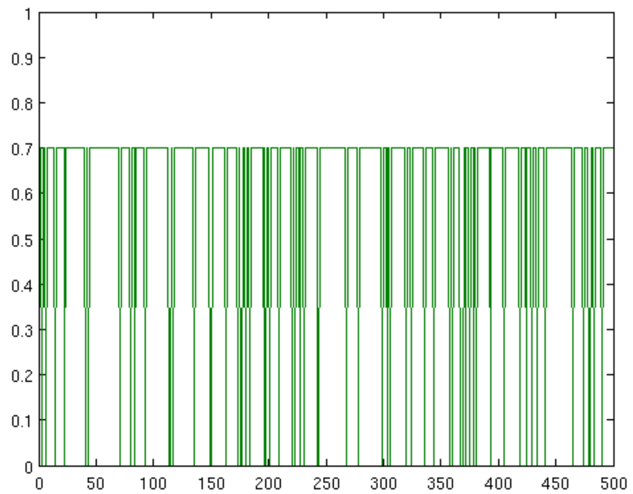
$$\frac{dT_2^*}{dt} = (1-f\epsilon_1) k_2 V T_2 - \delta T_2^* - m_2 E T_2^*$$

$$\frac{dV}{dt} = (1-\epsilon_2) N_T \delta (T_1^* + T_2^*) - cV - \left[(1-\epsilon_1)\rho_1 k_1 T_1 + (1-f\epsilon_1)\rho_2 k_2 T_2\right] V$$

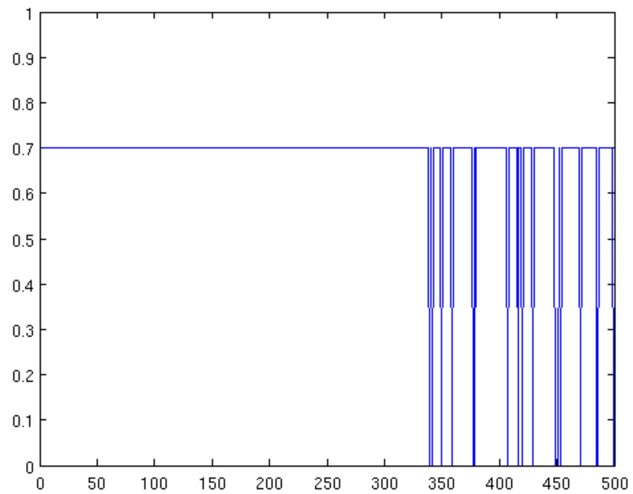
$$\frac{dE}{dT} = \lambda_E + \frac{b_E (T_1^* + T_2^*)}{T_1^* + T_2^* + K_b} E - \frac{d_E (T_1^* + T_2^*)}{T_1^* + T_2^* + K_d} E - \delta_E E$$

Appendix

- *A patient does not take his antiretroviral therapy in average once every eight days*
- *CGRL is run on the trajectory generated by this patient*



CGRL



Patient's treatment

CGRL suggestion