# Dynamic Treatment Regimes using Reinforcement Learning: A Cautious Generalization Approach

Raphael Fonteneau* Susan Murphy† Louis Wehenkel* Damien Ernst*

*Dept. of Electrical Engineering and Computer Science, University of Liège, Belgium
†Dept. of Statistics, University of Michigan, USA

## ABSTRACT

The treatment of chronic-like illnesses such has HIV infection, cancer or chronic depression implies long-lasting treatments that can be associated with low quality outcome, painful side effects and expensive costs. To enhance these treatments, clinicians often adopt what we call Dynamic Treatment Regimes (DTRs). DTRs are sets of sequential decision rules defining what actions should be taken at a specific instant to treat a patient based on information observed up to that instant. Since a few years, a growing research community is working on the development of formal methods (mainly issued from mathematics, statistics and control theory) that allow to infer from clinical data high-quality DTRs. We propose in this framework a consistent algorithm of quadratic complexity [3] that infer from clinical data a sequence of treatment actions by maximizing a recently proposed lower bound on the return depending on the initial state [2]. The algorithm (called CGRL for Cautious Generalization for Reinforcement Learning) has cautious generalization properties, i.e. it avoids taking treatment actions for which the sample of clinical data is too sparse to make safe generalization.

## 1 PROBLEM STATEMENT

- Discrete-time system dynamics over $T$ stages

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

where for all $t$, the state $x_t$ is an element of the normed vector state space $\mathcal{X}$ and $u_t$ is an element of the finite (discrete) action space $\mathcal{U}$,

- 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 $x_t$,

- The system dynamics $f$ and the reward function $\rho$ are unknown,
- The system dynamics $f$ and the reward function $\rho$ are Lipschitz continuous, i.e. that there exist finite constants $L_f, L_\rho \in \mathbb{R}$ such that: $\forall x, x' \in \mathcal{X}, \forall u \in \mathcal{U}$,

$$\|f(x,u) - f(x',u)\| \leq L_f \|x - x'\| \ ,$$
$$|\rho(x,u) - \rho(x',u)| \leq L_\rho \|x - x'\| \ ,$$

- Two constants $L_f$ and $L_\rho$ satisfying the above-written inequalities are known,
- Data : a set of one-step transitions

$$\mathcal{F} = \{(x^l, u^l, r^l, y^l)\}_{l=1}^{|\mathcal{F}|}$$

where each one-step transition is such that $y^l = f(x^l, u^l)$ and $r^l = \rho(x^l, u^l)$,

- Each action $a \in \mathcal{U}$ appears at least once in $\mathcal{F}$:

$$\forall a \in \mathcal{U}, \ \exists (x,u,r,y) \in \mathcal{F} \ : u = a \ ,$$

- For every initial state $x$, the return over $T$ stages of a sequence of actions $(u_0, \ldots, u_{T-1}) \in \mathcal{U}^T$ is defined as

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

## 2 OBJECTIVE

- An optimal sequence of actions $u_0^*(x), \ldots, u_{T-1}^*(x)$ is such that

$$J^{u_0^*(x),\ldots,u_{T-1}^*(x)}(x) = J^*(x) \doteq \max_{(u_0,\ldots,u_{T-1}) \in \mathcal{U}^T} J^{u_0,\ldots,u_{T-1}}(x) \ .$$

- The goal is to compute, for any initial state $x \in \mathcal{X}$, a sequence of actions $(\hat{u}_0^*(x), \ldots, \hat{u}_{T-1}^*(x)) \in \mathcal{U}^T$ such that $J^{\hat{u}_0^*(x),\ldots,\hat{u}_{T-1}^*(x)}$ is as close as possible to $J^*(x)$.

## 3 LOWER BOUND ON THE RETURN OF A GIVEN SEQUENCE ACTIONS

**Lemma 3.1** *Let $u_0, \ldots, u_{T-1}$ be a sequence of actions. Let $\tau = [(x^{l_t}, u^{l_t}, r^{l_t}, y^{l_t})]_{t=0}^{T-1} \in \mathcal{F}^T_{u_0,\ldots,u_{T-1}}$ where $\mathcal{F}^T_{u_0,\ldots,u_{T-1}}$ is the set of all sequences of one-step system transitions $[(x^{l_0}, u^{l_0}, r^{l_0}, y^{l_0}), \ldots, (x^{l_{T-1}}, u^{l_{T-1}}, r^{l_{T-1}}, y^{l_{T-1}})]$ for which $u^{l_t} = u_t, \ \forall t \in [\![0, T-1]\!]$. Then,*

$$J^{u_0,\ldots,u_{T-1}}(x) \geq B(\tau, x) \ ,$$

*with*

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

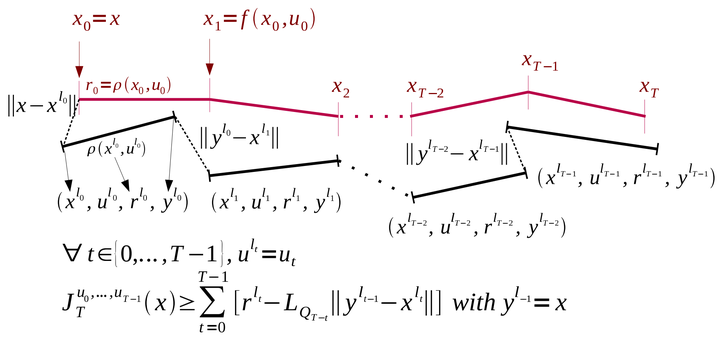


FIG. 1: A graphical interpretation of the different terms composing the bound on $J^{u_0,\ldots,u_{T-1}}(x)$ computed from a sequence of one-step transitions.

**Definition 3.2 (Highest lower bound for $u_0, \ldots, u_{T-1}$)**

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

**Definition 3.3 (Sample sparsity of $\mathcal{F}$)** *For $\mathcal{X}$ bounded, let $\mathcal{F}_a = \{(x^l, u^l, r^l, y^l) \in \mathcal{F} | u^l = a\}$. $\exists \ \alpha \in \mathbb{R}^+$ :*

$$\forall a \in \mathcal{U} \ , \sup_{x' \in X} \left\{ \min_{(x^l,u^l,r^l,y^l) \in \mathcal{F}_a} \|x^l - x'\| \right\} \leq \alpha \ . \qquad (1)$$

*The smallest $\alpha$ which satisfies equation (1) is named the sample sparsity and is denoted by $\alpha^*$.*

**Theorem 3.4 (Tightness of highest lower bound)** $\exists \ C > 0 : \ \forall (u_0, \ldots, u_{T-1}) \in \mathcal{U}^T$,

$$J^{u_0,\ldots,u_{T-1}}(x) - B^{u_0,\ldots,u_{T-1}}(x) \leq C\alpha^* .$$

## 4 THE CGRL ALGORITHM

- The CGRL algorithm computes for each initial state $x$ a sequence of actions $\hat{u}_0^*(x), \ldots, \hat{u}_{T-1}^*(x)$ that belongs to $\mathfrak{B}^*(x)$ where

$$\mathfrak{B}^*(x) = \{(u_0, \ldots, u_{T-1}) \in \mathcal{U}^T | \ B^{u_0,\ldots,u_{T-1}}(x) = \max_{(u'_0,\ldots,u'_{T-1}) \in \mathcal{U}^T} B^{u'_0,\ldots,u'_{T-1}}(x)\}.$$

- Finding an element of $\mathfrak{B}^*(x)$ can be reformulated as a shortest path problem (see Figure 2).

## 5 CONSISTENCY

**Theorem 5.1 (Consistency of CGRL algorithm)** *Let*

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

*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*

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

*Then*

$$C\alpha^* < \epsilon(x) \implies (\hat{u}_0^*(x), \ldots, \hat{u}_{T-1}^*(x)) \in \mathfrak{J}^*(x) \ .$$

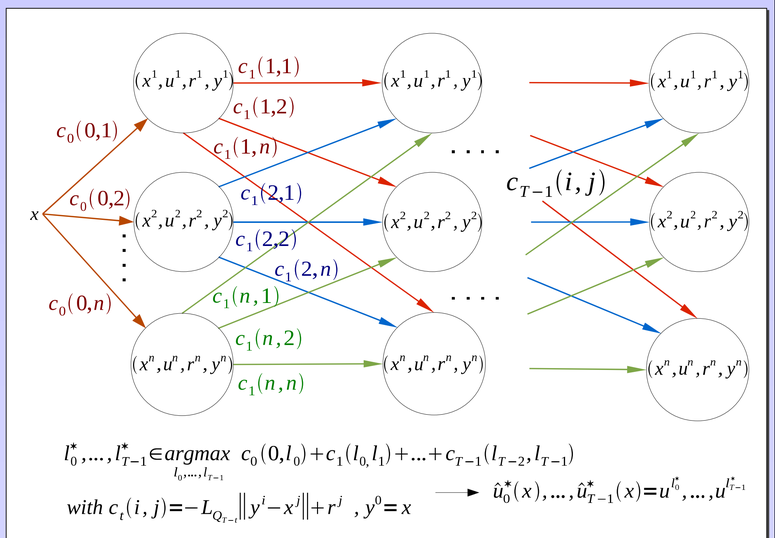


FIG. 2: A graphical interpretation of the CGRL algorithm (notice that $n = |\mathcal{F}|$)

## 6 PRELIMINARY VALIDATION

**The puddle word benchmark** The CGRL algorithm is compared with the Fitted Q Iteration(FQI) algorithm [1] on two samples $\mathcal{F}_1$ ("normal" sample) and $\mathcal{F}_2$ (no information about the puddle).

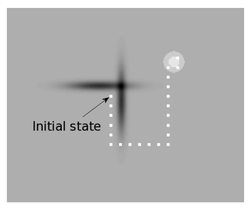


FIG. 3: CGRL with $\mathcal{F}_1$.

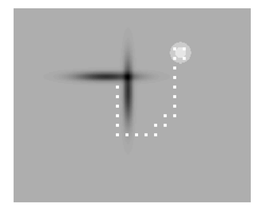

FIG. 4: FQI with $\mathcal{F}_1$.

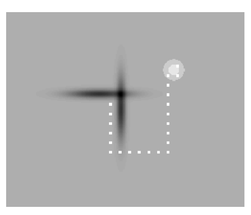

FIG. 5: CGRL with $\mathcal{F}_2$.

FIG. 6: FQI with $\mathcal{F}_2$.

**HIV infection** Database generation: A patient does not take his antiretroviral therapy in average once every eight days. CGRL is run on the trajectory generated by this patient.

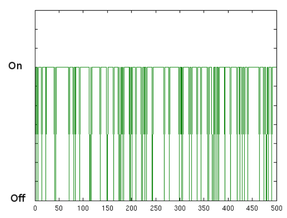


FIG. 7: Treatment evolution for generating the database

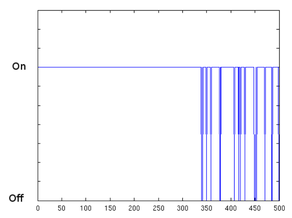


FIG. 8: Treatment evolution computed by the CGRL algorithm

## 7 FUTURE WORK

- Extension of the CGRL algorithm to a stochastic framework / on-line learning framework,
- Derivation of the CGRL algorithm to address the exploitation / exploration tradeoff,
- Selecting concise sets of transitions.

## Acknowledgement

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. RF acknowledges the financial support of the FRIA. DE is a research associate of the FRS-FNRS. We also acknowledge financial support from NIH grants P50 DA10075 and R01 MH080015. The scientific responsibility rests with its authors.

## References

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

[2] R. Fonteneau, S. Murphy, L. Wehenkel, and D. Ernst. 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.

[3] R. Fonteneau, S.A. Murphy, L. Wehenkel, and D. Ernst. 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.