313-indepth-rl

Apprentissage par renforcement de stratégies de traitement du VIH

Sommaire

- Modèle
  - Thérapie ciblée pour patients atteints du VIH
- Algorithmes
  - Fitted Q-iteration w/ Extra Trees (historical)
  - DQN
  - DQN fed with ReplayBuffer from ExtraTree Fitted Q-iteration
  - PPO
- Limites et perspectives

Plan de la présentation :
https://github.com/SuReLI/hiv_experiment/blob/main/2023/tentative%20de%20plan%20de%20pr%C3%A9sentation.md

Modèle

On modélise le **traitement d'un patient atteint du VIH**.
Deux papiers (2004 et 2006) s'intéressent au **Structured Treatment Interruption (STI)**

D'après le papier de 2004, **l'état du patient** est caractérisé par :

- **6 paramètres,**
- **un ensemble d'équations différentielles** avec 2 états localement stables (healthy / unhealthy).

Le traitement est une combinaison de **2 médicaments**.

Performances de la stratégie : fonction des paramètres du patient et de la quantité de médicament prescrite.

Intérêt du STI :

- *Toxicité des médicaments à court et long terme*
- *Efficacité accrue après reprise du traitement*
- *Traitement plus léger (Non observance = principale cause d' échec du traitement)*
- *Coût*

Modèle (1)

2004 : Adams BM, Banks HT, Kwon HD, Tran HT. “**Dynamic multidrug therapies for hiv: optimal and sti control approaches”,** *doi: 10.3934/mbe.2004.1.223*

2006 : D. Ernst, G. -B. Stan, J. Goncalves and L. Wehenkel, **"Clinical data based optimal STI strategies for HIV: a reinforcement learning approach"**, *doi: 10.1109/CDC.2006.377527.*

- On modélise le traitement d’un patient atteint du VIH
- L’objectif est Structured Treatment Interruption (STI)
  - Toxicité (short term and long term side effects/toxicities)
  - Efficacité
  - Arrêt traitement
- 6 paramètres du patient
- Système d’équations différentielles
- Traitement avec 2 médicaments
  - 4 actions possibles
- Modèle de récompense : maximiser l’efficacité de la réponse immunitaire tout en limitant le recours à des médicaments

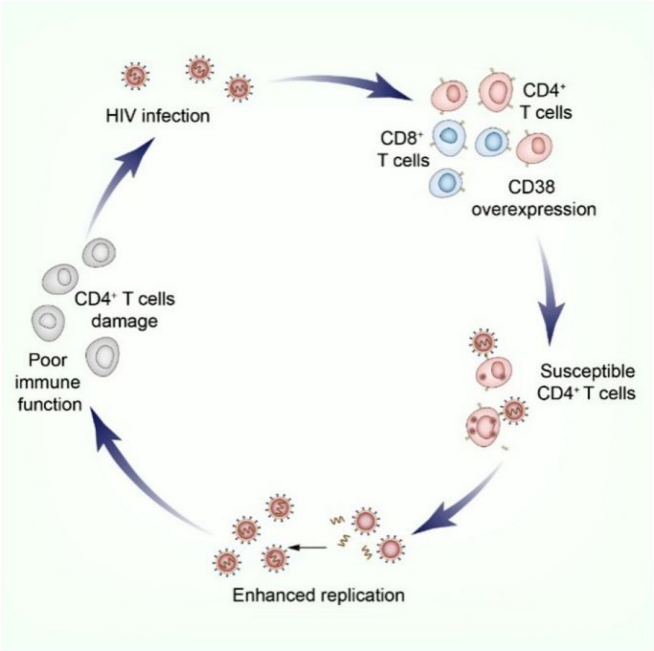
Contexte - HIV

Le VIH s’attaque au système immunitaire

- lymphocytes T auxiliaires (CD4+ T cells), rôle important dans la réponse immunitaire adaptative
- macrophages et cellules dendritiques.

Sans traitement, Syndrome Immunodéficience Acquise (SIDA)

- après 8 ans en moyenne
- décès du patient par maladies opportunistes



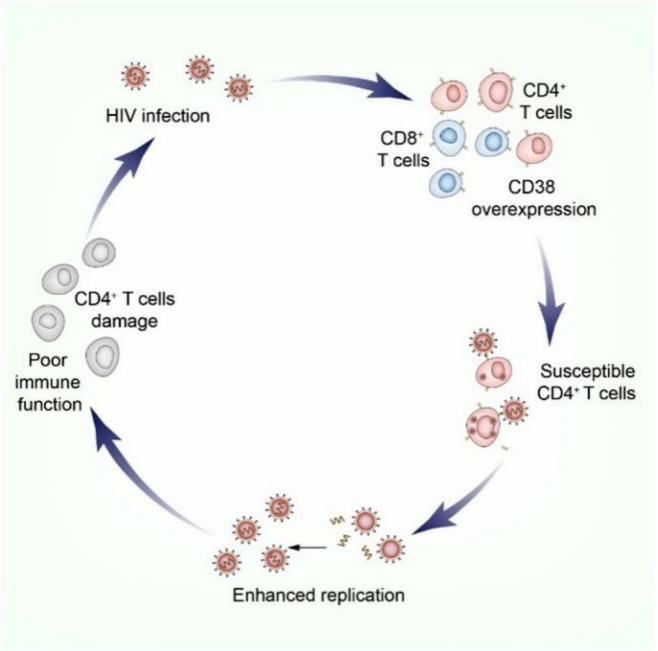
Contexte - HIV

Plusieurs défis lors du traitement

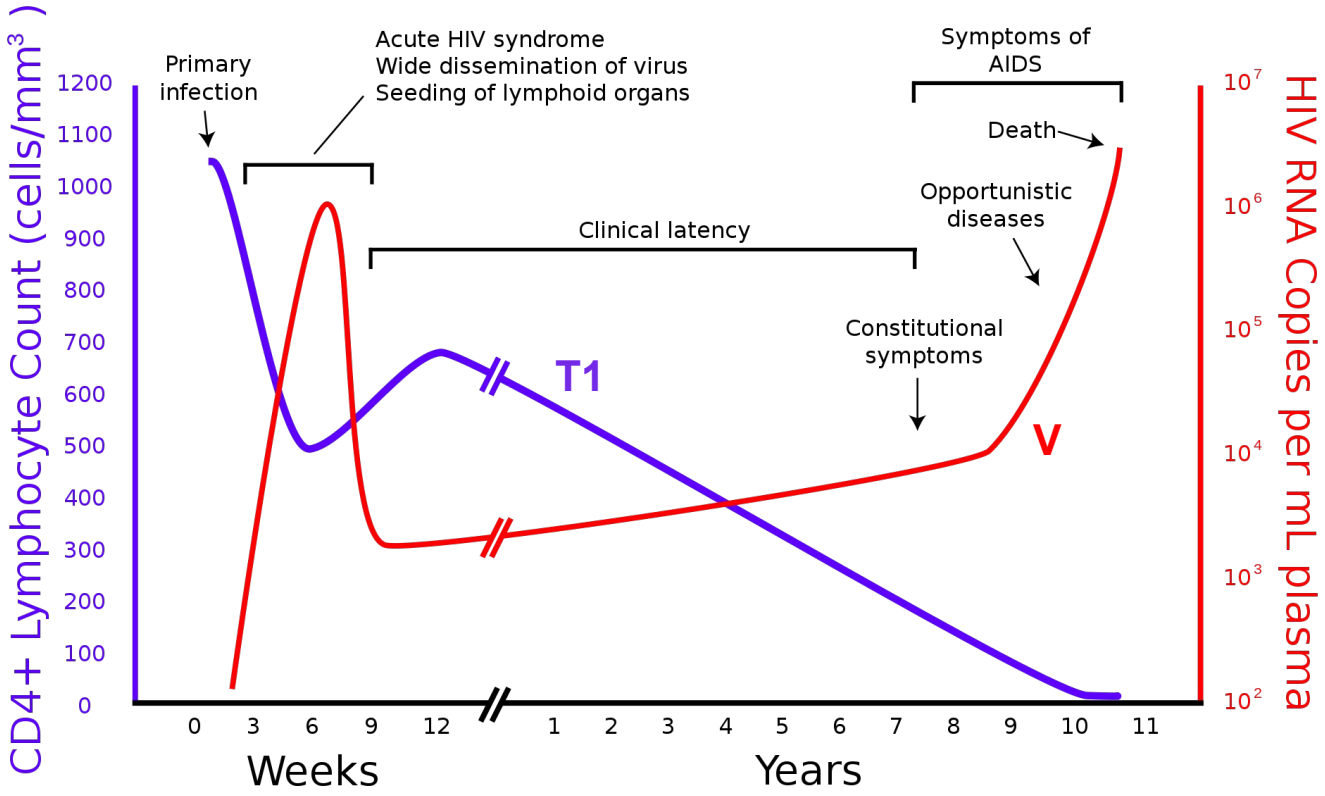
- Empêcher la diminution des CD4+
- Diminuer les risques de mutations

Le traitement repose sur les ART (Antiretroviral Therapy)

- Cocktail de médicaments (habituellement 3)
- Éviter les mutations en diminuant la quantité de virus



Contexte - Évolution VIH sans traitement



AIDS is defined as an HIV infection with either a $CD4^+$ T cell count below **200 cells per µL** or the occurrence of specific diseases associated with HIV infection. *(Mandell, Bennett, and Dolan (2010). Chapter 118.)*

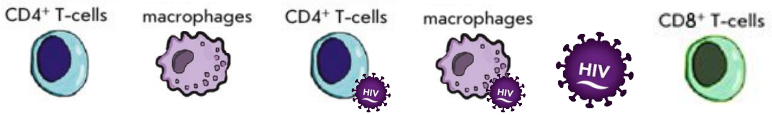
CD4+ normal range : 500 to 1500 cells/mm³ (https://www.ncbi.nlm.nih.gov/books/NBK513289/)

Modèle de la Simulation

Dans le modèle, la stratégie est réévaluée tous les 5 jours.

Définition de l'État du Patient

T1 T2 T1* T2* V E



Définition du Traitement

ϵ_1 RT Inhibitors

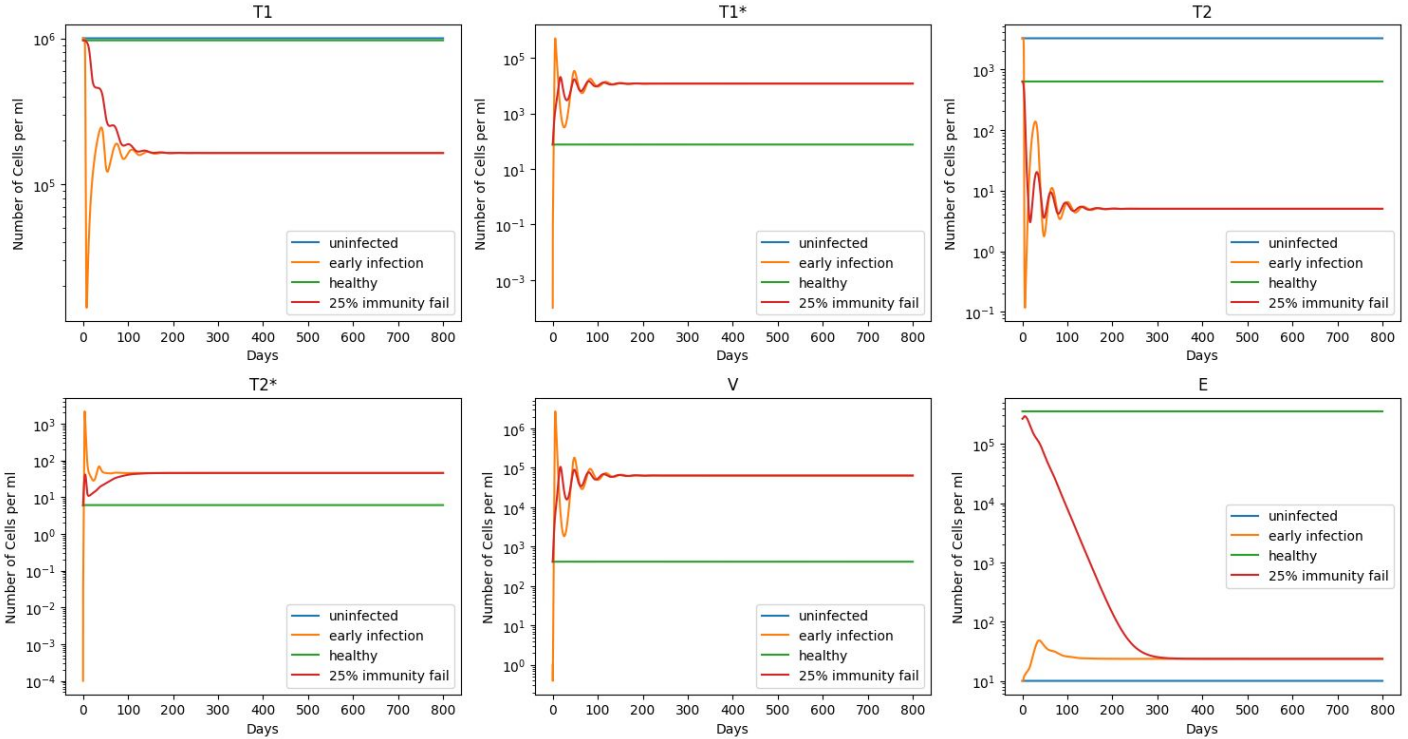
ϵ_2 Protease Inhibitors

Modèle de la Récompense

$$R_t(s,\, a) \;=\; -\left(Q\,V_t \;+\; R_1\,\epsilon_{1_t}^2 \;+ R_2\,\epsilon_{2_t}^2 - S\,E_t\right)$$

Visualisation du Simulateur

On simule ici l'état de 4 patients sans traitement pendant une durée de 800 jours.



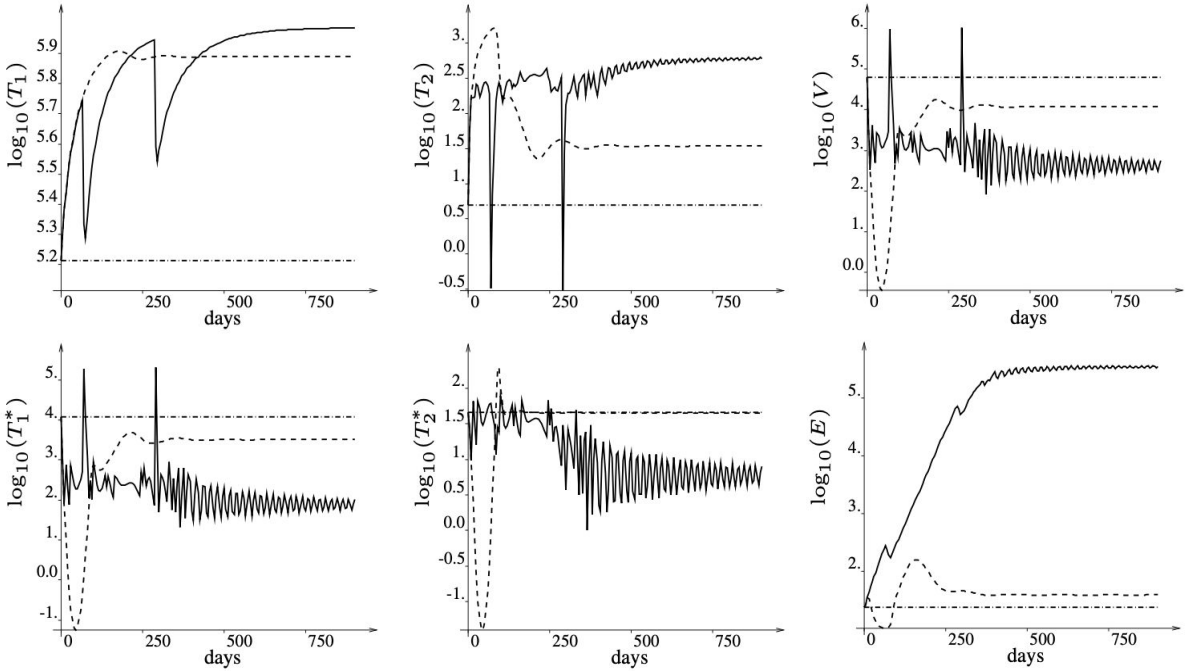
Comparaison entre différents Traitements (*Ernst et al,* 2006)

Ici on compare un traitement pour un patient dans l'état "unhealthy", avec un traitement ininterrompu et pas de traitement. La stratégie obtenue par RL provoque l'interruption du traitement à 2 reprises

Traitement avec 2 interruptions

Sans Traitement

Traitement Ininterrompu

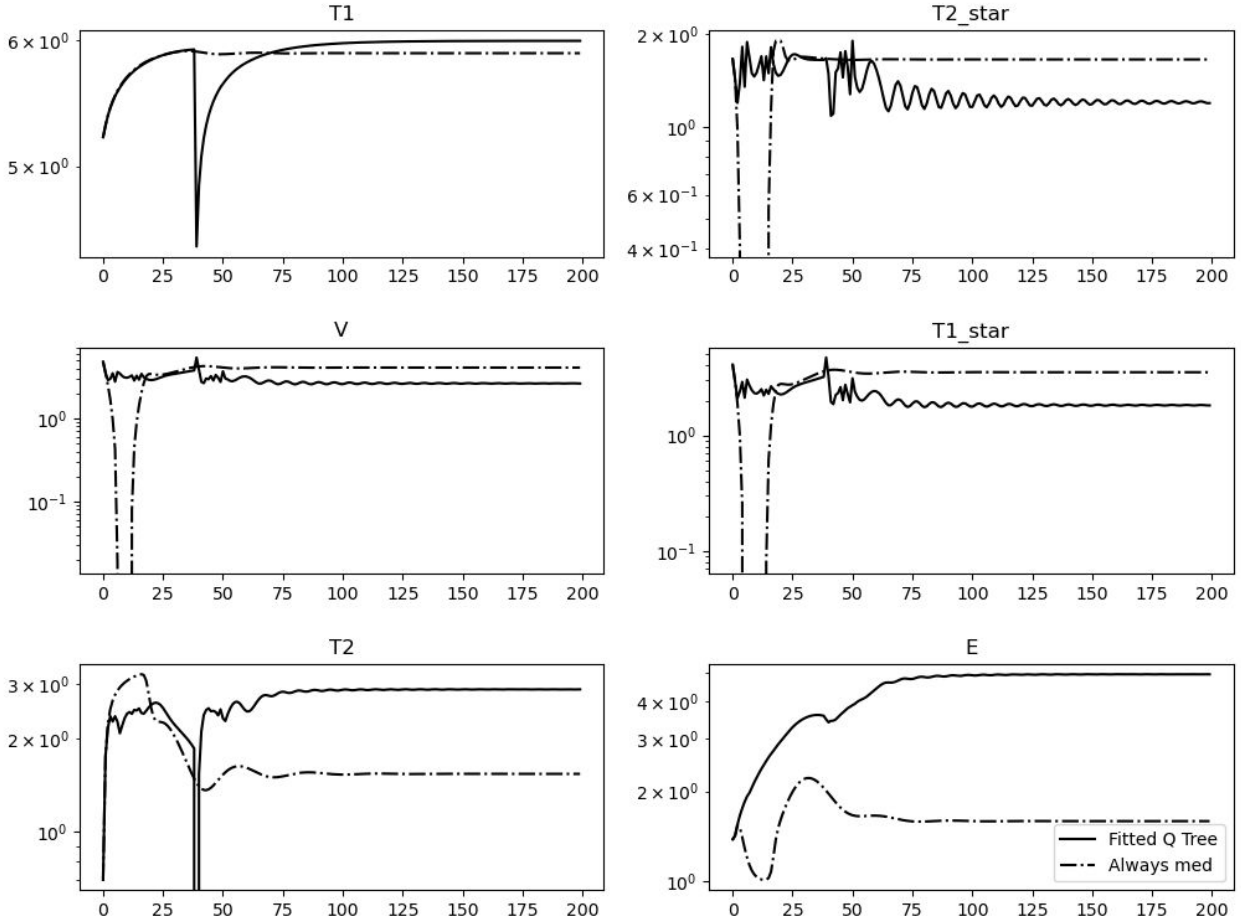


L'approche de *Ernst et al 2006*: Fitted Q-Iteration with Extra-trees

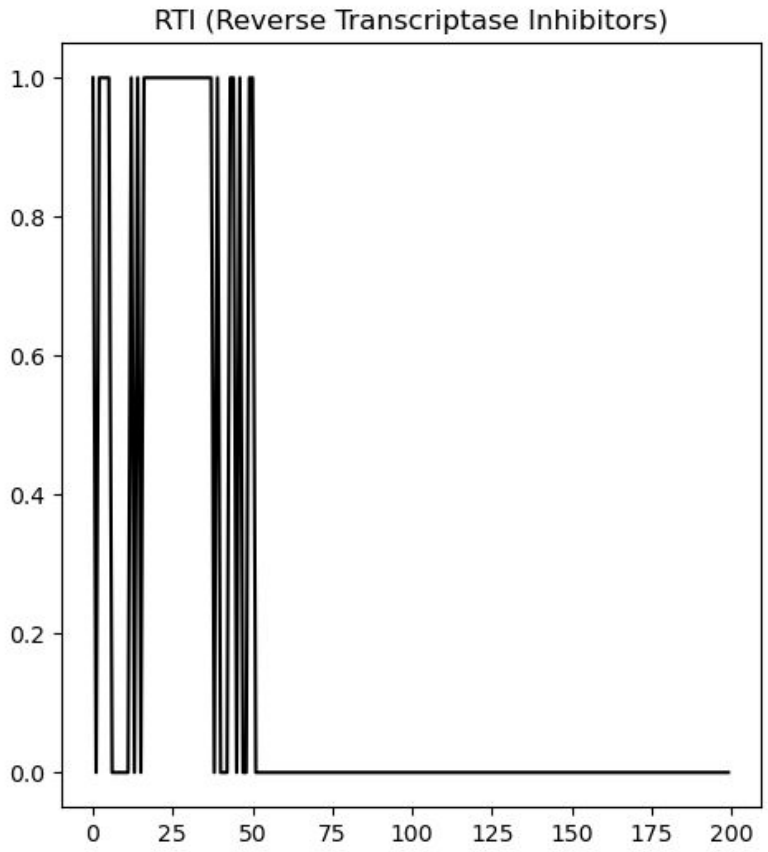
- Utilisation d’un replay buffer initial RB_0 : 30 patients simulés sur 200 itérations (~ 3 ans) avec actions aléatoires → 6000 couples {(s, a) , r}, premier fitting

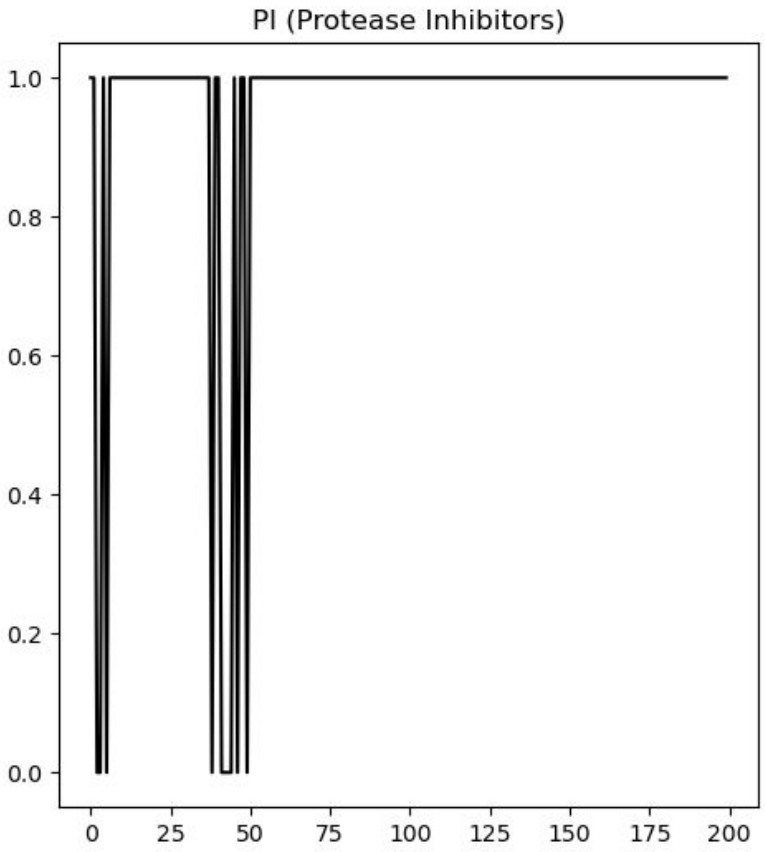
- Fitting itératif replay buffer stacké (nouveaux couples : epsilon greedy (0.15)) → sur 10 epochs

L'approche de *Ernst et al 2006*: Fitted Q-Iteration with Extra-trees



Politique de médicaments utilisée





L'approche de *Ernst et al 2006*: Fitted Q-Iteration with Extra-trees

Limites, améliorations, alternatives…

- Dynamique mal connue
- Définition du problème de contrôle optimal
- Observations partielles

Passage à l’échelle avec un algorithme robuste

L'approche de *Ernst et al 2006*: Fitted Q-Iteration with Extra-trees

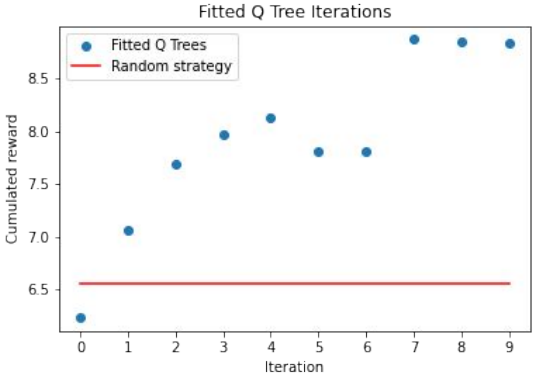
Tentative de visualisation de la robustesse de l’algorithme :

- Approche en réalisant du Domain Randomization
- Entraînement sur 30 patients différents, possédant des paramètres différents
- Paramètres choisis aléatoirement selon une loi uniforme

L'approche de *Ernst et al 2006*: Fitted Q-Iteration with Extra-trees

Training

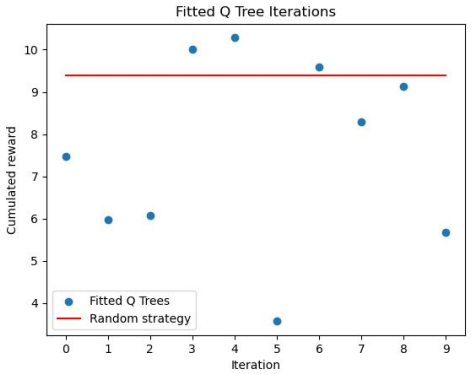
30 Patients Standards



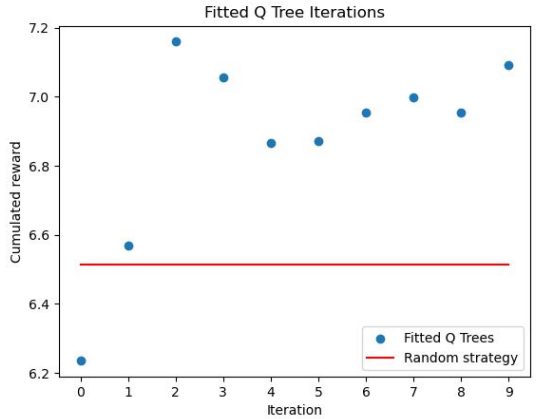
Modalité du test

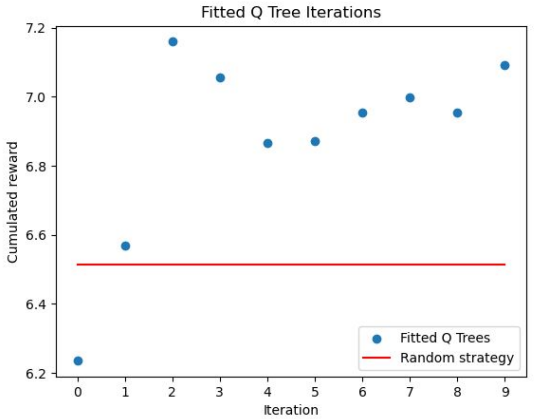
Sur un patient Standard

Sur un patient aux paramètres aléatoires



30 Patients avec paramètres aléatoires





Deep Q-network (DQN)

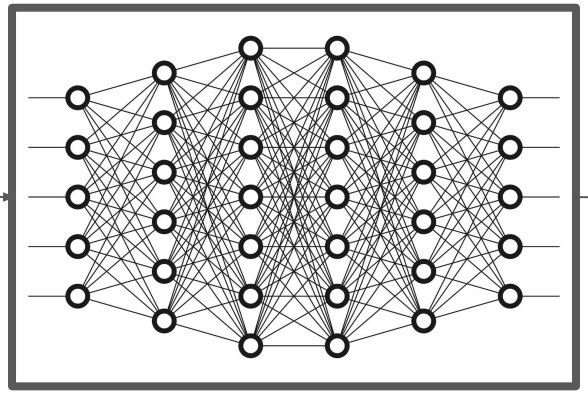
Objectif : Définir une une politique d'action optimale (prise ou non de médicaments) en fonction de l'état du patient à un instant donné.

Idée :

- Utiliser des réseaux de neurones pour apprendre la valeur Q à donner à chaque action en fonction de l'état du patient. L'apprentissage se fait par expérience en simulant des épisodes pour un patient atteint.
- Une fois ce modèle mis au point, la meilleure action à réaliser est celle avec la plus grande valeur.

s = (T1, T1*, T2, T2*, V, E)

État du patient



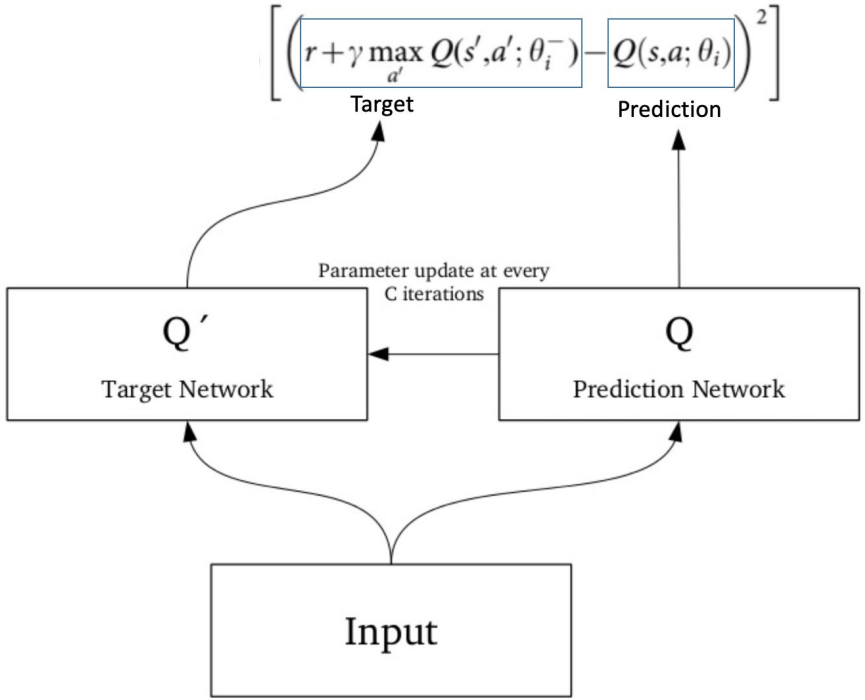
(Q(s,a1), Q(s, a2), ...)

Valeurs des actions a1, a2, … sachant s l'état du patient

Deep Q-network (DQN): Algorithme

Caractéristiques de l'algorithme:

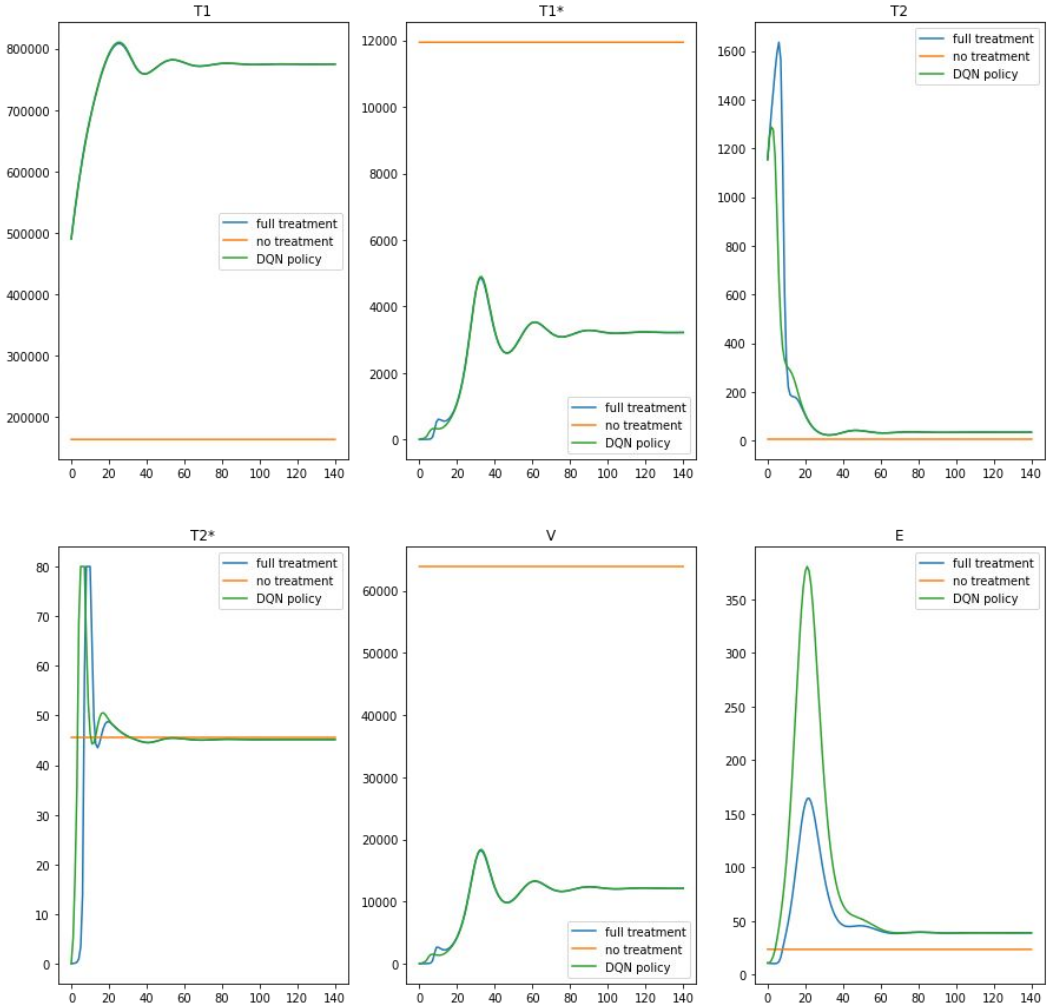
- Exploration d'actions aléatoire
- Deux réseaux appris en parallèle
- Discount factor sur les récompenses



Deep Q-network (DQN): Résultats

Observation 1:

Politique sans interruption de traitement par défaut



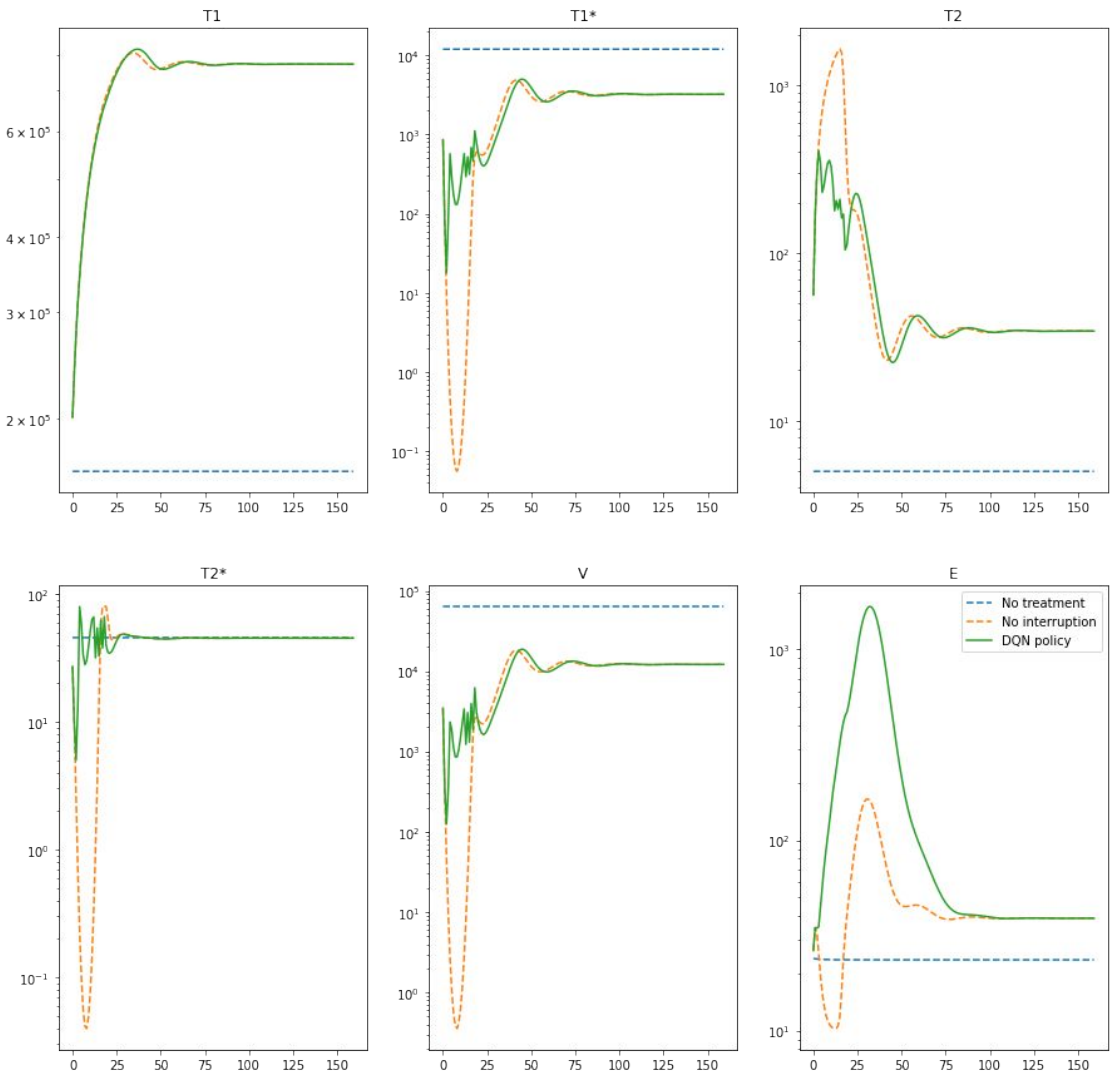
Deep Q-network (DQN): Résultats

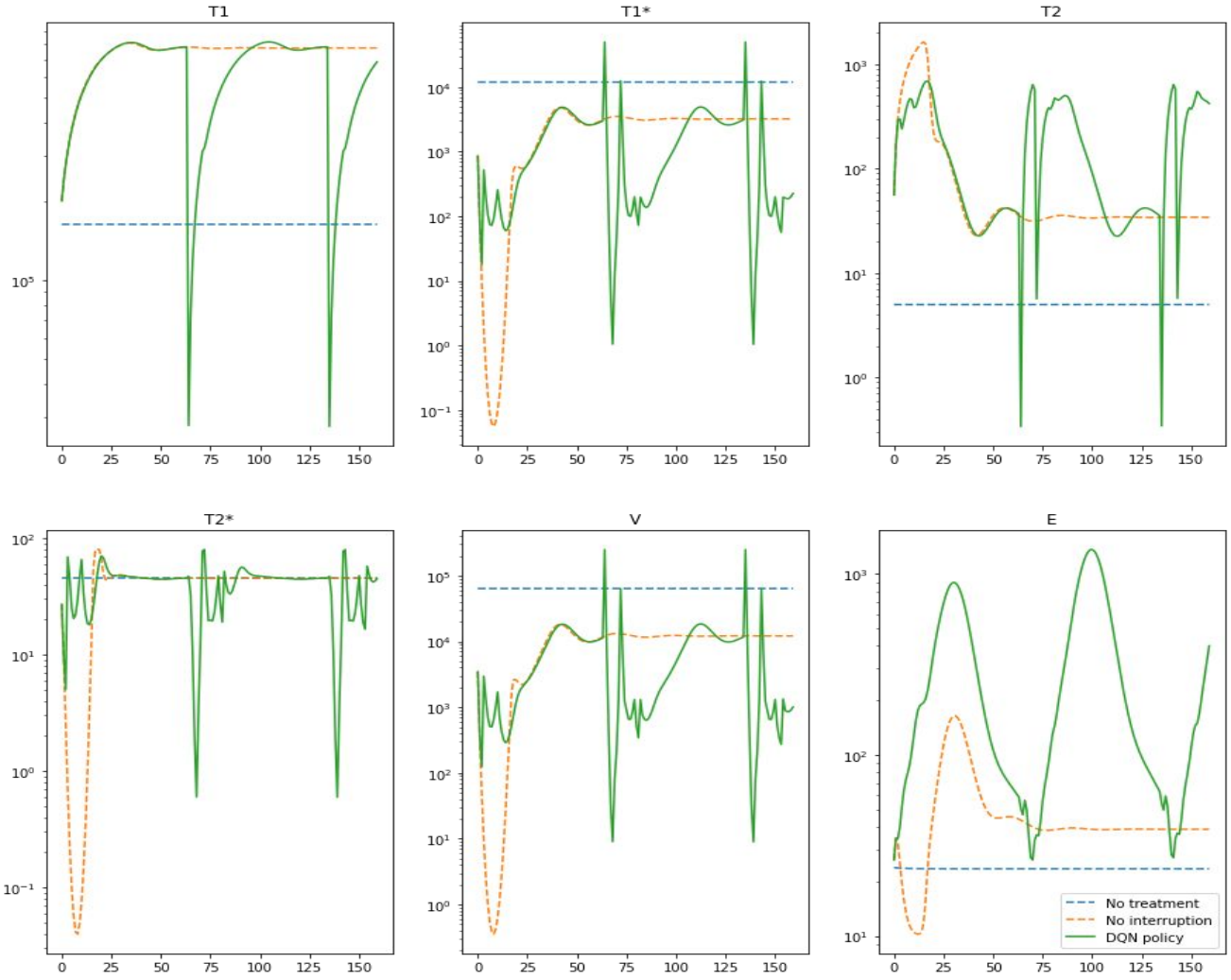
Observation 2:

Nouveaux paramètres.

Interruptions visibles.

Retour à l'état "unhealthy"




T1
T1*
T2
T2*
V
E
No treatment
No interruption
DQN policy

Deep Q-network (DQN): Limitations

Limites, améliorations, alternatives…

Premier plateau atteint assez rapidement : traitement continu

Difficulté d’exploration

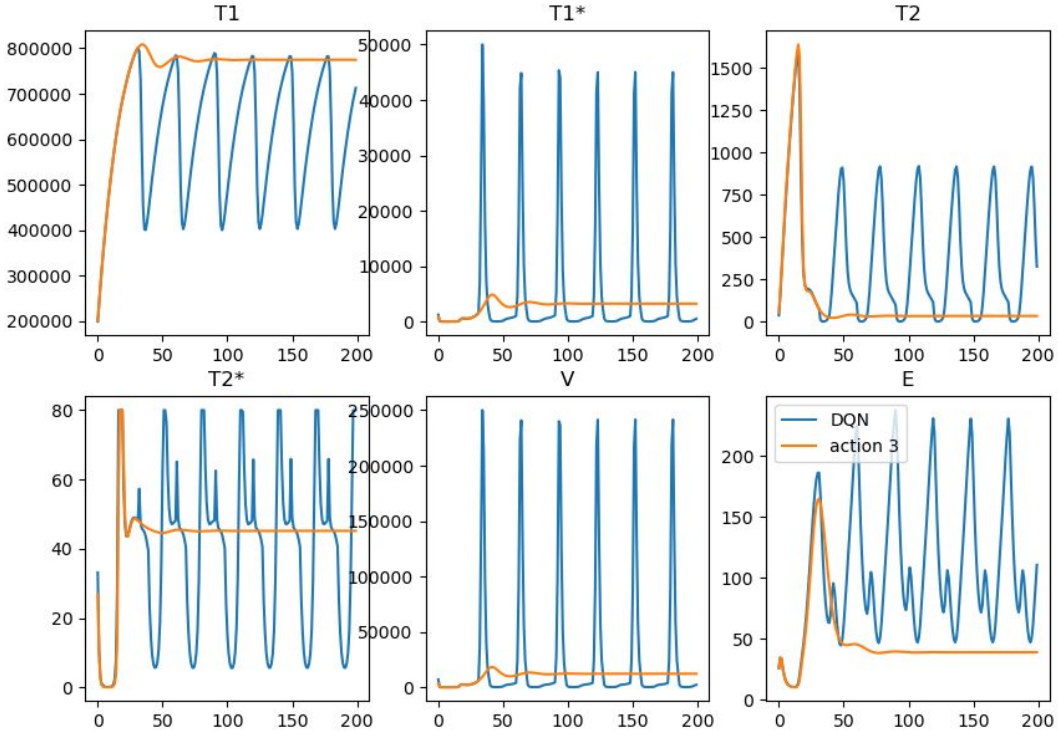
DQN avec ReplayBuffer du Fitted Q-Iteration

Motivations: exploration difficile par DQN, résultats très bons avec Xtrees.

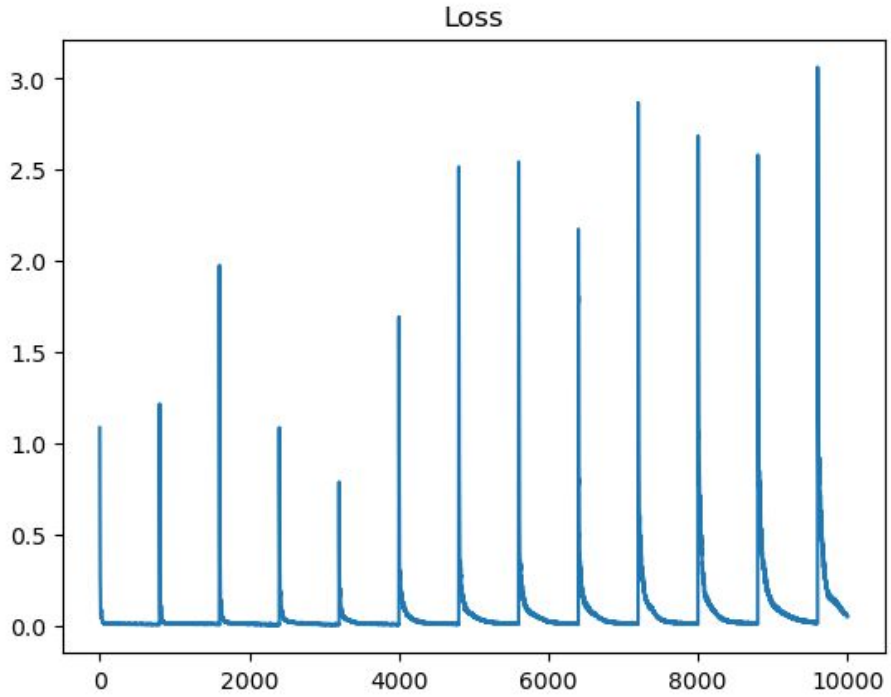
Entraînement offline d’un DQN

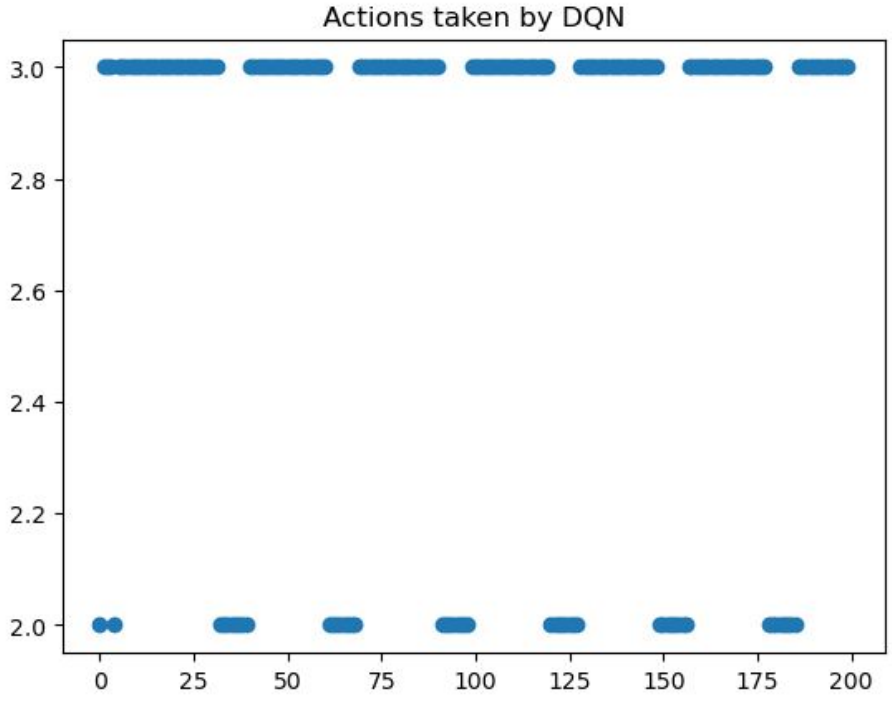
DQN avec ReplayBuffer du Fitted Q-Iteration

Reward : 17 millions



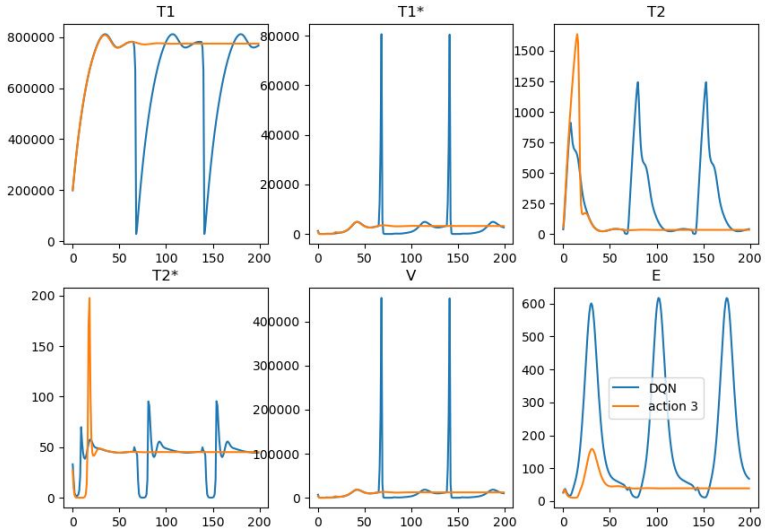
DQN avec ReplayBuffer du Fitted Q-Iteration

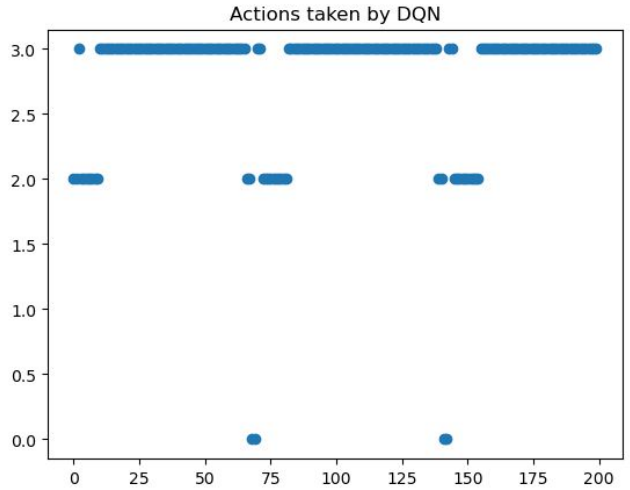




DQN avec ReplayBuffer du Fitted Q-Iteration

20k epochs,





Proximal Policy Optimization : PPO

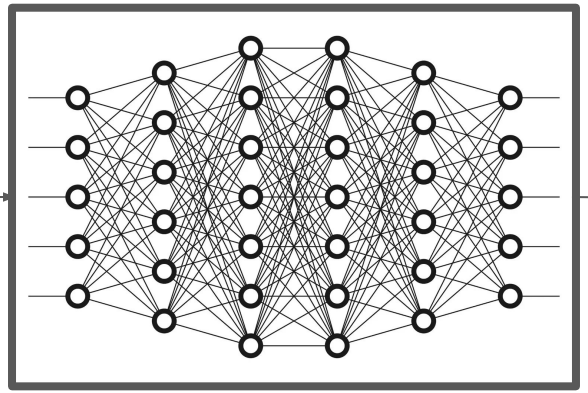
Objectif : Définir une une politique d'action optimale (prise ou non de médicaments) en fonction de l'état du patient à un instant donné.

Idée :

- Utiliser des réseaux de neurones pour apprendre la "valeur" à donner à chaque action en fonction de l'état du patient. L'apprentissage se fait par expérience en simulant des épisodes pour un patient atteint.
- Une fois ce modèle mis au point, la meilleure action à réaliser est celle avec la plus grande valeur.

s = (T1, T1*, T2, T2*, V, E)

État du patient



(Q(s,a1), Q(s, a2), ...)

Valeurs des actions a1, a2, … sachant s l'état du patient

Proximal Policy Optimization : PPO

Algorithme: le réseau de neurones est ici entraîné par descente de gradient (méthode dite de *Policy Gradient)*. Cela est rendu possible grâce au *policy gradient theorem* qui permet de calculer le gradient de la police.

Algorithm 5 PPO with Clipped Objective

Input: initial policy parameters θ_0, clipping threshold ϵ
for $k = 0, 1, 2, \ldots$ **do**
  Collect set of partial trajectories $\mathcal{D}_k$ on policy $\pi_k = \pi(\theta_k)$
  Estimate advantages $\hat{A}_t^{\pi_k}$ using any advantage estimation algorithm
  Compute policy update

$$\theta_{k+1} = \arg\max_{\theta} \mathcal{L}_{\theta_k}^{CLIP}(\theta)$$

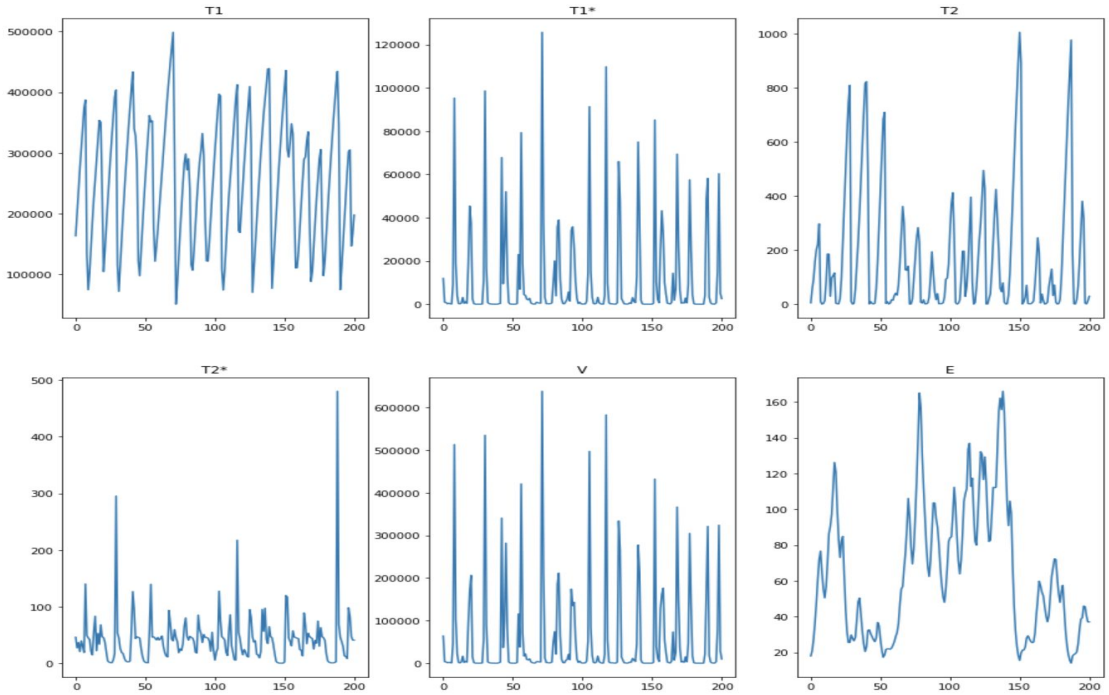
  by taking K steps of minibatch SGD (via Adam), where

$$\mathcal{L}_{\theta_k}^{CLIP}(\theta) = \mathop{\mathrm{E}}_{\tau \sim \pi_k} \left[\sum_{t=0}^{T} \left[\min(r_t(\theta)\hat{A}_t^{\pi_k}, \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t^{\pi_k}) \right] \right]$$

end for

Proximal Policy Optimization : PPO

Résultats : On obtient après beaucoup d'entraînement, on obtient seulement une récompense moyenne de 28M soit ~100 fois moins que le résultat escompté



Proximal Policy Optimization : PPO

Limitations :

Exploration difficile

Calculs lourds -> temps de calcul conséquents(~2h pour 500 epochs)

Le modèle ne converge pas vraiment, en moyenne la récompense augmente au fil des epochs mais forte retombée de temps en temps.

Les résultats dépendent et de l'état initial, et de la répartition des récompenses

Limites du Modèle global

- 1 Patient avec paramètres très spécifiques (Traitement Spécifique)
  - Robustesse de la méthode
- 4 Actions de Traitement (Discrètes, avec 2 médicaments)
  - Habituellement trithérapie
- Pas de Bruit / Incertitude
  - Dans le simulateur, le patient va réagir de manière déterministe au traitement - pas le cas dans la vrai vie
  - Pas d'étude de l'influence du bruit sur la stratégie trouvée.
- Pertinence du modèle de récompense ?

Bibliographie:

2004 : Adams BM, Banks HT, Kwon HD, Tran HT. “**Dynamic multidrug therapies for hiv: optimal and sti control approaches”,** *doi: 10.3934/mbe.2004.1.223*

2006 : D. Ernst, G. -B. Stan, J. Goncalves and L. Wehenkel, **"Clinical data based optimal STI strategies for HIV: a reinforcement learning approach"***, doi: 10.1109/CDC.2006.377527.*