Proceedings of the 45th IEEE Conference on Decision & Control
Manchester Grand Hyatt Hotel
San Diego, CA, USA, December 13-15, 2006

Clinical data based optimal STI strategies for HIV: a reinforcement learning approach

Damien Ernst, Guy-Bart Stan, Jorge Gonçalves, and Louis Wehenkel

Abstract— **This paper addresses the problem of computing optimal structured treatment interruption strategies for HIV infected patients. We show that reinforcement learning may be useful to extract such strategies directly from clinical data, without the need of an accurate mathematical model of HIV infection dynamics. To support our claims, we report simulation results obtained by running a recently proposed batch-mode reinforcement learning algorithm, known as fitted Q iteration, on numerically generated data.**

I. Introduction

Human Immunodeficiency Virus (HIV) is a retrovirus that may lead to the lethal Acquired Immune Deficiency Syndrome (AIDS). After initial contact and inclusion of the HIV particle into a cell of the immune system (e.g. $CD4^+$ T-lymphocytes and macrophages), there is a cascade of intracellular events leading to the production of massive numbers of new viral particles, the death of infected cells, and ultimately the devastation of the immune system.

Since the first identification of such unusual immune system failure in 1981, many advances have been made in the design of anti-HIV drugs and treatments. Current anti-HIV drugs can be roughly grouped into two main categories: Reverse Transcriptase Inhibitors (RTI) and Protease Inhibitors (PI). The action of RTIs is to prevent HIV RNA from being converted into DNA, thereby blocking the virus replication process initiated in the infected cell. The protease inhibitors work at the final stage of viral replication and attempt to prevent HIV from making new copies of itself by interfering with the HIV protease enzyme. As a result, the new copies of HIV are not able to infect new cells.

Typical treatments for acutely infected HIV patients utilize two or more drugs. Generally, these drug cocktails consist of one or more RTIs in combination with a PI. Despite the great success of these drug cocktails in reducing and maintaining viral loads below the detection limit, their long-term use yields substantial complications. Patients taking these drugs experience many common and sometimes highly undesirable side effects, often leading to poor compliance. Furthermore, the HIV mutates into new viral strains that become with time resistant to current drugs, resulting in the need to change drugs or even in the inability to find appropriate pharmaceutical treatments.

Concerns about this long term use of drugs have brought attention for the need of efficient drug-scheduling strategies. Idealistically, a drug-scheduling strategy should bring the immune system into a state that allows it to independently (without help from any drug) maintain immune control over the virus. Also, this transfer to a drug-independent viral control situation should be done with as low as possible drug-related systemic effects for the patients.

One such strategy, currently receiving a lot of attention, is structured treatment interruption (STI), in which patients are cycled on and off drug therapy (Bonhoeffer et al., 2000; Lisziewicz et al., 2000). STI strategies are often well-received by patients since they offer them periods of relief from treatment. During interruptions, viral load set points typically rebound to a high level, consequently activating an adaptive immune response. In some remarkable cases, it has been reported that repeated STI stimulations have enabled patients to maintain immune control over the virus in the absence of treatment (Lisziewicz et al., 1999).

More recently, several authors have addressed the problem of designing STI treatments by exploiting mathematical models of HIV infection dynamics (Adams et al., 2004; Bajaria et al., 2004). These models are usually represented by a set of Ordinary Differential Equations (ODEs), and deduction of STI strategies from them is done by using methods from control theory. Modelling the HIV infection dynamics is however a complex task. Not only does one have to select the right parametric system of ODEs, but one must also fit their parameters to reflect quantitatively biological observations. An interesting alternative would be to infer STI strategies directly from clinical data, without having to specify and identify a model of the HIV infection dynamics.

Typically, when a patient undergoes a STI treatment, clinical data representing the time-evolution of the patient's state ($CD4^+$ T cell count, systemic costs of drugs, etc.) are recorded at specific, discrete-time instants. Such clinical data may be seen as trajectories of the immune system responding to the treatment.

The problem of inferring from trajectories of a system an appropriate way to control it has been extensively studied in control theory and computer science. One way to approach it is to first state an optimality criterion and then search for control strategies optimizing this criterion. In particular, the classical approach consists of using the trajectories to identify an analytical model, and deriving a controller from this model and from the optimality criterion (Bitmead

Damien Ernst is with the Hybrid Systems Control Group of Supélec-IETR, Rennes, France

Guy-Bart Stan and Jorge Gonçalves are with the Control Group of the Department of Engineering, University of Cambridge, United Kingdom

Louis Wehenkel is with the Systems and Modelling Group of the Department of Electrical Engineering and Computer Science, University of Liège, Belgium

Email addresses: damien.ernst@supelec.fr; gvs22@eng.cam.ac.uk; jmg77@eng.cam.ac.uk; l.wehenkel@ulg.ac.be

1-4244-0171-2/06/$20.00 ©2006 IEEE.

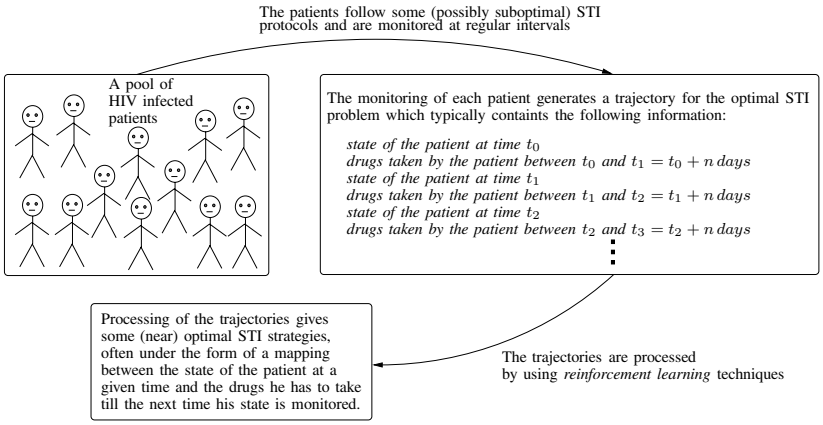

Fig. 1. Determination of optimal STI strategies from clinical data by using reinforcement learning algorithms: the overall principle.

et al., 1990). Reinforcement Learning (RL), on the other hand computes control strategies directly from the measured trajectories, without the need for identifying a priori a model of the system dynamics (Sutton & Barto, 1998).

In this paper, we aim at investigating the feasibility of using RL to determine (close-to-)optimal HIV-STI strategies from clinical data alone, in other words, without relying on the identification of an accurate model of the HIV infection dynamics. In this approach, illustrated in Figure 1, HIV-infected patients follow during clinical trials various STI protocols. Their states are monitored every n days and the trajectories gathered from this monitoring are processed by the RL algorithm to compute new STI strategies.

The paper is structured as follows. Section II formalizes the problem of learning optimal strategies from a set of trajectories and introduces the algorithms used in our simulations. Section III reports simulation results obtained by using the RL-based approach to determine STI strategies from clinical data optimal. Instead of actual clinical data, we have used synthetic ones obtained from simulations with an ODE model of the HIV infection dynamics. In Section IV, we suggest ways to overcome difficulties that may arise when relying on real-life data rather than numerically generated ones and, finally, Section V concludes.

II. Learning from a sample of trajectories: the RL approach

We start this section by formulating the problem of learning the solution of an optimal control problem from a sample of trajectories. We consider deterministic discrete-time optimal control problems for which the aim is to minimize a sum of discounted costs over an infinite time horizon. After formulating the problem, we remind some classical results from dynamic programming theory and introduce the fitted Q iteration algorithm. We refer the reader to (Bertsekas, 2000) for a comprehensive textbook on dynamic programming and to (Ernst et al., 2005) for a complement of information on the fitted Q iteration algorithm.

A. Problem formulation

Consider a system having a deterministic *discrete-time dynamics* described by:

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

where for all t, x_t is an element of the state space X and u_t is an element of the action space U. Let $c(x, u)$ be a (real-valued) cost function whose infinite norm is bounded by some positive constant B_c, and γ be a discount factor ($0 \leq \gamma < 1$).

Given a stationary control strategy $\mu(\cdot) : X \rightarrow U$, and assuming $x_0 = x$ and $x_{t+1} = f(x_t, \mu(x_t))$, forall t, we define the discounted infinite horizon cost function associated to μ by

$$J^{\mu}(x) = \lim_{N \rightarrow \infty} \sum_{t=0}^{N-1} \gamma^t c(x_t, \mu(x_t)). \tag{2}$$

The objective is to find an optimal stationary strategy μ^*, i.e. a strategy that minimizes J^{μ} for all x.

In order to compute such a strategy, we do not assume that the system dynamics (1) is known. However, we suppose available a (finite) set of (finite duration) system trajectories (in the form $(x_0, u_0, x_1, u_1, x_2, \cdots, x_{T-1}, u_{T-1}, x_T)$) as well as the cost-function $c(x, u)$. Reinforcement learning techniques compute from this kind of information an *approximation* $\hat{\mu}^*$ of the optimal stationary strategy since, except for very special conditions, the exact optimal strategy μ^* can not be deduced from such a limited amount of information on the system dynamics.[1]

The *fitted Q iteration* algorithm which we exploit in this paper, actually relies on a slightly weaker assumption, namely that a set of *one-step* system transitions is given, each one providing the knowledge of a new sample of information (x_t, u_t, x_{t+1}). We denote this set of transitions by $\mathcal{F} = \{(x_t^l, u_t^l, x_{t+1}^l)\}_{l=1}^{\#\mathcal{F}}$.

[1] RL actually handles the more general problem when the cost function is also unknown and replaced by sample values; it also carries over to stochastic systems.

B. Some dynamic programming results

The sequence of functions $Q_N : X \times U \rightarrow \mathbb{R}$ defined by the recurrence equation

$$Q_N(x,u) = c(x,u) + \gamma \min_{u' \in U} Q_{N-1}(f(x,u),u'), \quad \forall N > 1 \tag{3}$$

with $Q_1(x,u) \equiv c(x,u)$, converges in infinity norm to the Q-function, defined as the (unique) solution of the Bellman equation:

$$Q(x,u) = c(x,u) + \gamma \min_{u' \in U} Q(f(x,u),u'). \tag{4}$$

A stationary strategy μ^* that satisfies

$$\mu^*(x) = \arg\min_{u \in U} Q(x,u) \tag{5}$$

is an optimal strategy.

Let us denote by μ^*_N the stationary strategy

$$\mu^*_N(x) = \arg\min_{u \in U} Q_N(x,u). \tag{6}$$

The following bound on the suboptimality of μ^*_N with respect to μ^* holds (see (Ernst et al., 2005)):

$$\left\| J^{\mu^*_N} - J^{\mu^*} \right\|_\infty \leq \frac{2\gamma^N B_c}{(1-\gamma)^2}. \tag{7}$$

C. The fitted Q iteration algorithm

From the set of transitions $\mathcal{F}$, the fitted Q iteration algorithm computes the functions $\hat{Q}_1$, $\hat{Q}_2$, ..., $\hat{Q}_N$ which constitute approximations of the functions Q_1, Q_2, ..., Q_N defined by Eqn (3). This computation is done iteratively by solving a sequence of standard batch-mode supervised learning problems. The training sample for the k^{th} ($k \geq 2$) supervised learning problem of the sequence is

$$\left\{ \left((x^l_t, u^l_t),\ c(x^l_t, u^l_t) + \gamma \min_{u \in U} \hat{Q}_{k-1}(x^l_{t+1}, u) \right) \right\}_{l=1}^{\#\mathcal{F}}$$

with $\hat{Q}_1(x,u) \equiv c(x,u)$. Based on this training sample, the supervised learning (regression) algorithm produces the function $\hat{Q}_k$ that is used to determine the next training sample and from there, the next function of the sequence. Once the approximation functions $\hat{Q}_1$, $\hat{Q}_2$, ..., $\hat{Q}_N$ have been computed, the (sub-optimal) stationary strategy

$$\hat{\mu}^*_N(x) = \arg\min_{u \in U} \hat{Q}_N(x,u) \tag{8}$$

is taken as approximation of the optimal stationary strategy $\mu^*(x)$.

As batch-mode supervised learning algorithm, we have chosen the Extra-Trees algorithm (Geurts et al., 2006). This algorithm builds a model in the form of the average prediction of an ensemble of regressions trees obtained by randomization. It has three parameters: the number M of trees composing the ensemble, the minimum number of elements required to split a node n_{min} and the maximum number of cut-directions evaluated at each node K. These values have been chosen respectively equal to 50, 2 (the trees are fully developed) and the dimensionality of the input space (equal to 8 (6 state variables + 2 control variables) for the problem treated in Section III).

III. Simulation results

In this section we present the results we have obtained by using the RL-based approach on artificially generated data. We first define the kind of STI strategies we are looking for, in terms of the class of strategies considered and their optimality criterion. Then, we describe the simulation protocol behind the data generation and, finally, we discuss the obtained STI-strategy. Our work in this section is directly inspired from (Adams et al., 2004).

A. Kinds of STI strategies targeted

As in (Adams et al., 2004), we consider bi-therapy treatments combining a fixed RTI and a fixed PI. The protocol allows to revise drug administration every five days based on clinical measurements, by choosing one of the four possible on-off combinations for the next five days: RTI and PI on, only RTI on, only STI on, RTI and PI off. These four cocktails hence define the set of actions U of our optimal control problem.

In terms of optimality criterion, we seek STI strategies that minimize a sum of discounted instantaneous costs over an infinite horizon with the instantaneous cost at time t being given by:

$$c(x_t, u_t) = QV_t + R_1\epsilon^2_{1_t} + R_2\epsilon^2_{2_t} - SE_t \tag{9}$$

where $Q = 0.1$, $R_1 = 20000$, $R_2 = 2000$, $S = 1000$, $\epsilon_{1_t} = 0.7$ (resp. $\epsilon_{1_t} = 0$) if the RTI is cycled on (resp. off) at time t, and $\epsilon_{2_t} = 0.3$ (resp. $\epsilon_{2_t} = 0$) if the PI is cycled on (resp. off) at time t. V is the number of free HI viruses (in copies/ml) and E the number of cytotoxic T-lymphocytes (in cells/ml). Cytotoxic T-lymphocytes constitute the specific immune response of the body to HI viruses. The decay factor γ has been chosen equal to 0.98, which means that costs occurring after one year weight for approximately three-quarter less than costs occurring at instant $t = 0$.

We refer the reader to (Adams et al., 2004) for a discussion of rationale behind this cost function.[2]

B. Artificial generation of the clinical data

In order to evaluate the ability of RL to compute "good" STI strategies, we will apply the fitted Q iteration algorithm described in Section II-C on artificially generated data.

To obtain data which mimic real-life clinical data, we have used time-domain simulations of the nonlinear ODE model published in (Adams et al., 2004), which was validated and identified from real-life clinical data. In order to provide insight into the physical problem that is tackled, we briefly

[2]In (Adams et al., 2004), optimal strategies are computed by assuming that the dynamics of the HIV immune response are known. On the contrary, here we compute strategies from the sole knowledge of samples of transitions $\mathcal{F}$. Furthermore, we consider an optimal control problem with infinite time horizon and discounted costs while in (Adams et al., 2004) a finite horizon and undiscounted costs are considered. As a consequence, decisions made by strategies derived in our approach depend only on the current state of a patient. In (Adams et al., 2004) they also depend on the time elapsed since the beginning of the treatment, which means that patients presenting exactly the "same medical states" but at different stages of their treatment may undergo different STI strategies, which we believe is not appropriate.

discuss the main characteristics of this model, before defining the data generation procedure itself.

The dynamic model has six state variables that represent respectively the number of healthy CD4$^+$ T-lymphocytes (referred to as T_1), the number of healthy macrophages (T_2), the number of infected CD4$^+$ T-lymphocytes (T_1^*), the number of infected macrophages (T_2^*), the number of free virus particles (V) and the number of HIV-specific cytotoxic T-cells (E). Note that these variables are assumed to be measured every five days, in order to select the drug combination for the next five days.

As shown in (Adams et al., 2004), in the absence of treatment (i.e. $\epsilon_{1_t} = \epsilon_{2_t} \equiv 0$), the system of ordinary differential equations exhibits three physical equilibrium points (and several non physical ones (omitted here) for which one or more state variables are negative). These equilibrium points are, respectively, an unstable equilibrium point

$$(T_1, T_2, T_1^*, T_2^*, V, E) = (10^6, 3198, 0, 0, 0, 10)$$

which represents an uninfected state, and two locally stable equilibria corresponding to HIV-infected states. The HIV-infected equilibria may be categorized as:

1) a "healthy" locally stable equilibrium point

   $$(T_1, T_2, T_1^*, T_2^*, V, E) = (967839, 621, 76, 6, 415, 353108)$$

   which corresponds to a small viral load, a high CD4$^+$ T-lymphocytes count and a high HIV-specific cytotoxic T-cells count,
2) the "non-healthy" locally stable equilibrium point

   $$(T_1, T_2, T_1^*, T_2^*, V, E) = (163573, 5, 11945, 46, 63919, 24)$$

   for which T-cells are depleted and the viral load is very high.

Numerical simulations show that the basin of attraction of the healthy steady-state is relatively small in comparison with the one of the non-healthy steady-state. Furthermore, perturbation of the uninfected steady-state by adding as less as one single particle of virus per ml of blood plasma leads to asymptotical convergence towards the non-healthy steady-state.

During the data collection process, we assume that the (simulated) patients are monitored (and the medication protocol revised) every five days. The monitoring period for each patient is assumed to last for 1000 days.

The generation procedure of the clinical data is iterative. At the first iteration, we consider thirty patients in "non-healthy" steady-state. Every five days, the physiological data of each of these thirty patients (assumed here to be summarized by the quantities T_1, T_2, T_1^*, T_2^*, V, and E) are recorded and a new type of medication is randomly selected in U. The monitoring of each patient generates a trajectory $(x_0, u_0, x_1, \cdots, x_{199}, u_{199}, x_{200})$ from which we can extract $1000/5 = 200$ samples (x_t, u_t, x_{t+1}).

At the second step of the iterative process, we also consider a set of thirty patients in "non-healthy" steady-state and, once again, we record their physiological data every five days. Nevertheless, contrary to the first step, each five days, the corresponding drug cocktail is not selected at random anymore. Instead, the medication for these new thirty patients is determined by the following STI strategy: in 85% of the cases we use the strategy $\hat{\mu}^*_{400}$ computed by the fitted Q iteration algorithm[3] applied on the $6,000$ element set generated by the monitoring of the previous 30 patients, while in the remaining 15% cases we use a type of medication randomly selected in U.

At the third iteration, another set of thirty trajectories are generated in identical conditions, except that the corresponding STI strategy uses now in 85% of the cases a strategy $\hat{\mu}^*_{400}$ inferred from all the samples gathered previously (i.e. $12,000$ samples). By repeating this iterative procedure ten times, we have generated a total of 300 trajectories (10 sets of 30 patients) to which correspond $60,000$ samples (x_t, u_t, x_{t+1}).

The reader may wonder why we interlaced the generation of the samples with the computation of $\hat{\mu}^*_{400}$ and used this newly computed strategy to generate additional samples. There are two main reasons behind this choice. First, we wanted to simulate a situation in which STI strategies administered to patients were not chosen totally at random but rather benefit, at least partially, from the knowledge clinicians may already have about "good" STI strategies. Second, by using some knowledge already acquired about $\hat{\mu}^*_{400}$, we tend to gather much more information alongside the optimal trajectories. As a consequence, with a fairly small number of clinical trials we can converge rather quickly to close-to-optimal STI strategies.

C. Results

On Figure 2, we have represented the evolution of the cell counts, number of free viruses and immune effectors of a patient treated from "non-healthy" steady-state by the STI strategy inferred from the set of $60,000$ samples by the fitted Q iteration algorithm. As desired, the computed (close-to-)optimal STI strategy is able to bring the patient to the domain of attraction of the "healthy" drug-free steady-state. On the same figure, trajectories that would have been observed by putting the patient always on or always off both drugs have also been plotted. Compared to these two strategies, the RL-based STI strategy leads to higher T-cell counts, lower virus load, and significantly boosts the specific anti-HIV immune response.

In Figure 3 it can be seen that with the RL computed STI strategy the patients get active treatment, with some periods of relief, during approximately 380 days and are always put off both drugs afterwards (definitive treatment interruption after 380 days).

Usually, the quality of the strategies determined by RL increases with the number of trajectories since each additional trajectory generally provides additional information about the

[3]In all the simulation results reported in this paper, the fitted Q iteration algorithm is iterated 400 times and $\hat{\mu}^*_{400}$ is taken as approximation of the optimal stationary strategy $\hat{\mu}^*$. Side simulations have shown that the computed strategy remained mostly unchanged by increasing the number of iterations.

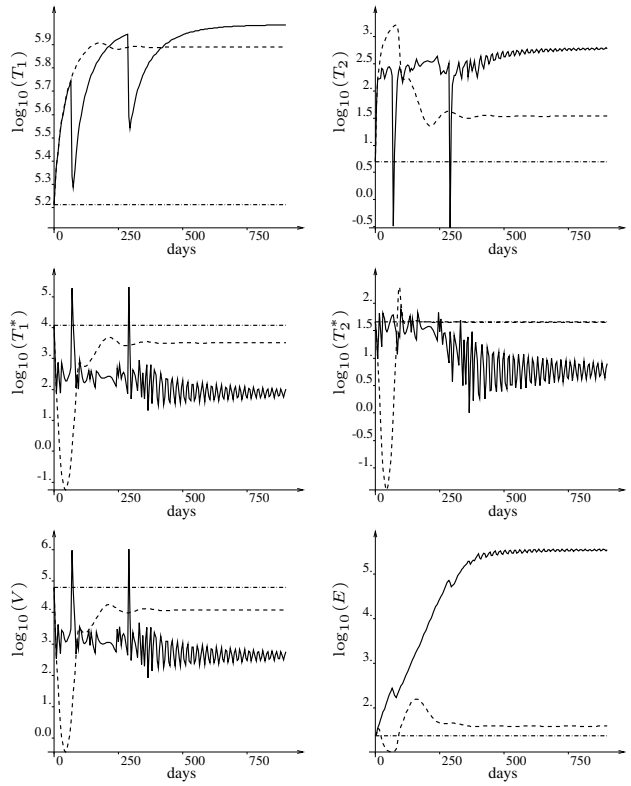

Fig. 2. The curves represent the time evolution of the different cells count (T_1, T_2, T_1^*, T_2^*), of the number of free virus particles (V) and of the number of immune effectors (E) for a patient being treated from "non-healthy" steady-state. The solid curve ($-$) corresponds to the STI strategy plotted on Fig. 3 and computed by the reinforcement learning algorithms. The dashed curves ($--$) represent the time evolution of these variables when there is no interruption in the treatment (i.e. $\epsilon_{1_t} = 0.7$ and $\epsilon_{2_t} = 0.3$, $\forall t \geq 0$) and the dotted curves ($-\cdot$) represent their time evolution when there is no treatment (i.e. $\epsilon_{1_t} = \epsilon_{2_t} = 0$, $\forall t \geq 0$).

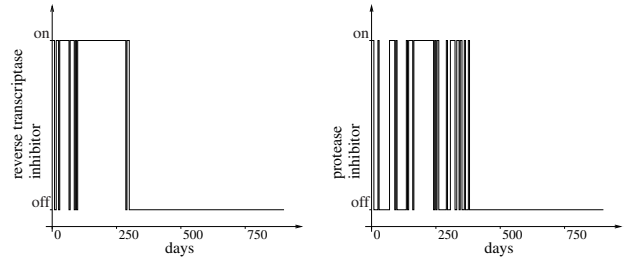

Fig. 3. Representation of the STI treatment for a patient treated from early stage of infection. The STI treatment is computed by the reinforcement learning algorithms on clinical data generated by 300 patients.

underlying problem. This is illustrated on Figure 4 where we have plotted infinite horizon costs associated with strategies computed by considering an increasing number of patients in the clinical trials. Note that in this particular case, STI strategies that put the patient always on (or alway off) both drugs produce larger costs than those obtained by using the STI strategy derived from only ten trajectories.

Overall, these results suggest that reinforcement learning can indeed infer appopriate STI strategies from a sample of transitions reflecting the instantaneous response of patients to drug administration at different stages of their treatment,

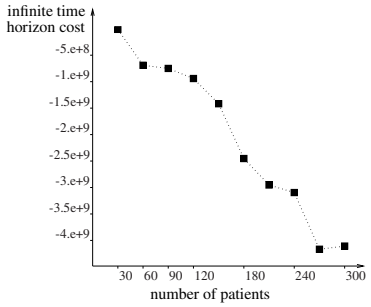

Fig. 4. Influence of the number of patients participating to the clinical trials on the infinite horizon cost corresponding to the computed STI strategies. Data generation follows the protocol described in Section III-B. To compute the infinite horizon cost associated to a given number of patients, we run RL on the trajectories generated by these patients and estimate $J^{\hat{\mu}^*_{400}}(x)$ obtained when a patient intitially in the "non-healthy" steady-state is treated by the learned strategy ($J^{\hat{\mu}^*_{400}}(x) = \sum_{t=0}^{\infty} \gamma^t c(x_t, \hat{\mu}^*_{400}(x_t))$ with $x = (T_{1_0}, T_{2_0}, T^*_{1_0}, T^*_{2_0}, E_0, V_0) = (163573, 5, 11945, 46, 63919, 24)$).

without explicit knowledge of the underlying dynamics.

IV. From numerically simulated to real-life patients

In the previous section, we have reported some results obtained by using numerical simulations to reproduce the clinical evolution of HIV-infected patients. In this section, we discuss the four main difficulties we expect to face when dealing with real-life patients.

The HIV/immune system interaction dynamics may be different from one patient to the other. When generating the clinical data, we have implicitly assumed that the dynamics of the interaction between HIV and the patients' immune system were the same for every patient. In real life conditions, these dynamics may substantially vary from one patient to the other. Some reasons for these discrepancies are: variance in the patients' immune systems, existence of different types of HIV infections, individual differences in the assimilation of the drugs, etc. We believe that one appropriate approach to address such a difficulty would be to add to the state vector x relevant information about the specifics of each patient's case (e.g. general medical condition, type of HIV virus (HIV-1, HIV-2), presence of drug-resistant HIV strains, etc.).

Proper statement of the optimal control problem. Different elements need to be defined when stating the optimal control problem: the time discretization, the cost function and the decay factor. These elements should be chosen to lead to desirable optimal trajectories and good learning speed. When working in a numerical environment, trial-and-error type of approaches can help to choose these elements. Trial-and-error approaches can however not be used on real patients. Thus, we will need to call for medical expertise in order to state properly the optimal control problem, but we also believe that some specific tools should be built to help in this task.

Partial observability. In our example, we have assumed that all the state variables were directly observable. When dealing with real patients, such an assumption is not fully realistic

since, among others, it is not possible with current technology to distinguish between healthy and non-healthy $CD4^+$ T-lymphocytes and macrophages. It is therefore clear that some partial observability issues will arise when processing real-life data. We refer the reader to (Murphy, 2000) for a survey of solution techniques for partial observable discrete-time optimal control problems.

Corrupted measurements. Collected clinical data are not necessarily thorough and accurate. Furthermore, the patients may not necessarily comply with the prescribed treatment. This may lead to uncertainties and measurement corruption which may significantly degrade the quality of the results obtained. One solution to mitigate the adverse effects of corrupted measurements would be to design some preprocessing algorithms able to filter out highly corrupted data.

V. Conclusions

In this paper, we have considered the problem of computing structured treatment interruption strategies for HIV infected patients from clinical data only. In the envisioned protocol, the clinical data would be generated by monitoring at regular time intervals the state of various patients during their treatment, and these data would be exploited by reinforcement learning to determine an optimal drug prescription strategy.

To investigate the validity of such a purely data driven approach, we have generated clinical data artificially by relying on a plausible mathematical model of the HIV infection dynamics. Based on a sufficient amount of simulated data, we found that reinforcement learning was indeed able to derive STI therapies which appear as excellent when used to "treat" simulated patients.

These encouraging results suggest that reinforcement learning techniques could also help to design effective real-life STI strategies from actual clinical data. The next step of this research will be to study more extensively, still by simulations, various difficulties that could be encountered when applying this approach in real-life. In particular, we expect that many problems will arise such as those related to corrupted data, variance in HIV viruses, inter-individual differences of the immune responses, and inability to count specific types of immune cells playing a critical role in the HIV infection.

Finally, although we target the development of model-free methods, we would like to stress the usefulness even in this kind of research of plausible analytical models of the dynamic response of patients to treatments. While we believe that it might not be possible to derive accurate enough dynamic models for the direct derivation of appropriate treatment strategies, it is clear that even approximate or highly simplified models may be very useful to gain understanding of a problem and to design an appropriate way to apply reinforcement learning to it. As a matter of fact, only after extensive "in silico" experiments one will gain enough confidence to start using this kind of approach in actual "in vivo" conditions.

Acknowledgment

Damien Ernst gratefully acknowledges the financial support of the FNRS (French acronym for the Belgian National Fund of Scientific Research) of which he was a postdoctoral researcher. Guy-Bart Stan thanks the European Commission for supporting his research through a FP6 Marie-Curie Intra-European Fellowship.

References

Adams, B., Banks, H., Kwon, H.-D., & Tran, H. (2004). Dynamic multidrug therapies for HIV: Optimal and STI control approaches. *Mathematical Biosciences and Engineering*, *1*, 223–241.

Bajaria, S., Webb, G., & Kirschner, D. (2004). Predicting differential responses to structured treatment interruptions during HAART. *Bull. Math. Biol.*, *66*, 1093–1118.

Bertsekas, D. (2000). *Dynamic Programming and Optimal Control*, vol. I. Belmont, MA: Athena Scientific. 2nd edition.

Bitmead, R., Gevers, M., & Werts, V. (1990). *Adaptive Optimal Control: The Thinking Man's GPC*. Prentice Hall International.

Bonhoeffer, S., Rembiszewski, M., Ortiz, G., & Nixon, D. (2000). Risks and benefits of structured antiretroviral drug therapy interruptions in HIV-1 infection. *AIDS*, *14*, 2313–2322.

Ernst, D., Geurts, P., & Wehenkel, L. (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, *6*, 503–556.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *36*, 3–42.

Lisziewicz, J., Rosenberg, E., & Liebermann, J. (1999). Control of HIV despite the discontinuation of anti-retroviral therapy. *New England J. Med.*, *340*, 1683–1684.

Lisziewicz, J., Rosenberr, E., & Liebermann, J. (2000). Structured treatment interruptions to control HIV-1 infection. *The Lancet*, *354*, 287–288.

Murphy, K. (2000). *A survey of POMDP solution techniques* (Technical Report). University of California at Berkeley.

Sutton, R., & Barto, A. (1998). *Reinforcement Learning, an Introduction*. MIT Press.