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

Guy-Bart Stan

Department of Engineering (Control Group), University of Cambridge, U.K.

UCL Seminar, April 17, 2007

Collaborators: D. Ernst, J. Gonçalves and L. Wehenkel
1 Introduction
   - HIV: Definition and Treatments

2 STI
   - STI: Definition
   - STI: A glimpse at today’s practice
   - More advanced: Model based design of optimal STI
   - More advanced: Clinical data based design of optimal STI

3 The RL approach
   - RL: Problem formulation
   - RL: Some dynamic programming results
   - RL: Fitted Q iteration
   - Illustration: Using Fitted Q iteration to compute an optimal STI
   - From simulation to reality

4 Conclusions
Outline

1 Introduction
   - HIV: Definition and Treatments

2 STI
   - STI: Definition
   - STI: A glimpse at today’s practice
   - More advanced: Model based design of optimal STI
   - More advanced: Clinical data based design of optimal STI

3 The RL approach
   - RL: Problem formulation
   - RL: Some dynamic programming results
   - RL: Fitted Q iteration
   - Illustration: Using Fitted Q iteration to compute an optimal STI
   - From simulation to reality

4 Conclusions
Human Immunodeficiency Virus (HIV) is a retrovirus at the source of the Acquired Immune Deficiency Syndrome (AIDS).

- HIV particles target cells of the immune system (mostly CD4+ lymphocytes and macrophages).
- Inclusion of HIV particles in immune cells lead to massive production of new viral particles, death of the infected cells and, ultimately, devastation of the immune system.
Current anti-HIV drugs

Two main categories:
1. Reverse Transcriptase Inhibitors (RTI);
2. Protease Inhibitors (PI).

Figure: Taken from http://www.cellsalive.com/hiv0.htm
Treatments for infected patients

Highly Active Anti-Retroviral Therapy (HAART): combination of two or more drugs. Usually one or more RTIs in combination with a PI.

- Two main concerns about the long-term use of anti-retroviral drugs:
  1. undesirable side effects (leading to poor compliance);
  2. mutation of the virus (need to change drugs or even inability to find appropriate pharmaceutical treatments).

⇒ Need for “efficient” drug-scheduling strategies. Idealistcally, a drug-scheduling strategy should bring as quickly as possible the immune system to a drug-independent state where it has immune control over the virus (with low amount of drugs and low systemic effects).
Outline

1 Introduction
   - HIV: Definition and Treatments

2 STI
   - STI: Definition
   - STI: A glimpse at today’s practice
   - More advanced: Model based design of optimal STI
   - More advanced: Clinical data based design of optimal STI

3 The RL approach
   - RL: Problem formulation
   - RL: Some dynamic programming results
   - RL: Fitted $Q$ iteration
   - Illustration: Using Fitted $Q$ iteration to compute an optimal STI
   - From simulation to reality

4 Conclusions
Structured Treatment Interruption (STI): Definition

**STI** ⇒ cycle the patient on and off drug therapy.

- STI strategies often well received by patients since they offer them period of relief from treatment.
- In some remarkable cases, STI strategies have enabled the patients to maintain immune control over the virus in the absence of treatment.

**Goal of this research:** compute optimal STI strategies.
SMART study: If CD4⁺ T-cells count falls below a certain threshold, put the patient on drugs. Otherwise put him off drugs.

- This practice has met some problems:

**Figure:** Taken from http://www.cpcra.org/docs/pubs/2006/croi2006-smart.pdf

**Conclusion**

Episodic use of ART based on CD4+ cell count levels as per the SMART study design is inferior to continuous ART for the management of treatment-experienced patients. Thus, this strategy should not be recommended.
Model based design of optimal STI

Some authors have proposed to design STI treatments by exploiting mathematical models of the HIV infection.

- Models are under the form of a set of Ordinary Differential Equations (ODEs).
- Deduction of STI strategies is done by using model-based methods from the control theory.

But modelling of the HIV dynamics is a difficult task:
- selection of the right parametric system of ODEs;
- fitting of the parameters to quantitatively reflect biological observations.
Clinical data based design of optimal STI

Compute “good” STI strategies directly from clinical data (without modelling the HIV infection dynamics).

- Clinical data contain time evolution of patient’s state (CD4$^{+}$ T-cells count, systemic costs of the drugs, etc.) recorded at discrete-time instant in correlation with sequence of drugs administered.

- Clinical data can be seen as trajectories of the immune system responding to the administered treatment.
Outline

1. Introduction
   - HIV: Definition and Treatments

2. STI
   - STI: Definition
   - STI: A glimpse at today’s practice
   - More advanced: Model based design of optimal STI
   - More advanced: Clinical data based design of optimal STI

3. The RL approach
   - RL: Problem formulation
   - RL: Some dynamic programming results
   - RL: Fitted $Q$ iteration
   - Illustration: Using Fitted $Q$ iteration to compute an optimal STI
   - From simulation to reality

4. Conclusions
The RL approach

Inferring appropriate control policies from trajectories
Problem of inferring appropriate control policies from trajectories has been studied in control theory and computer science.

Classical way to approach it: state an optimality criterion and search for strategies optimizing this criterion.

- **Model-based (control theory) approach**: infer a model and derive from this model and the chosen optimality criterion an optimal strategy.

- **Reinforcement learning (computer science) approach**: compute optimal strategies directly from the trajectory, without identifying a model a priori.
The patients follow some (possibly suboptimal) STI protocols and are monitored at regular intervals. Processing of the trajectories gives some (near) optimal STI strategies, often under the form of a mapping between the state of the patient at a given time and the drugs he has to take till the next time his state is monitored.

The monitoring of each patient generates a trajectory for the optimal STI problem which typically contains the following information:

- state of the patient at time $t_0$
- drugs taken by the patient between $t_0$ and $t_1 = t_0 + n$ days
- state of the patient at time $t_1$
- drugs taken by the patient between $t_1$ and $t_2 = t_1 + n$ days
- state of the patient at time $t_2$
- drugs taken by the patient between $t_2$ and $t_3 = t_2 + n$ days

The trajectories are processed by using reinforcement learning techniques.
RL: Problem formulation

- **Discrete-time dynamics:**
  \[ x_{t+1} = f(x_t, u_t) \quad t = 0, 1, \ldots \]
  where \( x_t \in X \) and \( u_t \in U \).

- **Cost function:** \( c(x_t, u_t) : X \times U \to \mathbb{R} \) with \( \|c(x_t, u_t)\|_{\infty} < B_c \).

- **Disc. inf. horizon cost assoc. to stationary policy \( \mu : X \to U \):**
  \[ J^\mu(x) = \lim_{N \to \infty} \sum_{t=0}^{N-1} \gamma^t c(x_t, \mu(x_t)) \]
  with \( x_0 = x \) and \( 0 \leq \gamma < 1 \).

- **Objective:** Find an optim. stationary policy \( \mu^* \), i.e. a policy that min. \( J^\mu(x) \) for all \( x \).

**We do not know:** The discrete-time dynamics \( f(\cdot, \cdot) \).

**We know instead:** A set of trajectories
\((x_0, u_0, x_1[, c(x_0, u_0)], \cdots, x_T-1, u_T-1, x_T[, c(x_T-1, u_T-1)]))\).
Dynamic Programming paradigm

\[ x_{t+1} = f(x_t, u_t) \]

\[ c(x_t, u_t) \]

\[ \gamma \times \text{"Best (Next Move)"} \]

Example: Shortest Path problem
The Q function

- Sequence of functions $Q_N: X \times U \rightarrow \mathbb{R}$

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

with $Q_1(x, u) \equiv c(x, u)$, converges as $N \rightarrow \infty$ to the Q-function, unique solution of the Bellman equation:

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

- Relations between $J^{\mu^*}(x)$ and $Q(x, u)$

$$J^{\mu^*}(x) = \min_u Q(x, u)$$

$$Q(x, u) = c(x, u) + \gamma J^{\mu^*}(f(x, u))$$
Optimal and suboptimal policies

- Nec. and suff. optimality condition: $\mu^*(x) \in \arg\min_{u \in U} Q(x, u)$.
- Stationary policy $\mu_N^*$: $\mu_N^*(x) \in \arg\min_{u \in U} Q_N(x, u)$.
- Bound on the subopt. of $\mu_N^*$ w.r.t. $\mu^*$: $\| J^{\mu_N^*} - J^{\mu^*} \|_\infty \leq \frac{2\gamma^N B_c}{(1-\gamma)^2}$

where $J^{\mu_N}(x) = \sum_{t=0}^{N-1} \gamma^t c(x_t, \mu(x_t))$ with $x_0 = x$ and $0 \leq \gamma < 1$.

(This requires consistency, i.e. $\| J^{\hat{\mu}_N^*} - J^{\mu_N^*} \|_\infty$ arbitrarily small as $\#\mathcal{F} \to \infty$)
RL: Fitted $Q$ iteration

- Trajectories
  $$(x_0, u_0, x_1, [c(x_0, u_0)], \ldots, x_{T-1}, u_{T-1}, x_T, [c(x_{T-1}, u_{T-1})])$$
  transformed into a set of one-step system transitions
  $$F = \left\{ (x^l_t, u^l_t, x^l_{t+1}, [c(x_t, u_t)]) \right\}_{l=1}^{\#F}.$$
- Fitted $Q$ iteration computes from $F$ the functions
  $\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_N$, approximations of $Q_1, Q_2, \ldots, Q_N$.
- Computation done iteratively by solving a sequence of standard supervised learning (SL) problems. Training sample for the $k^{th}$ ($k \geq 2$) problem is
  $$\left\{ \left( (x^l_t, u^l_t), \hat{Q}_k(x^l_t, u^l_t) = c(x^l_t, u^l_t) + \gamma \min_{u \in U} \hat{Q}_{k-1}(x^l_{t+1}, u) \right) \right\}_{l=1}^{\#F}$$
  with $\hat{Q}_1(x, u) \equiv c(x, u)$.
  From the $k^{th}$ training sample, the supervised learning algorithm outputs $\hat{Q}_k(\cdot, \cdot)$.
- $\hat{\mu}_N^*(x) \in \arg \min_{u \in U} \hat{Q}_N(x, u)$ is taken as approximation of $\mu^*(x)$.

Remark: in our simulations, the SL method used is an ensemble of regression trees method named Extra-Trees.
We present results we have obtained by using the RL-based approach on artificially generated data.

The model used to generate these data is taken from
B.M. Adams, H.T. Banks, H.-D. Kwon and H.T. Tran.
(2004). “Dynamic multidrug therapies for HIV: Optimal and
STI Control Approaches”. Mathematical Biosciences and
Engineering, 1, 223-241.
Illustration: A math. model as subst. for real-life patients

\[ \begin{align*}
\dot{T}_1 &= \lambda_1 - d_1 T_1 - (1 - \epsilon_1)k_1 VT_1 \\
\dot{T}_2 &= \lambda_2 - d_2 T_2 - (1 - f\epsilon_1)k_2 VT_2 \\
\dot{T}_1^* &= (1 - \epsilon_1)k_1 VT_1 - \delta T_1^* - m_1 ET_1^* \\
\dot{T}_2^* &= (1 - f\epsilon_1)k_2 VT_2 - \delta T_2^* - m_2 ET_2^* \\
\dot{V} &= (1 - \epsilon_2)N_T \delta (T_1^* + T_2^*) - cV - [(1 - \epsilon_1)\rho_1 k_1 T_1 + (1 - f\epsilon_1)\rho_2 k_2 T_2]V \\
\dot{E} &= \lambda_E + \frac{b_E(T_1^* + T_2^*)}{(T_1^* + T_2^*) + K_b} E - \frac{d_E(T_1^* + T_2^*)}{(T_1^* + T_2^*) + K_d} E - \delta_E E
\end{align*} \]

\(T_1\) (\(T_1^*\)) = number of non-infected (infected) CD4\(^+\) lymphocytes

\(T_2\) (\(T_2^*\)) = number of non-infected (infected) macrophages

\(V\) = number of free HI viruses

\(E\) = number of HIV-specific cytotoxic \(T\)-lymphocytes

\(\epsilon_1\) and \(\epsilon_2\) = control actions corresponding to RTI and the PI
Illustration: Some insight into this model

In absence of treatment ($\epsilon_1 = \epsilon_2 = 0$), three physical equilibrium points:

1. “uninfected” locally unstable equilibrium:

   \[(T_1, T_2, T_1^*, T_2^*, V, E) = (10^6, 3198, 0, 0, 0, 10)\]

2. “healthy” locally stable equilibrium:

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

   (small viral load, a high CD4$^+$ T-lymphocytes count, high HIV-specific cytotoxic T-lymphocytes count)

3. “non-healthy” locally stable equilibrium:

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

   (T-cells depleted, viral load very high)
Illustration: State-space interpretation

- Healthy equilibrium point (locally stable with small basin of attraction)
- Non-healthy equilibrium point (locally stable with very large basin of attraction)
- Non-infected equilibrium point (locally unstable)
- Small perturbation (1 particle of HIV per ml of blood)
- Control action – drug scheduling ($u_t$)

Guy-Bart Stan - UCL Seminar, April 17, 2007
Illustration: Kinds of STI strategies targeted

- **Bi-therapy** treatments combining a fixed RTI and a fixed PI.
- Revise drug administration every five days based on clinical measurements.
- **Four possible on-off combinations** for the next five days: RTI and PI on, only RTI on, only PI on, RTI and PI off.
- We seek STI strategies that minimize $J^\mu$ defined by

$$c(x_t, u_t) = 0.1V_t + 20000\epsilon_1^2 t + 2000\epsilon_2^2 t - 1000E_t$$

$\epsilon_1 t = 0.7$ (resp. $\epsilon_1 t = 0$) if the RTI is cycled on (resp. off) at time $t$

$\epsilon_2 t = 0.3$ (resp. $\epsilon_2 t = 0$) if the PI is cycled on (resp. off) at time $t$

$V$: number of free HI viruses

$E$: number of cytotoxic T-lymphocytes

Decay factor $\gamma$: chosen equal to 0.98.
Monitoring of patients: every five days during 1000 days.

Medication: revised every five days based on the information generated by the monitoring.

Iterative generation of the clinical data (ten iterations):

- First iteration (random actions): Thirty patients in “non-healthy” steady-state. Every five days, physiological data \((T_1, T_2, T_1^*, T_2^*, V, E)\) are recorded and a new type of medication is randomly selected. Monitoring of each patient generates a trajectory \((x_0, u_0, x_1, \cdots, x_{199}, u_{199}, x_{200})\).

- Second iteration (\(\epsilon\)-greedy actions): Other set of thirty patients in “non-healthy” steady-state.
  - In 85% of the cases, use strategy \(\hat{\mu}_{400}^*\) computed by fitted Q iteration on previously generated trajectories;
  - In the remaining 15% select medication at random.

- Third to tenth iterations (\(\epsilon\)-greedy actions): similar to 2nd iteration.
Illustration: Simulation results

**Figure:** RL-computed STI treatment for a patient initially in “non-healthy” steady state.
Illustration: Simulation results
We expect to face **four main difficulties**: 

- The HIV/immune system dynamics may be different from one patient to the other
- Difficulty to state properly the optimal control problem
- Partial observability
- Corrupted measurements
Outline

1 Introduction
   - HIV: Definition and Treatments

2 STI
   - STI: Definition
   - STI: A glimpse at today’s practice
   - More advanced: Model based design of optimal STI
   - More advanced: Clinical data based design of optimal STI

3 The RL approach
   - RL: Problem formulation
   - RL: Some dynamic programming results
   - RL: Fitted $Q$ iteration
   - Illustration: Using Fitted $Q$ iteration to compute an optimal STI
   - From simulation to reality

4 Conclusions
Reinforcement learning algorithms seem to be promising tools to extract from clinical data, good STI strategies. Lot of work is however still needed to apply this approach in real-life conditions. But 40 millions of people are living with HIV/AIDS. Isn’t it a good reason to keep working hard?

Figure: Taken from UNAIDS. AIDS epidemic update: December 2005. “UNAIDS/05.19E”
$x_{t+1} = f(x_t, u_t, w_t), \quad t = 0, 1, \ldots$
Reinforcement Learning framework

- **Cost function:** \( c(x_t, u_t, w_t) : X \times U \times W \rightarrow \mathbb{R} \) with \( \|c(x_t, u_t, w_t)\|_\infty < B_c \).

- **Discounted infinite horizon cost associated to stationary policy \( \mu : X \rightarrow U \):**

\[
J^\mu(x) = \lim_{N \to \infty} \mathbb{E}_{w_t} \left( \sum_{t=0}^{N-1} \gamma^t c(x_t, \mu_N(x_t), w_t) | x_0 = x \right)
\]

with \( 0 \leq \gamma < 1 \).

- **Objective:** Find an optimal stationary policy \( \mu^*(x) \), i.e. a policy that minimizes \( J^\mu(x) \) for all \( x \).
Dynamic Programming paradigm

\[ x_{t+1} = f(x_t, u_t, w_t) \]

\[ u_{t+1} \]

\[ c(x_t, u_t, w_t) \]

\[ \gamma \times \text{"Best (Next Move)"} \]

Guy-Bart Stan - UCL Seminar, April 17, 2007
Some dynamic programming results

The \( Q \) function

- Sequence of functions \( Q_N: X \times U \rightarrow \mathbb{R} \)

\[
Q_N(x, u) = E_w \left( c(x, u, w) + \gamma \min_{u' \in U} Q_{N-1}(f(x, u, w), u') \right), \quad \forall N > 1
\]

with \( Q_1(x, u) \equiv E_w(c(x, u, w)) \), converges as \( N \rightarrow \infty \) to the \( Q \)-function, unique solution of the Bellman equation:

\[
Q(x, u) = E_w \left( c(x, u, w) + \gamma \min_{u' \in U} Q(f(x, u, w), u') \right).
\]

- Relations between \( J^\mu_\ast(x) \) and \( Q(x, u) \)

\[
J^\mu_\ast(x) = \min_u Q(x, u)
\]

\[
Q(x, u) = E_w \left( c(x, u, w) + \gamma J^\mu_\ast(f(x, u, w)) \right)
\]
Some dynamic programming results

Optimal and suboptimal policies

- Nec. and suff. optimality condition: $\mu^*(x) \in \arg\min_{u \in U} Q(x, u)$.
- Stationary policy $\mu^*_N$: $\mu^*_N(x) \in \arg\min_{u \in U} Q_N(x, u)$.
- Bound on the subopt. of $\mu^*_N$ w.r.t. $\mu^*$: $\|J^{\mu_N^*} - J^{\mu^*}\|_\infty \leq \frac{2\gamma^N B_c}{(1-\gamma)^2}$

where $J^{\mu_N}(x) = \mathbb{E}_{w_t} \left( \sum_{t=0}^{N-1} \gamma^t c(x_t, \mu_N(x_t), w_t) \mid x_0 = x \right)$

with $0 \leq \gamma < 1$.

(This requires consistency, i.e. $\|J^{\mu_N^*_F} - J^{\mu^*_N}\|_\infty$ arbitrarily small as $\#F \to \infty$)