# Parcours éducatif optimal d'un patient: étude par simulation d'algorithmes adaptatifs

Xavier Goblet

Jeolis Solutions, 63000 Clermont-Ferrand, France
xavier.goblet@lojelis.com

**Résumé.** Dans le cadre d'une éducation thérapeutique digitalisée où l'aspect motivationnel est prépondérant, une personnalisation du parcours patient doit se rapprocher d'une trajectoire optimale dans un large espace d'activités ludiques, caractérisées par un niveau de difficulté et des compétences à maîtriser. La machine doit apprendre des succès ou échecs pour faire dynamiquement des recommandations des futures étapes et offrir l'expérience la plus adaptée à chaque patient. En abordant cette problématique sous l'angle des algorithmes adaptatifs, nous proposons une approche originale. Nous étudions deux familles d'algorithmes (règles logiques et bandits Thompson Sampling) en comparant leurs performances à travers un environnement de simulation. Les premiers résultats montrent un avantage pour le bandit TS quelles que soient les caractéristiques d'apprentissage du patient.

## 1 Introduction

Selon l'OMS, l'Éducation Thérapeutique du Patient (ETP) vise à "aider les patients à acquérir ou maintenir les compétences dont ils ont besoin pour gérer au mieux leur vie avec une maladie chronique" et se déroule à travers un suivi régulier sur un temps long. Les principaux freins à une ETP efficace sont : le manque de temps des praticiens et des patients, les contraintes économiques du monde de la santé, les contraintes écologiques (par exemple : limiter les déplacements physiques), une accessibilité grandissante d'informations sur internet (pas toujours fiables), un contenu peu adapté à l'individu, une faible adhésion si la motivation du patient n'est pas maintenue. Des retours d'expériences de praticiens montrent que les confinements successifs ont eu un impact négatif sur les patients en ayant modifié le suivi des programmes ETP, ce qui plaide pour une véritable e-ETP allant au-delà de simples consultations de pages web et d'échanges téléphoniques. L'objectif de ce travail est d'apprendre automatiquement un modèle pour personnaliser le parcours patient, uniquement à partir des retours qu'il donne (succès, difficulté...) à la suite de l'exécution de différentes activités. Nous montrons que cette problématique se formalise comme un problème d'apprentissage par renforcement dans un contexte de récompenses binaires et nous explorons alors différents algorithmes adaptatifs. Nous proposons aussi d'utiliser un environnement de simulation afin de comparer leurs performances. La section 2 présente quelques travaux e-ETP et quelques principes en relation. Après avoir formalisé notre problématique, la section 3 explore deux familles d'algorithmes

adaptatifs, suivie en section 4 d'une simulation, une interprétation préliminaire des résultats et quelques perspectives. Nous concluons dans une dernière courte section.

## 2 Travaux reliés

A notre connaissance, il existe peu de travaux concernant la digitalisation intelligente de l'ETP. Les travaux de Goblet et Rey (2020) posent les briques d'une telle e-ETP en s'appuyant sur des principes psychologiques, pédagogiques et l'intelligence artificielle. Retenons principalement que l'aspect motivant se traduit par le fait que chaque patient doit explorer un large espace d'activités ludiques sous la forme de défis personnels. Un défi doit permettre de mettre en oeuvre des compétences ou comportements du patient en relation avec la maladie chronique. Chaque défi, associé à un domaine spécifique, se caractérise obligatoirement par un niveau de difficulté et peut être conditionné pour exécution à des contraintes. A l'échéance d'un défi, le patient évalue sa réussite ou son échec. Ces retours permettent ensuite de proposer au patient de choisir son prochain défi parmi des recommandations personnalisées. Les concepts et les règles sont implémentés en utilisant les langages symboliques OWL2 et SWRL du web sémantique. Contrairement au logiciel KidBreath (Sauzéon et al., 2021), ETP à destination d'enfants asthmatiques, où la personnalisation du parcours est issue de techniques adaptatives des environnements d'apprentissage humain, en particulier des algorithmes de bandits manchots (cf. les travaux précurseurs de Clément et al. (2014) et Clément (2018)). Déployé en condition réelle d'usage, la durée trop courte de l'expérimentation n'a pas permis de montrer une amélioration significative sur l'efficacité pédagogique. Par contre, les auteurs constatent que la personnalisation adaptative a un plus fort impact sur la motivation des enfants. Il existe une multitude d'algorithmes de bandits plus ou moins complexes. Dans le cadre préliminaire de notre étude, nous avons privilégié la simplicité du bandit probabiliste versus le bandit fréquentiste (Kaufman, 2016). Le premier algorithme de bandit mentionné dans la littérature est un algorithme probabiliste avec inférence bayésienne (aussi appelé Thompson Sampling) proposé dans le contexte des essais cliniques (Thompson, 1933). Tombé dans l'oubli pendant des décennies, il a été redécouvert indépendamment par plusieurs auteurs dans les années 2000 et reconnu depuis pour ses très bonnes performances pratiques et ses facilités d'adaptation et d'extension (Russo et al., 2018). L'échantillonnage de Thompson est maintenant utilisé dans différentes applications nécessitant la génération de séquences adaptatives (Lin, 2020).

L'engagement, la motivation et l'adaptation sont aussi présents en psychologie cognitive avec pour objectif la personnalisation du parcours d'un individu apprenant tout au long de sa vie ou dans une perspective thérapeutique. La "théorie du flow" (ou expérience optimale) de Csikszentmihalyi (1975) et le défi approprié (Case-Smith et O'Brien, 2010) se complètent mutuellement. Réaliser une activité où le niveau de difficulté se situe au-delà des compétences a pour effet d'augmenter l'anxiété, alors qu'une activité avec un niveau en dessous des compétences de l'individu va entraîner l'ennui. Entre ces deux situations se situe l'activité considérée comme fournissant l'expérience optimale, c'est-à-dire un équilibre entre le niveau de difficulté de l'activité (défi) et les compétences de l'individu. Peu de travaux s'intéressent à la définition dynamique d'une trajectoire optimale dans cet espace du flow.

# 3 Notre contribution

Opérer une trajectoire, si possible optimale, dans un large espace d'activités peut se formaliser dans le cadre de l'apprentissage par renforcement. Un agent dispose de plusieurs actions pouvant modifier son environnement (ici un patient) dont l'impact est observé sous la forme de récompenses binaires (dites de Bernoulli) : succès ou échec. L'agent doit réaliser une prise de décision séquentielle basée sur les actions et récompenses observées précédemment pour choisir la prochaine activité. Une bonne stratégie exploite cette information de sorte à réaliser un compromis entre exploration (essayer les actions peu jouées) et exploitation (favoriser les actions ayant obtenu des bonnes performances). Maximiser l'espérance des récompenses revient à maximiser le nombre moyen de succès pour chaque patient.

## 3.1 Définition des actions

Le modèle ORALOOS (Goblet et Rey, 2020) caractérise les défis en leur associant une valeur numérique représentant un niveau de difficulté. La psychologie expérimentale montre que le niveau de difficulté d'une tâche, activité ou exercice est le levier principal sur lequel on peut agir pour rester dans le flow et qui plus est lorsque l'espace d'activités à explorer par l'individu est important (Baranes et al., 2014). Dans ce cadre, les actions sont les façons d'agir sur ce niveau de difficulté et consistent à l'incrémenter, le diminuer ou conserver la même valeur. En s'inspirant d'une pédagogie béhavioriste, ORALOOS propose des règles de recommandation évaluées à chaque retour d'un défi réalisé par le patient :

- R+ (renforcement) : en cas de succès, proposer au patient des défis avec un niveau de difficulté supérieur au niveau du défi évalué ;
- R- (remédiation) : en cas d'échec, proposer au patient des défis avec un niveau de difficulté inférieur au niveau du défi courant ;
- R= (palier) : quel que soit son retour, proposer au patient des défis de même niveau de difficulté que le défi courant évalué.

Tous les défis proposés au patient appartiennent au domaine du défi évalué et doivent être exécutables au moment de l'évaluation. La règle palier introduit une prise de décision aléatoire implémentée dans un algorithme *ad hoc* que nous qualifions de semi-adaptatif.

## 3.2 Stratégie semi-adaptative des règles béhavioristes

Les paramètres $\alpha$ et $\beta$ de l'algorithme sont des compteurs de succès et d'échecs, respectivement, pour chaque règle et utilisés pour établir les métriques de comparaison. C'est un algorithme stochastique (dépendant du hasard) et semi-adaptatif car tenant compte uniquement du retour patient sur l'activité courante pour définir la prochaine activité. Si ces règles permettent un premier niveau de personnalisation en tenant compte du retour immédiat du patient, elles n'exploitent pas toutes les activités et retours précédents. Or, cette historique (difficile à écrire uniquement sous une forme logique) est une mine d'informations concernant le comportement du patient qu'il faut exploiter pour une meilleure personnalisation.

**Algorithm 1** RandomBehaviorist({R+,R-,R=}, $\alpha$, $\beta$)

Init : $r_0 \leftarrow randomchoice(\{0, 1\})$ ▷ tirage aléatoire d'un élément parmi un ensemble fini
**for** $t = 1, 2...$ **do**
  **if** $r_{t-1} = 1$ **then**
    $R_i \leftarrow randomchoice(\{R+, R =\})$
  **end if**
  **if** $r_{t-1} = 0$ **then**
    $R_i \leftarrow randomchoice(\{R-, R =\})$
  **end if**
  Appliquer $R_i$ et observer la récompense $r_t \leftarrow \{0, 1\}$
  Mettre à jour les compteurs de la règle tirée :
  $(\alpha_{R_i}, \beta_{R_i}) \leftarrow (\alpha_{R_i} + r_t, \beta_{R_i} + 1 - r_t)$
**end for**

## 3.3 Stratégie adaptative du bandit Thompson Sampling

Une exploitation optimale de cette historique patient est possible en utilisant un algorithme de bandit pour choisir la prochaine activité. Comme évoqué en préambule, ce type d'algorithme permet aussi de répondre simplement au compromis exploitation / exploration. La stratégie d'échantillonnage de Thompson (TS dans la suite) consiste à estimer la meilleure action parmi plusieurs possibles via des distributions de probabilité. L'algorithme TS s'appuie sur les principes de l'inférence bayésienne pour la prise de décision séquentielle à partir d'une série d'observations. Dans le contexte spécifique de Bernoulli (récompenses binaires), la loi Beta possède les caractéristiques nécessaires à son utilisation avec TS :

— Beta est une loi de probabilités continue définie sur l'intervalle $[0, 1]$ ;
— Beta se définit aussi par les deux paramètres $\alpha$ et $\beta$, pseudo-compteurs de succès et d'échecs respectivement ;
— La loi Beta est dite conjuguée car, si l'on "entre" une loi Beta en *a priori* dans la machine bayésienne, on "sort" une loi Beta en *a posteriori* ;
— En l'absence d'observations au démarrage, on utilise souvent comme croyance a priori une loi Beta(1,1) uniforme qui retourne toujours 1, dite non informative.

En associant une loi Beta pour chaque bras k, l'algorithme TS Bernoulli (Russo et al., 2018) s'écrit :

**Algorithm 2** BernTS(K, $\alpha$, $\beta$)

**for** $t = 1, 2...$ **do**
  **for** $k = 1, ..., K$ **do** ▷ Échantillonner les distributions de probabilité
    Sample $\hat{\theta_k} \sim Beta(\alpha_k, \beta_k)$
  **end for**
  $x_t \leftarrow \text{argmax}_k \hat{\theta_k}$ ▷ Sélectionner le meilleur bras
  Appliquer $x_t$ et observer la récompense $r_t \leftarrow \{0, 1\}$
  Mettre à jour la distribution du bras choisi :
  $(\alpha_{x_t}, \beta_{x_t}) \leftarrow (\alpha_{x_t} + r_t, \beta_{x_t} + 1 - r_t)$
**end for**

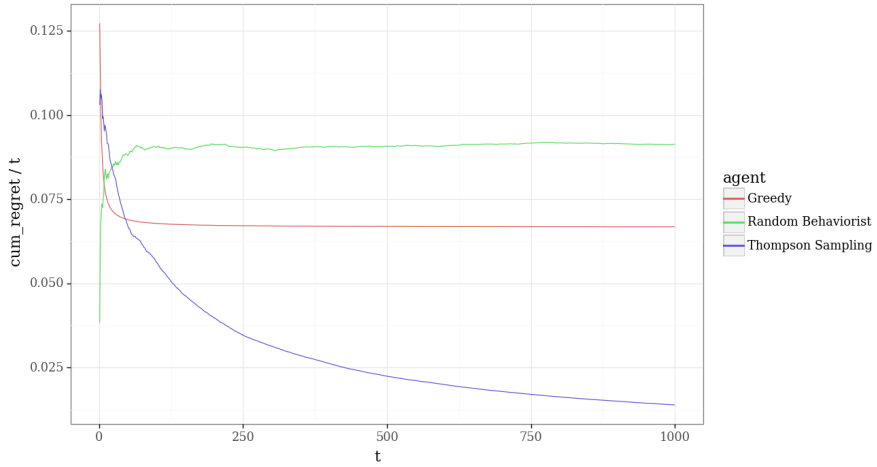



FIG. 1 – *Comparaison du regret cumulé.*

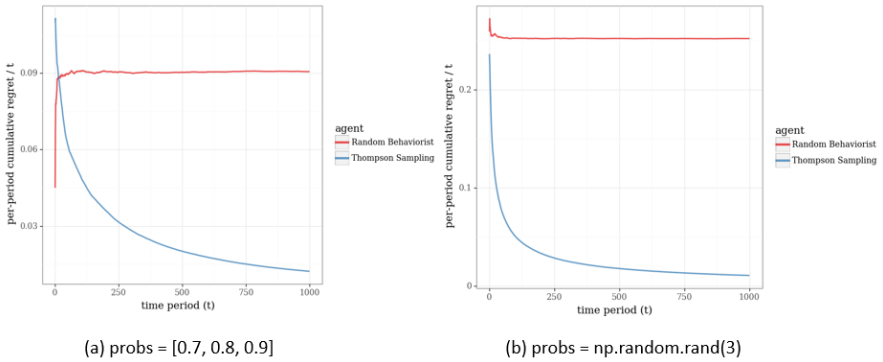
L'idée de TS est de choisir le bras qui maximise la récompense attendue par rapport à une croyance tirée au hasard. Cette technique repose sur l'intuition que si le nombre de tirages pour un bras donné correspond à sa probabilité estimée d'être le bras optimal, on peut alors obtenir un bon compromis entre l'exploitation et l'exploration des possibles. Dans notre contexte ETP, les bras du bandit TS correspondent aux règles R+, R- et R= du modèle ORALOOS. Nous obtenons une stratégie totalement adaptative car, nativement, le bandit tient compte des résultats des choix précédents pour chaque patient.

# 4 Expérimentations

Les expériences ont été mises en place en utilisant l'environnement de simulation accompagnant l'article "A tutorial of Thompson Sampling" (Russo et al., 2018). Le code propose différents bandits prédéfinis comme Thompson Sampling, des agents gloutons et il est facilement adaptable pour des besoins spécifiques. Afin de comparer les performances des différents algorithmes, nous utilisons principalement la métrique du regret cumulé, dont l'objectif d'un agent est de le minimiser. Nous comparons les algorithmes RandomBehaviorist (algo. 1) et BanditTS (algo. 2) par rapport à une stratégie gloutonne (Greedy) qui sert de référence.

## 4.1 Comparaison du regret cumulé des trois algorithmes

La première expérience consiste à comparer les performances des trois algorithmes pour un même patient simulé avec des probabilités d'actions constantes en valeur et en ordre sur toutes les itérations. Par exemple, [0.7, 0.8, 0.9] correspond respectivement à la probabilité de succès en application de R+, R- ou R=. La figure 1 présente le regret cumulé calculé pour les trois algorithmes sur 1000 itérations. Première information : la stratégie béhavioriste obtient de moins bonnes performances que la stratégie gloutonne en terme de regret cumulé, avec un




(a) probs = [0.7, 0.8, 0.9] (b) probs = np.random.rand(3)

FIG. 2 – *Variations des probabilités du patient.*

profil similaire à celle-ci. La stratégie Thompson Sampling est la meilleure des trois, avec un regret cumulé qui tend vers 0 en fin. Ces résultats sont conformes aux travaux de (Chapelle et Li, 2011) et (Granmo, 2010) autour de l'échantillonnage de Thompson qui se révèle particulièrement efficace dans le contexte des bandits Bernouilli stationnaires.

## 4.2 Variations de différents paramètres

La prochaine expérience (cf. figure 2) a pour objectif de comparer le regret cumulé des stratégies Béhavioriste et TS en faisant varier les probabilités de succès du patient simulé, ce qui consiste à regarder le comportement des agents dans un contexte stationnaire avec probabilité fixe pour tous les lots de 1000 itérations (figure 2-a) et non stationnaire (figure 2-b ; tirage aléatoire des trois probabilités entre chaque lot). Quelles que soient les valeurs de probabilités appliquées au patient, l'algorithme TS reste performant.

Une dernière expérience (cf. figure 3) a été de comparer le comportement de TS en faisant varier les paramètres (valeurs initiales) de la loi *a priori* Beta. Rappelons que toutes les expériences précédentes se sont faites avec la loi non informative et uniforme Beta(1, 1). Globalement, Beta(1,1) reste le paramétrage le plus efficace pour l'algorithme TS même si on constate un croisement à partir de la période 500 et une toute relative moins bonne performance que Beta(10, 4). Dans les toutes premières itérations, Beta(1,1) reste beaucoup moins chaotique pour de meilleures performances en terme de regret cumulé. Nous constatons aussi que l'algorithme TS est plutôt sensible au paramétrage $\beta$ du cumul des échecs : une valeur haute, relativement à $\alpha$, entraîne systématiquement une dégradation des performances.

## 4.3 Quelques perspectives

La figure 1 montre que dans les premières itérations, l'algorithme TS est un peu moins performant qu'une stratégie gloutonne. Afin de traiter la problématique du démarrage à froid (faible nombre d'observations), nous pouvons d'abord utiliser un glouton et dès que le regret n'évolue plus, basculer sur un bandit TS en transférant les paramètres $\alpha$ et $\beta$. Au-delà d'une simple variation aléatoire du patient, nous devons aussi étudier une simulation plus complexe

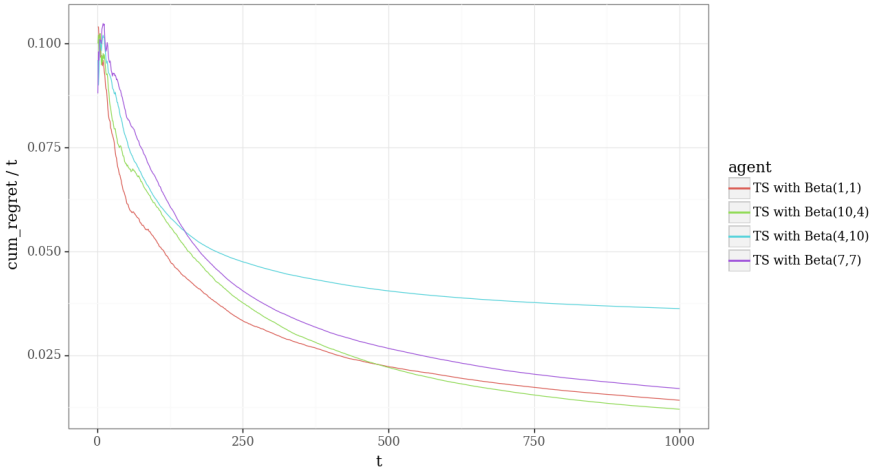



FIG. 3 – *Variations des paramètres de Beta.*

afin d'étayer les performances des bandits TS. Il existe différentes approches possibles et il a été démontré qu'une adaptation du Thompson Sampling reste efficace dans ce cadre (Gupta et al., 2011).

# 5 Conclusion

Personnaliser optimalement le parcours patient dans un contexte d'éducation thérapeutique nécessite l'utilisation d'algorithmes adaptatifs. Nous avons étudié une approche *ad hoc* d'inspiration béhavioriste et un bandit Thompson Sampling issu de la littérature scientifique. En simulation, nous avons appris que l'algorithme *ad hoc* est moins performant qu'une approche gloutonne avec un comportement similaire et que le bandit TS reste le plus efficace quels que soient les paramètres utilisés. Le faible coût opérationnel de ce dernier nous permet d'envisager une utilisation en contexte réel dans une prochaine étape de ce travail.

# Références

Baranes, A. F., P.-Y. Oudeyer, et J. Gottlieb (2014). The effects of task difficulty, novelty and the size of the search space on intrinsically motivated exploration. *Frontiers in Neuroscience 8:317*.

Case-Smith, J. et J. C. O'Brien (2010). *Occupational therapy for children*. Maryland Heights: Mosby/Elsevier.

Chapelle, O. et L. Li (2011). An empirical evaluation of thompson sampling. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, et K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, Volume 24. Curran Associates, Inc.

Clément, B. (2018). *Adaptive Personalization of Pedagogical Sequences using Machine Learning*. Thèse de doctorat, Université de Bordeaux.

Clément, B., D. Roy, P.-Y. Oudeyer, et M. Lopes (2014). Online optimization of teaching sequences with multi-armed bandits. *7th International Conference on Educational Data Mining, London, United Kingdom*.

Csikszentmihalyi, M. (1975). *Beyond boredom and anxiety: Experiencing flow in work and play*. San Francisco, CA: Jossey-Bass.

Goblet, X. et C. Rey (2020). Suivi thérapeutique intelligent par recommandation à base d'ontologie et de règles. *Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle, Afia (Ed)*, 50–57.

Granmo, O.-C. (2010). Solving two-armed bernoulli bandit problems using a bayesian learning automaton. *International Journal of Intelligent Computing and Cybernetics*.

Gupta, N., O.-C. Granmo, et A. Agrawala (2011). Thompson sampling for dynamic multi-armed bandits. In *2011 10th International Conference on Machine Learning and Applications and Workshops*, Volume 1, pp. 484–489.

Kaufman, E. (2016). Modèles de bandit : une histoire bayésienne et fréquentiste. In S. de Mathématiques Appliquées et Industrielles (Ed.), *Revue Matapli*, Volume 109.

Lin, F. (2020). Adaptive quiz generation using thompson sampling. *Third Workshop Eliciting Adaptive Sequences for Learning (WASL 2020)*.

Russo, D. J., B. Van Roy, A. Kazerouni, I. Osband, et Z. Wen (2018). A tutorial on thompson sampling. *Foundations and Trends in Machine Learning 11:1*, 1–96.

Sauzéon, H., B. Clément, C. Mazon, D. Roy, et P.-Y. Oudeyer (2021). Conception d'un système tutoriel intelligent (sti) basé sur les progrès d'apprentissage : des étapes formelles aux étapes d'expérimentations chez les enfants. In C. M. Marie Lefevre (Ed.), *10e Conférence sur les Environnements Informatiques pour l'Apprentissage Humain*, pp. 20–27.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika 25:3-4*, 285–294.

## Summary

As part of a digital therapeutic education where the motivational aspect is preponderant, a personalization of the patient journey must approach an optimal trajectory in a large space of playful activities defined by a level of difficulty and skills to master. The machine must learn from successes or failures to dynamically make recommendations for future steps and deliver the most appropriate experience for each patient. By approaching this problem from the angle of adaptive algorithms, we propose an original approach. We study two families of algorithms (logical rules and bandits Thompson Sampling) by comparing their performances through a simulation environment. The first results show an advantage for the TS bandit regardless of the patient's learning characteristics.