Politiques de transmission basées sur l'apprentissage par renforcement dans les réseaux cellulaires dynamiques et aléatoires

Qiong Liu, Philippe Mary, Jean-Yves Baudais

Politiques de transmission basées sur l'apprentissage par renforcement dans les réseaux cellulaires dynamiques et aléatoires

Qiong LIU, Philippe MARY, Jean-Yves BAUDAIS

Univ Rennes, INSA Rennes, CNRS, IETR-UMR 6164, F-35000 Rennes, France
{qiong.liu, philippe.mary, jean-yves.baudais}@insa-rennes.fr

Résumé – Dans cet article, nous proposons des politiques de transmission de l'information tenant compte de l'état du canal, de la file d'attente des émetteurs et des interférences entre les signaux, pour la liaison descendante des réseaux cellulaires dynamiques. Les réseaux cellulaires à grande échelle avec des distances de transmission aléatoires sont considérés. Le problème est formulé selon un processus de décision de Markov à horizon infini et il est résolu en ligne en utilisant l'apprentissage par renforcement (AR). Le but est de minimiser les coûts de transmission tout en limitant le coût d'attente dans les buffers. Nous montrons qu'il existe un compromis, qui dépend de l'intensité du trafic, entre la probabilité de stabilité et les coûts de transmission. Les résultats de simulation montrent que les politiques basées sur l'AR conservent la même probabilité de stabilité que celle de l'algorithme glouton, mais avec un coût de transmission inférieur.

Abstract – In this work, we provide transmission policies considering the channel state information, the queue states and aggregate interference in dynamic downlink cellular networks. Large-scale networks with multi-cells and random link distances are considered. The problem is formulated with a infinite horizon Markov decision process, and solved online using reinforcement learning (RL). The goal is to minimize the transmission costs while limiting the waiting cost in the buffer. We show there exists a trade-off between the stable probability and transmitting costs which depends the traffic intensity. The numerical results reveals that the RL-based policies holds the same stable region compared to greedy algorithm however with a lower transmission cost.

1 Introduction

La géométrie stochastique fournit un cadre mathématique pour analyser les performances des réseaux sans fil à grande échelle, en considérant le déploiement de ces réseaux comme une réalisation d'un processus spatial ponctuel [1]. Ces dernières années, la géométrie stochastique a été associée à la théorie des files d'attente pour modéliser les systèmes dynamiques spatio-temporels [2, 3]. Cependant, l'interaction entre les files d'attente rend le problème complexe à analyser puisque l'état de chaque file d'attente dépend de l'état de toutes les autres.

Dans un réseau de communication dynamique, les stratégies de transmission doivent être adaptées en fonction de l'état du réseau afin de satisfaire un critère d'optimalité. Cependant, il peut être très difficile de dériver analytiquement la stratégie de transmission optimale lorsque le système devient complexe. Ces dernières années, l'apprentissage par renforcement (AR) est revenu sur le devant de la scène dans son application à la gestion des ressources radio du réseau, dès lors que l'interaction de l'agent avec son environnement est modélisée par un processus décisionnel markovien (PDM) [4]. L'intérêt de cette approche est qu'elle permet de trouver une politique de transmission optimale, en un sens qui sera clarifié ultérieurement, dans un environnement incertain, sans modèle physique explicite de la communication pour effectuer l'allocation de ressources mais seulement par le biais d'essais et d'erreurs de la part de l'agent.

Dans le domaine des communications numériques, l'AR a été appliqué aux systèmes point à point afin de minimiser l'énergie consommée sous contrainte de délai, e.g. [5, 6], aux réseaux IoT pour la gestion du spectre et de récupération d'énergie [7], ainsi qu'au problème du *caching* dans les réseaux cellulaires à l'aide de l'AR profond [8]. Ces références ne sont pas exhaustives, cependant aucun travail n'a encore combiné l'apprentissage par renforcement avec la géométrie stochastique dans un réseau dynamique pour l'évaluation des performances moyennes de l'AR.

Dans cet article, nous nous intéressons à la région de stabilité d'un réseau cellulaire lorsqu'un agent apprend sa politique de transmission dans un environnement dynamique avec des interférences inter-cellules. Les principales contributions sont :

1. Nous montrons qu'une politique de transmission basée sur le *Q-learning* et SARSA [1] permet d'atteindre la même région de stabilité que la politique gloutonne à un coût moindre ;
2. Il existe un compromis entre la probabilité de stabilité et le coût de transmission. Ce compromis dépend de l'intensité du trafic et de la densité de station de base.

2 Modèle et hypothèses

2.1 Topologie

On considère la liaison descendante d'un réseau cellulaire où la position des stations de base est modélisée par un processus ponctuel de Poisson homogène Φ de densité λ. La densité des utilisateurs est suffisamment grande pour que chaque cellule contienne au moins un utilisateur et celui-ci est associé à la station de base la plus proche. On considère que l'utilisateur typique est positionné à l'origine. Toutes les stations de base sont supposées transmettre dans la même bande.

1. *current State, current Action, next Reward, next State, next Action*.

2.2 Modèle de canal

On considère un canal à évanouissement par bloc, où les gains du canal entre toutes paires d'émetteur et récepteur sont supposés indépendants et identiquement distribués suivant une loi exponentielle. Les coefficients des canaux restent constants pendant la transmission d'un paquet et varient d'un paquet à l'autre. La fonction d'affaiblissement de propagation est $g(x) = \|x\|^{-\alpha}$, où α est l'exposant d'affaiblissement. Par conséquent, le rapport signal sur interférence (RSI) de l'utilisateur typique est

$$\gamma_t = \frac{H_{x_0,t} \|x_0\|^{-\alpha}}{\sum\limits_{x \in \Phi \setminus x_0} \beta_{x,t} H_{x,t} \|x\|^{-\alpha}} \tag{1}$$

où $H_{x_0,t}$ et $H_{x,t}$ sont les gains d'évanouissement à l'instant t entre l'utilisateur typique et, respectivement, sa station de base située à x_0, et la station interférente située à x. Le paramètre $\beta_{x,t}$ vaut 1 si l'émetteur en x est actif et 0 sinon.

2.3 File d'attente

Le temps est discrétisé en créneaux de tailles égales et indexés par $t = 1, 2, \cdots$. Chaque station de base dispose d'un buffer pour stocker les paquets en attente de transmission. La taille du paquet est fixée et sa transmission prend exactement un pas de temps. Les paquets arrivés à l'instant t sont considérés à l'instant $t+1$. L'équation dynamique des files d'attente est

$$B(t+1) = [B(t) - D(t)]^+ + S_y(t) \tag{2}$$

où $B(t)$ est la taille du buffer à l'instant t et $S_y(t)$ est le processus d'arrivée des paquets, qui suit une distribution de Bernoulli d'intensité $\xi \in [0,1]$. Le processus de départ $D(t)$ dépend du RSI à chaque instant : si un paquet est transmis avec succès, il est retiré du buffer sinon ce paquet reste dans le buffer pour une transmission ultérieure, jusqu'à ce qu'il soit reçu avec succès. On note $[v]^+$ la partie positive de v, i.e. $[v]^+ = \max(0, v)$.

3 Apprentissage par renforcement

L'apprentissage par renforcement est basé sur l'interaction d'un agent avec un environnement inconnu. La décision se fait par un processus d'essais et d'erreurs [9]. Dans chaque état s de l'espace $\mathcal{S}$ des états, l'agent sélectionne une action a dans l'ensemble $\mathcal{A}$ des actions possibles. Le choix de l'action a est dicté par sa politique π qui a pour distribution $\pi(a|s)$. En réponse, l'agent reçoit une récompense $r(s,a)$ et passe à l'état suivant $s' = T(s,a)$. Les interactions entre l'agent et l'environnement se poursuivent jusqu'à ce que l'agent ait appris une politique maximisant sa récompense cumulée sur le long terme.

On note $S(t) = [S_c(t), S_b(t), S_y(t)] \in \mathcal{S}$ l'état de l'environnement à l'instant t, où $S_b \in \mathcal{B} = \{0,1\}$ indique si le buffer est vide ou non. La variable $S_y(t) \in \mathcal{Y} = \{0,1\}$ vaut 1 si un nouveau paquet arrive dans le buffer et 0 sinon, avec $\mathbb{P}(S_y(t) = 1) = \xi, \forall t \in \mathbb{N}$. D'autre part, S_c représente l'état du canal, avec $S_c(t) = i$ si $\gamma_t \in [\theta_i, \theta_{i+1}[$, $i \in [0, M]$ et $\theta_0 = 0$, $\theta_{M+1} = +\infty$. L'espace des actions est $\mathcal{A} = \{0,1\}$. À chaque instant, la station typique décide de transmettre ou non un paquet mais si le buffer est vide, elle reste silencieuse. Si un paquet est transmis, i.e. $A(t) = 1$, le paquet est retiré du buffer avec une certaine probabilité $f_s(1, S_c(t))$ qui caractérise la probabilité de bonne réception :

$$f_s : \mathcal{A} \times \Gamma \to [0,1] \tag{3}$$

où $f_s(\cdot,\cdot)$ est une fonction croissante du RSI. On a

$$f_s(1, S_c(t)) = \begin{cases} 0, \text{ si } \theta_1 > \gamma_t \\ f_m, \text{si } \theta_m < \gamma_t < \theta_{m+1}, m \in [1, M-1] \\ f_M, \text{si } \gamma_t > \theta_M \end{cases} \tag{4}$$

et $f_s(0, S_c(t)) = 0$. De manière similaire qu'en [5], la fonction objectif de l'agent est modélisée par deux fonctions de coût :

1. Le coût de transmission est une fonction C non croissante avec le RSI, c'est-à-dire qu'il est moins coûteux de transmettre dans de bonnes conditions de transmission. On définit cette fonction par :
$$C(A(t), S(t)) = \begin{cases} f(A(t), S_c(t)), \text{ si } A(t) = 1 \\ 0, \text{ sinon} \end{cases} \tag{5}$$
2. Le coût de délai. On considère que maintenir une file d'attente à un coût fixe, avec
$$W(A(t), S(t)) = \begin{cases} 0, \text{ si } A(t) = 1, \\ w, \text{ si } A(t) = 0, S_b(t) > 0. \end{cases} \tag{6}$$

La politique d'allocation est une application entre l'état du système et une action. Elle peut-être déterministe ou aléatoire. Soit π une politique stationnaire, i.e. invariante dans le temps, et Ψ l'ensemble des politiques stationnaires. On définit deux coûts à *long terme* sachant un état initial $S(0) = s$ et une réalisation du réseau Φ comme

$$C_\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \eta^t C(A(t), S(t)) \middle| s, \Phi \right] \tag{7}$$

$$W_\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \eta^t W(A(t), S(t)) \middle| s, \Phi \right] \tag{8}$$

où $C_\pi(s)$ et $W_\pi(s)$ sont les coûts moyens de transmission et de délai en partant d'un état s pour une certaine réalisation du réseau et en suivant la politique π. L'espérance est prise sur la distribution dynamique du processus markovien sous-jacent. Le coefficient $\eta \in [0,1]$ traduit l'affaiblissement des récompenses futures sur le coût moyen à mesure que le temps passe.

Le problème d'apprentissage par renforcement consiste à trouver la politique π qui minimise le coût de transmission moyen sous contrainte de coût de délai, i.e.

$$\min_{\pi \in \Psi} C_\pi(s) \quad \text{t.q.} \quad W_\pi(s) \leq \delta, \ \forall s \in \mathcal{S}. \tag{9}$$

Le problème (9) peut se reformuler avec le lagrangien $L(t,\lambda) = C(A(t), S(t)) + \lambda W(A(t), S(t))$, $\lambda \geq 0$, et en définissant $L_\pi(s,\lambda) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \eta^t L(t,\lambda) \mid s, \Phi\right]$. Ainsi le problème (9) devient

$$\min_{\pi \in \Psi} L_\pi(s,\lambda) = \min_{\pi \in \Psi} C_\pi(s) + \lambda W_\pi(s) \tag{10}$$

et la politique optimale, dépendante de λ, est

$$\pi^*_\lambda = \arg\min_{\pi \in \Psi} L_\pi(s,\lambda). \tag{11}$$

Il reste à trouver le multiplicateur de Lagrange optimal.

Multiplicateur de Lagrange optimal Le multiplicateur de Lagrange optimal λ^* satisfait

$$\lambda^* = \arg\max_{\lambda} \min_{\pi \in \Psi} L_\pi(s, \lambda) - \lambda\delta \tag{12}$$

qui peut être atteint en utilisant les sous-gradients stochastiques [6], c'est-à-dire,

$$\lambda_{k+1} = \lambda_k + \frac{1}{k}(W_{\pi_\lambda}(s) - (1-\eta)\delta). \tag{13}$$

Le terme $(1-\eta)\delta$ convertit la contrainte d'attente actualisée δ en une contrainte d'attente moyenne. La séquence $(\lambda_1, \lambda_2, \cdots)$ converge vers λ^* sous des conditions légères [6].

4 Politique optimale de transmission

Pour une réalisation du PPP, la fonction état-action optimale $q^*(s,a) = \min_\pi q_\pi(s,a)$ est la valeur d'action minimale pouvant être atteinte en prenant l'action a à partir de l'état s et en suivant la politique π. Cette quantité suit l'équation de Bellman soit [9] :

$$q^*(s,a) = \mathbb{E}_{S',L}\left[L(0,\lambda) + \min_{a'} q^*(s',a') \;\middle|\; s, a, \Phi\right] \tag{14}$$

La politique optimale π^*, qui est stationnaire, est obtenue lorsque $q^*(a,s)$ est atteinte. Nous étudions les performances de trois politiques dans cet article : la politique obtenue avec le *Q-learning* [9], celle obtenue avec SARSA [10], et la politique gloutonne, cette dernière consistant simplement à transmettre tout le temps.

Le Q-learning et SARSA sont deux algorithmes itératifs qui permettent de converger vers la politique stationnaire optimale, au sens du problème lagrangien défini ci-dessus. Pour ce faire, la valeur de la fonction état-action est mise à jour à partir d'une différence incrémentale entre l'objectif et l'estimée précédente de la fonction état-action, soit :

$$\begin{aligned} q_{t+1}(S(t+1), A(t+1)) \leftarrow\ & q_t(S(t), A(t)) \\ & + \alpha_t \left[T_{t+1} - q_t(S(t), A(t))\right] \end{aligned} \tag{15}$$

où α_t est le pas de mise à jour avec $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$ afin d'assurer la convergence [9]. D'autre part, T_{t+1} est la valeur objectif de l'algorithme et prend une forme légèrement différente selon que l'on considère le Q-learning ou SARSA.

4.1 Q-learning

La valeur de l'objectif est

$$T_{t+1} = L(t+1, \lambda) + \eta \min_{a' \in \mathcal{A}} q_t(S(t+1), a'). \tag{16}$$

Dans la phase d'exploration de la table, la politique ϵ-gloutonne est utilisée pour sélectionner l'action à chaque instant ; c'est ce que l'on appelle la politique de *comportement*. Ainsi

$$a = \begin{cases} \arg\min_{a' \in \mathcal{A}} q(S(t), a'), \text{ avec la probabilité } 1 - \epsilon_t, \\ a' \in \mathcal{A}, \text{ avec la probabilité } \epsilon_t. \end{cases} \tag{17}$$

Mais la politique *cible*, celle indiquant l'action à prendre dans le prochain état, est gloutonne.

4.2 SARSA

La valeur de l'objectif est ici

$$T_{t+1} = L(t+1, \lambda) + \eta q_t(S(t+1), A(t+1)). \tag{18}$$

Contrairement au *Q-learning*, les politiques de *comportement* et *cible*, sont toutes deux ϵ-gloutonne, i.e. la prochaine action à prendre en observant l'état $S(t+1)$ est l'action a' qui maximise $q_t(S(t+1), a')$ avec une probabilité $1-\epsilon$, et une action au hasard avec la probabilité ϵ.

4.3 Probabilité de stabilité

Étant donné Φ et la politique π, la probabilité moyenne de succès de la transmission est

$$r_\Phi(s, \pi) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} \mathbb{E}\left[f_s(A(t), S_c(t)) \;\middle|\; s, \Phi, \pi\right] \tag{19}$$

et la probabilité de stabilité p_s est

$$p_s = \mathbb{P}_\Phi\left[r_\Phi(s.\pi) > \xi\right]. \tag{20}$$

Proposition 1. *La politique gloutonne, selon laquelle la BS est toujours active lorsque le buffer n'est pas vide, donne une borne supérieure à la probabilité de stabilité des politiques basées sur l'apprentissage par renforcement.*

Démonstration. La preuve, qui n'est pas détaillée ici, est basée sur la méta-distribution du RSI [11]. □

5 Évaluation expérimentale

Soit un PPP de densité $\lambda = 0.25$ stations/km^2 sur une surface de 900 km^2. La puissance de transmission est normalisée à 1 pour chaque station de base et $\alpha = 4$. Pour chaque réalisation du réseau, la fonction état-action de l'agent typique est mise à jour à l'aide d'une table pour les algorithme *Q-learning* et SARSA jusqu'à la convergence, c'est-à-dire lorsque, $\forall o > 0$ $|q_{t+1}(s,a) - q_t(s,a)| \leq o$, $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$. Ensuite, une nouvelle réalisation de réseau est tirée et le processus se répète. La simulation est répétée 5000 fois.

À chaque instant, au plus un paquet est transmis ou arrive dans la file d'attente (ou les deux). La qualité du canal est modélisée par trois régions distinctes, (s_c^1, s_c^2, s_c^3), séparées par deux seuils arbitrairement fixés à $\theta_1 = -1.47$ dB et $\theta_2 = 5.07$ dB. Lorsque $\gamma_t < \theta_1$, i.e. $s_c = s_c^1$, la transmission échoue ; lorsque $\theta_1 < \gamma_t < \theta_2$, i.e. $s_c = s_c^2$, le paquet est transmis avec succès avec la probabilité $f_1 = 0.5$; si $\gamma_t > \theta_2$, i.e. $s_c = s_c^3$, le paquet est transmis avec succès, i.e. $f_2 = 1$. La fonction de coût de transmission est définie de manière similaire qu'en [5], avec $c(a, s_1) = 1.7a$, $c(a, s_2) = 0.8a$ et $c(a, s_3) = 0.2a$, $\forall a \in \mathcal{A}$. Au temps t, si la station de base ne transmet rien alors que le buffer n'est pas vide, il y a un coût de délai $w(a) = 0.6a$. Le facteur d'oubli est $\eta = 0.98$, et le pas d'apprentissage est $\alpha_t = 0.01$. Enfin, ϵ_t dans (17) est fixe à $\epsilon_t = 1/t$.

La figure 1 compare le coût total et la probabilité de stabilité des politiques en fonction de l'intensité du trafic. Pour une même configuration de réseau, la politique basée sur l'AR est capable de maintenir la même région de stabilité à un coût de transmission inférieur à la politique gloutonne. Il n'y a pas de

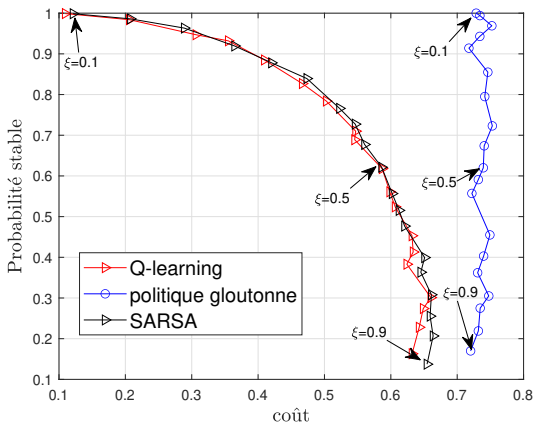

FIGURE 1 – Compromis probabilité de stabilité-coût total.

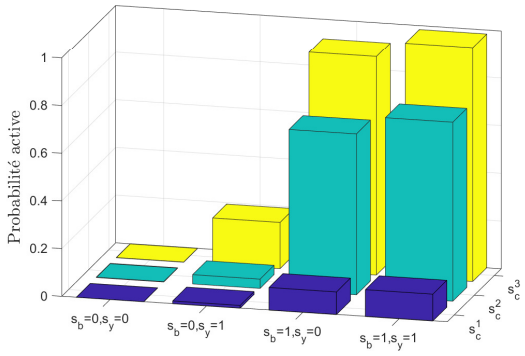

FIGURE 2 – Probabilité d'activité basée sur *Q-learning*, $\xi = 0.3$.

différence significative de performance entre les algorithmes *Q-learning* et SARSA qui convergent vers la politique optimale. On observe qu'il existe un compromis entre la probabilité de stabilité et le coût total des politiques basées sur l'AR. En effet, à mesure que l'intensité du trafic augmente, la probabilité de stabilité diminue et l'agent a tendance à être plus actif pour envoyer des paquets, ce qui augmente d'autant le coût de transmission. Les politiques basées sur AR permettent aux agents d'ajuster de manière flexible la politique de transmission en fonction de l'intensité du trafic et de la configuration du réseau, tandis que la politique gloutonne n'est sensible qu'aux états du buffer.

La figure 2 montre la politique moyenne suivie selon les 12 états possibles de l'environnement. La politique est complètement décrite par la probabilité d'activité, qui est définie comme $\bar{\pi}_s = \mathbb{E}_\Phi[\pi(a=1|s, r_\Phi(s,\pi) > \xi)]$, c'est-à-dire la probabilité que l'agent soit actif dans un état où le buffer ne diverge pas. Nous observons que l'agent est plus actif lorsque le RSI est bon et que le buffer n'est pas vide. Par exemple, dans l'état $[s_c^3, s_b = 1, s_y = 1]$, la probabilité d'activité est de 98.7%, car le coût de transmission est faible et la probabilité de succès est élevée. Dans l'état $[s_c^1, s_b = 0, s_y = 0]$, la probabilité d'activité de l'agent est de 0. L'agent a une probabilité de 66.7% d'être actif lorsqu'il rencontre un niveau modéré de RSI, tandis qu'avec la politique gloutonne, l'agent est toujours actif dans cet état.

6 Conclusion

L'étude des politiques optimales de transmission dans les réseaux cellulaires à grande échelle constitue un défi fondamental car l'environnement est dynamique. Récemment, l'AR est apparu comme une technique prometteuse pour résoudre les problèmes de programmation dynamique. Dans ce travail, nous introduisons l'AR pour les réseaux cellulaires PPP afin de rechercher des stratégies optimales de transfert en tenant compte de l'intensité du trafic, des interférences et de l'état du canal. Les résultats de l'expérience montrent que la politique de transmission basée sur RL garantit la même région stable que celle de la politique gloutonne mais à un coût inférieur. De nombreux travaux futurs peuvent être explorés sur la base de ce travail, comme la prise de décision distribuée basée sur l'AR multi-agents.

Références

[1] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Transactions on Communications*, vol. 59, no. 11, pp. 3122–3134, 2011.

[2] M. Gharbieh, H. Elsawy, A. Bader, and M. S. Alouini, "Spatio-temporal Stochastic Modeling of IoT Enabled Cellular Networks : Scalability and Stability Analysis," *IEEE Transactions on Communications*, vol. 65, no. 8, pp. 3585–3600, 2017.

[3] Q. Liu, J. Baudais, and P. Mary, "A tractable coverage analysis in dynamic downlink cellular networks," in *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2020.

[4] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y. C. Liang, and D. I. Kim, "Applications of Deep Reinforcement Learning in Communications and Networking : A Survey," *IEEE Communications Surveys and Tutorials*, vol. 21, no. 4, pp. 3133–3174, 2019.

[5] M. H. Ngo and V. Krishnamurthy, "Monotonicity of constrained optimal transmission policies in correlated fading channels with ARQ," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 438–451, 2010.

[6] N. Mastronarde and M. Van Der Schaar, "Joint physical-layer and system-level power management for delay-sensitive wireless communications," *IEEE Transactions on Mobile Computing*, vol. 12, no. 4, pp. 694–709, 2013.

[7] C. Man, L. Hang, L. Xuewen, and C. Shuguang, "Reinforcement learning-based multiaccess control and battery prediction with energy harvesting in IoT systems," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2009–2020, 2019.

[8] A. Sadeghi, F. Sheikholeslami, and G. B. Giannakis, "Optimal and scalable caching for 5G using reinforcement learning of space-time popularities," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 1, pp. 180–190, 2018.

[9] R. S. Sutton and A. G. Barto, *Reinforcement learning : An introduction*. MIT press Cambridge, 2018.

[10] H. Van Seijen, H. Van Hasselt, S. Whiteson, and M. Wiering, "A theoretical and empirical analysis of expected Sarsa," in *2009 ieee symposium on adaptive dynamic programming and reinforcement learning*. IEEE, 2009, pp. 177–184.

[11] M. Haenggi, "The meta distribution of the sir in poisson bipolar and cellular networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 4, pp. 2577–2589, 2016.