Construction de récompenses par apprentissage par préférences pour les modèles d'apprentissage par renforcement appliqués aux stratégies de traitements adaptatifs

Sophia Yazzourh[1] , Nicolas Savy [1], Philippe Saint-Pierre [1] et Michael Kosorok[2]

[1] *Institut de Mathématiques de Toulouse; UMR5219 - Université de Toulouse ; CNRS - UPS IMT,*
F-31062 Toulouse Cedex 9, France
[2] *Department of Biostatistics, University of North Carolina at Chapel Hill,*
Chapel Hill, NC 27599, U.S.A.

Résumé. Délivrer des traitements personnalisés à chaque étape des maladies chroniques est un objectif clé de la médecine de précision qui est formalisé par les "Dynamic Treatment Regimes". Ce cadre adapte les stratégies de traitement en se basant sur des règles de décision apprises à partir d'essais cliniques pour améliorer l'efficacité du traitement. L'utilisation de l'apprentissage par renforcement aide à déterminer ces règles en se basant sur leurs données individuelles et de leurs historiques médicaux. L'apprentissage de la stratégie de traitement repose sur des évaluations quantitatives du système appelées récompenses. Traditionnellement, ces récompenses sont déterminées par des experts qui sélectionnent une variable d'intérêt, mais qui peut être trop restrictive pour l'apprentissage de la règle de décision. Nous proposons une approche automatique et généralisée pour construire les récompenses, en utilisant l'apprentissage par préférences ou "Preference Learning".

Mots-clés. Apprentissage par renforcement; Stratégies de traitements adaptatifs; Médecine de précision; Apprentissage par préférences; Récompenses;

Abstract. Delivering personalized treatments at each stage of chronic diseases is a key goal of precision medicine, formalized by "Dynamic Treatment Regimes." This framework adjusts treatment strategies based on decision rules learned from clinical trials to enhance treatment effectiveness. Reinforcement learning helps determine these rules by using individual patient data and medical histories. Learning treatment strategy relies on quantitative system evaluations called rewards. Traditionally, experts select these rewards based on a variable of interest, which may be too restrictive for learning the decision rule. We propose an automatic and generalized approach to construct rewards using preference learning.

Keywords. Reinforcement Learning; Dynamic Treatment Regimes; Precision Medicine; Preference Learning; Rewards;

1 Introduction

La médecine moderne cherche à améliorer sa capacité à traiter de manière personnalisée chaque patient. C'est notamment la médecine de précision [Kosorok and Laber, 2019] qui initie une réflexion profonde sur cette question. Son objectif est d'améliorer la qualité des soins de santé en ajustant l'approche médicale selon l'état de santé spécifique et changeant de chaque patient. L'hétérogénéité des caractéristiques et des réactions parmi les populations de patients exigent des approches de traitement variées. Initialement, des modèles statistiques ont été implémentés pour répondre à cette problématique. Avec l'avènement du stockage de données et de la puissance de calcul, les méthodes d'apprentissage automatique ont également commencé à être appliquées comme le montre Yu et al. [2021] et Coronato et al. [2020].

Dans ce contexte, la médecine moderne s'intéresse de plus en plus à l'adaptation des traitements aux données individuelles des patients. La médecine de précision vise à améliorer la santé en mettant au coeur de la décision les informations spécifiques à chaque patient. Cette approche qui cherche à recommander des traitements personnalisés est appelée "Dynamic Treamtent Regimes" (DTR) et est explicitée dans Kosorok and Laber [2019] et Chakraborty and Murphy [2014]. Les DTR se basent sur l'historique médical et les réponses des patients aux traitements précédents.

Au fil des décennies, l'apprentissage automatique est devenue une solution incontournable pour résoudre des problèmes complexes à grande échelle. Dans le domaine du support à la décision, particulier pour les scénarios séquentiels, l'apprentissage par renforcement ou "Reinforcement Learning" (RL) [Sutton and Barto, 2018] se révèle être une approche particulièrement efficace. Ces méthodes s'adaptent aux conditions changeantes et optimisent les décisions sur plusieurs étapes. Elles répondent à des questions de processus de prise de décisions dynamiques. L'idée principale est de trouver une règle de décision, appelée "policy", visant à optimiser un objectif à long terme, en prenant des décisions successives pour maximiser le bénéfice global. Cet objectif d'optimisation est basé sur des récompenses qui sont des indications quantitatives sur l'état du patient. Leur construction ou formulation est alors cruciale dans l'apprentissage de la prise de décision.

De manière générale, les récompenses sont conçues par un expert du système qui propose de l'évaluer au travers d'un score. Dans le contexte des essais cliniques visant à aider les personnes obèses à perdre du poids, une récompense pourrait consister à mesurer leur indice de masse corporelle (IMC) [Linn et al., 2015]. Dans le contexte des soins critiques, un autre exemple serait d'évaluer les traitements en fonction du taux de survie ou de mortalité des patients [Roggeveen et al., 2021]. Certaines récompenses peuvent être conçus de manière plus subtile en faisant des compromis et des combinaisons de variables. Dans le cadre d'une simulation de cancer [Zhao et al., 2009], les récompenses sont évaluées en prenant en compte la taille de la tumeur, de la toxicité du traitement, du bien être du patient et du taux de survie.

Construire de manière manuelle une fonction récompense peut impliquer des choix arbitraires, voire très spécifiques au contexte mais également mener à des objectifs d'apprentissages trop restrictifs. Dans la simulation de [Zhao et al., 2009], lorsque le décès d'un patient survient, un score de -60 est arbitrairement attribué à cet événement. Une des manière de

généraliser et de construire automatiquement des récompenses est d'utiliser l'apprentissage par préférences ou "Preference Learning" (PL) [Fürnkranz et al., 2012]. L'approche repose sur l'expertise médicale, où le médecin exprime ses préférences quant aux trajectoires ou aux suivis médicaux des patients. Ces informations de comparaison deux à deux seront ensuite utilisées au travers d'un modèle probabiliste comme celui de Bradley-Terry [Akrour et al., 2012] pour construire par maximum de vraisemblance les récompenses.

Dans la suite nous proposons un aperçu de l'application du RL à l'optimisation des séquences de traitement et une amélioration de la construction des récompenses par apprentissage par préférences.

2 Apprentissage par renforcement

2.1 Processus de Décision

Ce contexte de modélisation est celui des processus de décision ou "Decision Process" (DP) qui sont le cadre initial des DTR. Un DP est un système dynamique au cours du temps $t \in \mathbb{T}$ qui navigue dans un espace d'états $\mathbb{S}$ contenant les covariables décrivant l'état du patient. Les possibles actions sont contenues dans l'espace $\mathbb{A}$ et représentent les traitements et leurs dosages associés. Le sous-ensemble mesurable non vides de $\mathbb{A}$, notée $\{\mathbb{A}(s)|s \in \mathbb{S}\}$, contient les actions réalisables qui peuvent être prises lorsque le système se trouve dans un état spécifique $s \in \mathbb{S}$. En d'autres termes tous les traitements ne sont pas admissibles ou disponibles pour un patient à une étape donnée. Le formalisme mathématiques est donné par la définition suivante :

Definition 2.1 (Processus de décision ou "Decision Process" (DP))**.** Un processus de décision $(S, A, \{\mathbb{A}(s)|s \in \mathbb{S}\}, \nu)$ sur $\mathbb{T}$ contient:

- une famille S de variables aléatoires à valeurs dans $\mathbb{S}$: $\{S_t, t \in \mathbb{T}\}$, où $\mathbb{S}$ est appelé espace d'états.
- une famille A de variables aléatoires à valeurs dans $\mathbb{A}$: $\{A_t, t \in \mathbb{T}\}$, où $\mathbb{A}$ est appelé espace d'actions.
- une famille $\{\mathbb{A}(s)|s \in \mathbb{S}\}$ des sous-ensembles mesurables non vides de $\mathbb{A}$, ensemble des actions réalisables lorsque le système se trouve dans l'état $s \in \mathbb{S}$. Cela requière à $\mathbb{K} = \{(s, a)|s \in \mathbb{S}, a \in \mathbb{A}(s)\}$ d'être un sous ensemble mesurable de $\mathbb{S} \times \mathbb{A}$.
- une distribution initiale ν sur $\mathbb{S}$.

Definition 2.2 (Histoires admissibles)**.** Pour tout $n \in \mathbb{N}$, une histoire admissible au temps n est un vecteur qui contient les états parcourus par le système ainsi que les actions prises jusqu'au temps n. L'ensemble des histoires admissibles au temps n est noté:

$$\mathbb{H}_0 = \mathbb{S} \qquad \mathbb{H}_n = \mathbb{K}^{n-1} \times \mathbb{S} \tag{1}$$

Un élément $h_n \in \mathbb{H}_n$ s'écrit $(s_0, a_0, \ldots, s_{n-1}, a_{n-1}, s_n)$ où pour tout $0 \leq j \leq n-1$, $(s_j, a_j) \in \mathbb{K}$.

Les histoires admissibles h_n observées sont les trajectoires de soins de différents patients et décrivent les informations des traitements et des covariables à chaque étape.

L'aspect principal à considérer dans le traitement du processus de décision est de déterminer la probabilité d'atteindre l'état s_{n_1} au temps $n+1$ étant donné l'historique et les décisions jusqu'au temps n. On l'exprime comme suit :

$$\mathbb{P}_\nu \left[S_{n+1} = s_{n+1} \mid H_n = h_n, A_n = a_n \right].$$

En pratique, le calcul de ces probabilités demande des ressources computationnelles significatives en raison de l'augmentation de la longueur du vecteur h_n, proportionnellement à n. Travailler directement avec une telle variable devient rapidement inabordable (généralement lorsque $n \geq 4$). De manière classique, cette difficulté est surmontée par l'hypothèse markovienne. Ainsi, le formalisme traditionnel du RL est souvent celui des processus de décision markovien.

2.2 Stratégie

Un des principaux concepts du RL est celui de stratégie, politique ou "policy". Elle correspond à la stratégie ou règle de décision médicale personnalisée que les DTR cherchent à déterminer à chaque étape d'intervention. Un seul traitement est administré par étape.

Definition 2.3. Une stratégie est une séquence $\pi = (\pi_n)_{n \in \mathbb{N}}$ de distributions conditionnelles de $\mathbb{A}$ sachant $\mathbb{H}_n$ définit, pour tout $\mathcal{A} \in \mathcal{B}(\mathbb{A})$ et pour tout $h_n \in \mathbb{H}_n$, par:

$$\pi_n(\mathcal{A}, h_n) = \mathbb{P}\left[A_n \in \mathcal{A} \mid H_n = h_n \right],$$

satisfaisant pour tout $n \in \mathbb{N}$, pour tout $h_n \in \mathbb{H}_n$:

$$\pi_n(\mathbb{A}(s_n), h_n) = 1,$$

et pour tout $n \in \mathbb{N}$, pour tout $h_n \in \mathbb{H}_n$ et pour tout $a_n \in \mathbb{A}(s_n)$

$$\pi_n(a_n, h_n) > 0.$$

La prise de décision implique la sélection d'une option en fonction des informations environnementales. Une politique représente un plan stratégique aligné sur un objectif spécifique définissant une séquence d'actions. Une stratégie π suggère une action a_n pour chaque état possible s_n en tenant compte de l'historique h_n.

2.3 Récompense et optimalité

L'objectif qui guide la stratégie est formalisé au travers d'un critère d'optimalité appelé récompense ou "reward". Dans un context d'optimisation de traitement, les récompenses

sont calibrées pour répondre à un objectif médical. Elles peuvent être déterminées par l'expertise des médecins ou selon des objectifs précis.

Definition 2.4. Les récompenses $\{R_n, n \in \mathbb{N}\}$ sont les éléments d'une famille de variables aléatoires bornées dans $\mathbb{R}$.

Afin de pouvoir évaluer et comparer les stratégies, deux mesures quantitatives doivent êtres introduites :

Definition 2.5 (Fonctions valeurs [Chakraborty and Murphy, 2014])**.** Étant donné un processus de décision $(S, A, \{\mathbb{A}(s)|s \in \mathbb{S}\}, \nu)$ sur $[0, \tau]$, $\{R_n, n \in \mathbb{N}\}$ une famille de récompenses et π une stratégie :

- A l'étape n la V-fonction ou "state-value function" est l'espérance de la somme des récompenses attendues depuis l'étape n sachant l'histoire h_n:

$$V_n^\pi(h_n) = \mathbb{E}_\nu^\pi \left[\sum_{j=n+1}^{\tau} R_j \mid H_n = h_n \right]. \tag{2}$$

- A l'étape n la Q-fonction ou "action-value function" est l'espérance de la somme des récompenses attendues depuis l'étape n sachant l'histoire h_n et l'action prise a_n:

$$Q_n^\pi(h_n, a_n) = \mathbb{E}_\nu^\pi \left[\sum_{j=n+1}^{\tau} R_j \mid H_n = h_n, A_n = a_n \right]. \tag{3}$$

L'objectif du RL est de déterminer les stratégies optimales notées π^* qui sont les politiques qui maximisent la somme des récompenses. En d'autres mots, celles qui maximisent le gain à long terme. Les stratégies optimales peuvent alors être déterminées en évaluant les fonctions valeurs optimales.

Theorem 2.1. Les stratégies optimales $\pi*$ sont les stratégies qui maximisent les fonctions valeurs pour tout $n \in \mathbb{N}$, pour tout $h_n \in \mathbb{H}_n$ et $a_n \in \mathbb{A}$, tel que

$$V_n^{\pi^*}(h_n) = V_n^*(h_n) = \max_\pi \; V_n^\pi(h_n) \qquad \text{et} \qquad Q_n^{\pi^*}(h_n, a_n) = Q_n^*(h_n, a_n) = \max_\pi \; Q_n^\pi(h_n, a_n).$$

3 Construction de récompenses par apprentissage par préférences

3.1 Apprentissage par préférences

L'apprentissage par préférences ou "Preference Learning" (PL) [Fürnkranz et al., 2012] offre une alternative à la construction de récompenses. Dans le cadre de l'optimisation des séquences de traitement, les récompenses sont ajustées pour s'aligner sur les objectifs médicaux. Cependant, cette approche se limite parfois à la maximisation d'une seule mesure

quantitative, ce qui peut être trop restrictif pour l'apprentissage. Le PL propose de construire des récompenses $\{R_n, n \in \mathbb{N}\}$ en se basant sur l'avis d'un expert. L'approche consiste à solliciter un avis médical en demandant la préférence du médecin entre deux éléments spécifiques. Ensuite, un modèle probabiliste permets de construire les récompenses qui seront par la suite incorporées dans une méthode d'apprentissage par renforcement classique afin de déterminer les stratégies $\pi*$ qui les maximisent.

La préférence ou comparaison des éléments deux à deux peut porter sur des trajectoires de patients h_n, des stratégies de traitements π, des états s_n ou des actions a_n. Dans notre modèle, nous nous intéressons aux préférences sur les histoires admissibles ou trajectoires de patients h_n. Une fois les préférences exprimées par l'expert, une relation d'ordre entre les éléments est établie. Puis le modèle probabiliste utilisé est celui de Bradley-Terry [Akrour et al., 2012], afin de comparer des trajectoires deux à deux, adapté ici aux DTR. Si l'on considère des préférences sur des trajectoires tel que $h^i \prec h^j$, notre objectif est de maximiser la probabilité suivante :

$$\mathbb{P}(h^i \prec h^j|\theta) = \frac{e^{\alpha(R_\theta(h^i)-R_\theta(h^j))}}{1+e^{\alpha(R_\theta(h^i)-R_\theta(h^j))}}.$$

On calcule par maximum de vraisemblance les récompenses paramétrées R_θ .

3.2 Application

On se place dans un contexte applicatif de simulation d'un cancer traité par chimiothérapie à 4 facteurs [Zhao et al., 2009] : (1) la croissance tumorale en l'absence de chimiothérapie ; (2) les résultats négatifs sur le bien-être des patients en réponse à la chimiothérapie ; (3) la capacité du médicament à tuer les cellules tumorales tout en augmentant la toxicité ; et (4) une interaction entre les cellules tumorales et le bien-être du patient. Pour chaque patient on a deux variables d'états $S_t = \{Y_t, X_t\}$ avec Y la taille de la tumeur et X la toxicité du traitement à chaque mois t tel que $t = 0, 1, 2, 3$. Le traitement A_t administré au mois t est un dosage compris entre 0 et 1 avec un pas de 0,1. Ce modèle se base sur le système d'équations différentielles suivant :

$$\begin{aligned}\Delta Y_t &= [0,15 \times \max(X_t, X_0) - 1,2 \times (A_t - 0,5)] \times \mathbb{1}(Y_t > 0)\\ \Delta X_t &= 0,1 \times \max(Y_t, Y_0) + 1,2 \times (A_t - 0,5)\end{aligned}$$

En utilisant l'indicatrice $\mathbb{1}(Y_t > 0)$, le modèle attribue le statut de rémission totale à un patient lorsque la taille de sa tumeur est réduite à zéro, indiquant ainsi l'absence de récidive.

La possibilité de décès d'un patient pendant un traitement est représentée par un modèle de survie. Pour chaque intervalle de temps $(t-1, t]$, le taux de survie est défini comme une fonction de la taille de la tumeur et de la toxicité : $\lambda(t) = \exp(-4+Y_t+X_t)$. Dans ce modèle, la taille de la tumeur et la toxicité ont une influence tout aussi importante sur la survie du patient. La probabilité de décès du patient pendant l'intervalle de temps $(t-1, t]$ est :

$$\mathbb{P}_{décès} = 1 - \exp(-\int_{t-1}^{t} \lambda(x)dx)$$

Dans le modèle proposé par Zhao et al. [2009], les récompenses sont accordées en fonction du statut de survie, du bien-être du patient et de la taille de la tumeur. On notera que si le patient décède, la "récompense" était de -60. Ce qui corresponds à un choix arbitraire. Notre projet est de construire ces récompenses en se basant sur les préférences d'un expert.

On se base sur le modèle de Fürnkranz et al. [2012], qui propose d'exprimer des préférences sur les trajectoires de la manière suivante :

- Si le patient h_j survit plus longtemps que le patient h_i : $h^i \prec h^j$
- Deux patients décédés au même temps t ne peuvent pas être comparées
- Si les patients survivent dans les deux trajectoires, on note T^i et T^j les toxicités maximales sur la durée de la trajectoire pour les patients i et j respectivement et D^i et D^j la taille de leurs tumeurs à la fin. Les trajectoires sont alors comparée par la dominance de Pareto suivante :
$$h^i \prec h^j \Leftrightarrow (T^j \leq T^i) \text{ et } (D^j \leq D^i)$$

4 Conclusion

La conception directe des récompenses nécessite une compréhension approfondie du processus de traitement médical. En revanche, l'apprentissage par préférences offre une alternative à leur création, permettant d'intégrer davantage d'informations d'évaluation que simplement une variable quantitative.

Les récompenses obtenues grâce à notre modèle probabiliste de comparaison de trajectoires reposent sur un modèle de régression. Ce modèle, facilement interprétable, nous permettra de mettre en évidence la sélection pondérée des covariables de l'espace d'état. Nous pourrons alors comparer les récompenses obtenues par l'apprentissage par préférences avec celles construites manuellement. Finalement, nous étudierons et comparerons les deux règles de décision respectives obtenues par apprentissage par renforcement.

References

Michael R Kosorok and Eric B Laber. Precision medicine. *Annual review of statistics and its application*, 6:263–286, 2019.

Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

Antonio Coronato, Muddasar Naeem, Giuseppe De Pietro, and Giovanni Paragliola. Reinforcement learning for intelligent healthcare applications: A survey. *Artificial Intelligence in Medicine*, 109:101964, 2020.

Bibhas Chakraborty and Susan A Murphy. Dynamic treatment regimes. *Annual review of statistics and its application*, 1:447–464, 2014.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Kristin A Linn, Eric B Laber, and Leonard A Stefanski. iqlearn: Interactive q-learning in r. *Journal of statistical software*, 64(1), 2015.

Luca Roggeveen, Ali El Hassouni, Jonas Ahrendt, Tingjie Guo, Lucas Fleuren, Patrick Thoral, Armand RJ Girbes, Mark Hoogendoorn, and Paul WG Elbers. Transatlantic transferability of a new reinforcement learning model for optimizing haemodynamic treatment for critically ill patients with sepsis. *Artificial Intelligence in Medicine*, 112:102003, 2021.

Yufan Zhao, Michael R Kosorok, and Donglin Zeng. Reinforcement learning design for cancer clinical trials. *Statistics in medicine*, 28(26):3294–3315, 2009.

Johannes Fürnkranz, Eyke Hüllermeier, Weiwei Cheng, and Sang-Hyeun Park. Preference-based reinforcement learning: a formal framework and a policy iteration algorithm. *Machine learning*, 89:123–156, 2012.

Riad Akrour, Marc Schoenauer, and Michèle Sebag. April: Active preference learning-based reinforcement learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, pages 116–131. Springer, 2012.