UNIVERSITÉ DU QUÉBEC À MONTRÉAL

APPARIEMENT INTELLIGENT DES APPAREILS D2D SOUS-JACENTS AUX RÉSEAUX CELLULAIRES ET ACTIVÉS POUR LE CACHE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

ACHRAF MOUSSAID

DÉCEMBRE 2020

REMERCIEMENTS

Au terme de ce mémoire, je présente mes sincères remerciements à Madame Pr. Halima ELBIAZE ma directrice de recherche, pour son soutien scientifique, moral et financier, ses encouragements et ses orientations précieuses, malgré les occupations et les responsabilités qu'elle assumait durant ma période du projet. Que ce travail soit le modeste témoignage de ma haute considération.

Je remercie aussi Dr. Wael JAAFAR qui m'a énormément aidé dans la réalisation de ce mémoire à travers ses conseils et ses remarques.

Mes vifs remerciements vont également aux membres du jury pour l'intérêt qu'ils ont porté à notre recherche en acceptant d'examiner notre travail et de l'enrichir par leurs propositions.

Sans oublier de remercier vivement le corps professoral de la faculté des sciences de l'UQAM. Et tous mes collègues au sein du laboratoire TRIME avec qui j'ai passé de très bons moments.

Mes remerciements s'adressent également à mes parents, ma soeur et mon grand-père qui ont été toujours présents pour moi et qui m'ont toujours soutenu et encouragé.

Finalement, je remercie Mr Taki LARAKI et Mr Rachid LARAKI pour tous leurs soutiens et conseils qui m'ont énormément aidé durant tout le long de mon parcours scolaire et universitaire.

Achraf MOUSSAID

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ACRONYMES

1G	Première Génération
2G	Deuxième Génération
3G	Troisième Génération
4G	Quatrième Génération
BS	Station de base
Cisco	Computer Information System COmpany
Dec-POMDP	Decentralized Partially Observable Markov Decision Process
DQN	Deep Q-Network
DUE	D2D User Equipment
IA	Intelligence Artificielle
MARL	Multi-Agent Reinforcement Learning
MARL	Multi-agents Reinforcement Learning
MAS	Multi-agent Systems
MDP	Markov Decision Process
ML	Machine Learning
MNOs	Mobile Network Operators ou Opérateurs de réseaux mobiles

MOS	Mean Opinion Score
RL	Reinforcement Learning
SINR	Signal-to-interference-plus-noise ratio
TRIME	Télécommunication, Réseaux, Informatique Mobile et Embarquée

RÉSUMÉ

La communication *Device-to-Device* (D2D) a suscité de l'intérêt en tant que technologie prometteuse pour les réseaux sans fil de prochaine génération, elle favorise l'utilisation de communications point à point entre équipements utilisateur (UE) sans passer par les stations de base (BS).

Les communications appareil à appareil (D2D) ont été proposées dans les réseaux cellulaires comme un paradigme complémentaire pour améliorer principalement la connectivité du réseau. Cependant, avec l'émergence de nouvelles applications, telles que la distribution de contenu et la publicité géolocalisée, de nouveaux cas d'utilisation de D2D dans les réseaux cellulaires ont été introduits. Plus précisément, les réseaux activés par le cache ont attiré l'attention en raison de leur capacité à réduire le trafic de liaison et à éliminer les transmissions redondantes de contenu.

Cependant, ce mode de communication a introduit un surcoût d'interférence supplémentaire, en raison de multiples communications simultanées sur les mêmes bandes de fréquences que les utilisateurs cellulaires (CU), ce qui est difficile à contrôler et à atténuer.

L'objectif de ce mémoire est de proposer une solution distribuée et intelligente pour le déchargement du trafic entre les utilisateurs D2D sous-jacents à un réseau cellulaire. Plus précisément, nous étudions le problème d'appariement distribué entre les utilisateurs demandeurs et les appareils de mise en cache à proximité. Etant donné que ce problème est NP-difficile, nous proposons une nouvelle approche d'apprentissage par renforcement multi-agent (MARL), basée sur l'algorithme QMIX, où chaque utilisateur demandeur est un agent capable de décider à quel appareil de cache appairer, tout en garantissant la qualité de service des utilisateurs cellulaires.

Grâce à des simulations, nous montrons l'efficacité de l'algorithme MARL proposé pour obtenir la meilleure stratégie d'appariement D2D dans le réseau cellulaire sous-jacent, par rapport aux approches de base. Enfin, l'impact de plusieurs paramètres a été étudié, comme la taille du réseau D2D, la taille de la bibliothèque de fichiers et les exigences de qualité de la communication.

Mots Clés : *Device-to-Device* (D2D), Apprentissage par renforcement

Multi-agent (MARL), QMIX, Déchargement du réseau.

CHAPITRE I

INTRODUCTION GÉNÉRALE

1.1 Contexte

La croissance explosive du nombre d'appareils mobiles et des applications Internet mobiles telles que les jeux mobiles, le visionnement des vidéos à haute définition (HD) et la réalité virtuelle apporte un grand confort à la société avec des progrès rapides dans la technologie et les services cellulaires. Ces progrès associés à la nécessité d'accéder aux données à tout moment, n'importe où et à partir de n'importe quel appareil, ont entraîné une augmentation de la demande de débits de données et d'exigence en qualité de service (QoS) (Jameel *et al.*, 2018). D'après Cisco (Cisco, 2020), il y avait 8,8 milliards d'appareils et de connexions mobiles dans le monde en 2018, qui passeront à 13,1 milliards en 2020, ce qui à son tour remettra en question la gestion actuelle des ressources réseau disponibles.

Les réseaux cellulaires ont jusqu'à présent été en mesure de maintenir une qualité de service acceptable pour une multitude d'applications et de fournir une bonne couverture aux utilisateurs dans la plupart des zones géographiques mêmes celles les plus isolées. Toutefois, les techniques utilisées par les réseaux actuels ne seront pas en mesure de répondre aux demandes croissantes des futurs utilisateurs mobiles, en termes de débit et de latence. Ces demandes se font encore plus difficile à satisfaire dans des environnements très achalandés où les utilisateurs sont

à proximité les uns des autres, comme dans les centres commerciaux ou dans les événements sportifs ou musicaux (Jameel *et al.*, 2018).

Dans ce cadre, des technologies tels que les réseaux hétérogènes, les communications de machine à machine, les réseaux d'appareil à appareil (D2D) et l'Internet des objets (IoT) ont été proposées pour faire l'objet d'une normalisation sous l'égide du développement de réseaux 5G. Les réseaux D2D sont considérés comme l'une des technologies clés pour les réseaux cellulaires 5G en raison du besoin inhérent d'un débit de données élevé, d'un délai de communication limité et d'une communication qui respecte la QoS (Ansari *et al.*, 2018).

Cependant, les communications D2D peuvent être plus utiles et bénéfiques si le contenu est stocké au niveau des appareils D2D. Dans ce contexte, les techniques de mise en cache qui stockent le contenu (vidéo, pages Web, etc) sur les périphériques de stockage près du bord du réseau sans fil (BS, équipement utilisateur) pour une utilisation future sont de bons candidats pour réduire le trafic sur le réseau d'amenée (*backhaul*) et éliminer les transmissions redondantes de contenu (Li *et al.*, 2018a).

1.2 Motivation et problématique

La technologie D2D reçoit une attention considérable, vu qu'elle apporte une multitude d'avantages, tels que l'amélioration du débit du réseau et l'amélioration de la robustesse aux défaillances de l'infrastructure (Andreev *et al.*, 2014). Plus important encore, les communications D2D permettent de réduire considérablement la charge de trafic de la BS, en permettant aux utilisateurs d'effectuer un partage direct du contenu et des services locaux, en particulier dans les zones denses.

De plus, la transmission directe entre les appareils peut être obtenue avec une puissance de transmission inférieure, ce qui se traduit par une efficacité énergé-

tique améliorée. En outre, la communication D2D peut offrir de nombreux autres avantages tels que le contrôle de la congestion et les garanties de la QoS. La communication D2D est particulièrement avantageuse pour améliorer la couverture cellulaire et le débit au niveau de la zone de bord de cellule où les signaux sont beaucoup plus faibles (Jameel *et al.*, 2018).

Bien que les communications D2D présentent de nombreux avantages, de nombreux défis restent à relever pour mettre en œuvre avec succès cette technologie. En particulier, les communications D2D nécessiteront des mécanismes efficaces de découverte des appareils, des algorithmes intelligents de sélection du mode de communication (D2D ou cellulaire), des techniques performantes de gestion des ressources, des procédures de gestion de la mobilité et des protocoles de sécurité robustes.

Une exigence de conception fondamentale pour les réseaux D2D et qu'on va utiliser dans ce mémoire est la découverte des appareils à proximité et d'établir une connexion directe avec eux. Pour accomplir cette tâche, les appareils échangent des signaux pour recueillir des informations telles que l'emplacement et la distance de l'appareil, l'état du canal et l'identificateur de l'appareil, etc. Ces informations sont utilisées par les appareils pour évaluer la faisabilité du regroupement en une paire les uns avec les autres. En règle générale, la découverte de périphériques dans les communications D2D est soit centralisée, où les appareils se découvrent à l'aide d'une entité centralisée ou typiquement à l'aide de la BS, soit distribuée, dans ce cas les appareils se localisent les uns les autres sans l'intervention de la BS et transmettent périodiquement des messages de contrôle pour localiser les appareils à proximité. Cependant, des problèmes de synchronisation, d'interférence et de puissance du signal de balise se posent fréquemment en mode distribué (Jameel *et al.*, 2018).

Le choix du mode de communication est considéré aussi comme un problème dans les communications D2D qu'on va aussi étudier dans ce mémoire, car deux appareils communicants peuvent fonctionner dans le même mode, différent ou hybride, ce qui rend la gestion du réseau plus complexe. Généralement, les appareils peuvent choisir l'un des quatre modes de communication dépendemment de la situation et de la disponibilité des ressources, on a le mode purement cellulaire, le mode partiellement cellulaire, le mode dédié, ou bien le mode sous-jacent (autrement dit, mode de partage) (Klügel et Kellerer, 2020). Donc choisir le mode de communication approprié et en prenant en considération le maintien de la QoS des utilisateurs cellulaires et la bonne communication entre les appareils cellulaires constituent un problème majeur qu'il faut bien gérer.

D'un autre côté, un autre problème étudié dans ce mémoire est la gestion des interférences entre les appareils cellulaires et les appareils D2D est considérée comme l'un des problèmes les plus critiques pour la communication D2D en mode de partage (où les mêmes ressources radio sont utilisées pour les communications cellulaires et D2D). Bien que le déploiement de la communication D2D en mode de partage améliore l'efficacité spectrale. Cependant, cela donne lieu à des problèmes de gestion des interférences puisque par rapport aux scénarios de communication cellulaire, le système nécessite de gérer de nouvelles situations d'interférence. Si l'interférence générée n'est pas bien contrôlée, cela détériorerait les avantages potentiels de la communication D2D car la capacité et l'efficacité cellulaires globales sont dégradées (Noura et Nordin, 2016).

La gestion de la mobilité constitue également une composante essentielle des communications D2D. Étant donné que les appareils peuvent changer d'emplacement tout en communiquant entre eux, leur connectivité peut être interrompue. Par conséquent, un mécanisme est nécessaire pour gérer la communication lorsque les appareils sont mobiles. Pour les applications D2D telles que le transfert de

données en masse ou le déchargement cellulaire entre des appareils à proximité, l'évaluation des modèles de mobilité des appareils et leur impact sur la fiabilité des communications reste aussi un défi majeur (Jameel *et al.*, 2018). Dernièrement, la sécurité est aussi un problème dans les communications D2D.

D'autre part, la mise en cache du contenu au niveau des appareils dans les communications D2D est avantageuse mais présente aussi des limitations. En effet, en plus d'être stockés dans la BS, les contenus les plus populaires peuvent très bien être stockés dans plus d'appareils, tout en étant demandés par plus d'appareils, ce qui peut causer un problème de stockage au niveau des appareils. Si certains utilisateurs trouvent leurs contenus souhaités mis en cache localement par leurs appareils voisins, un moyen prometteur est d'obtenir ces contenus des voisins via les communications D2D, ce qui engendra une gestion plus élevée d'interférences et d'allocation des ressources (Wu *et al.*, 2018).

1.3 Objectifs

Dans ce mémoire, nous étudions trois des problèmes majeurs des communications D2D cités dans la section précédente, comme la découverte des appareils, le choix du mode de communication, et la gestion d'interférence d'une part entre les utilisateurs D2D et les utilisateurs cellulaires, et entre les utilisateurs D2D d'une autre part. L'objectif général de ce projet de maîtrise est d'étudier le problème d'appariement des appareils D2D équipés de capacités de stockage pour mettre en cache des données, et qui sont sous-jacents à un réseau cellulaire, dans le but d'améliorer les performances en terme de débit total des utilisateurs D2D, tout en respectant les contraintes de communication cellulaire et la QoS des appareils D2D, ainsi que la disponibilité du contenu.

Prenant en considération la continuité des transmissions D2D, nous proposons

dans ce travail une nouvelle solution distribuée et intelligente pour le déchargement du trafic parmi les utilisateurs D2D sous-jacents à un réseau cellulaire, en utilisant l'apprentissage par renforcement multi-agents (MARL).

Pour résumer, nous promouvons dans ce mémoire une approche distribuée intelligente, où aucun échange d'informations n'est requis. Les décisions d'appariement sont gérées localement (au niveau des appareils D2D qui demandent du contenu), évitant ainsi les transmissions de contrôle vers et depuis la BS.

1.4 Contributions du mémoire et choix de méthode

Nos contributions peuvent être résumées comme suit :

1. Nous formulons le problème d'appariement D2D dans les réseaux cellulaires sous-jacents activés pour le cache comme un problème d'appariement en tenant compte des variations des conditions de canal et des emplacements des appareils.
2. Nous proposons une nouvelle solution MARL, basée sur la méthode basée sur la valeur QMIX, pour l'appariement D2D. On a choisi le MARL car c'est une méthode d'apprentissage automatique distribuée qui ne nécessite pas une connaissance globale de l'état du système, y compris l'emplacement des périphériques, l'état de la mise en cache, les conditions des canaux, etc., qui sont très difficiles à collecter en temps opportun et sans dégrader la qualité des transmissions, en particulier dans un système à grande échelle. Par conséquent, le MARL serait plus appropriées, où les décisions de mise en cache et de livraison sont prises localement, ce qui réduit la complexité de traitement et garantit un suivi plus précis des conditions du système.
3. Les performances de l'approche proposée sont comparées aux valeurs de référence en termes de débit total des appareils D2D, à savoir les algorithmes

d'apprentissage basés sur la valeur Q distribué et de réseau Q profond. Ensuite, l'impact de plusieurs paramètres est étudié, notamment le nombre d'appareils D2D, la taille de la bibliothèque de fichiers de contenu, la capacité de mise en cache, etc.

1.5 Plan du mémoire

Nous organisons ce mémoire comme suit :

Le deuxième chapitre fournit un aperçu des communications *D2D*, et plus précisément l'aspect de mise en cache de contenu dans les réseaux *D2D*. À partir de là nous définissons les problématiques et soulignons l'intérêt de l'utilisation de la mise en cache dans les réseaux *D2D*.

Le troisième chapitre présente les différents aspects de l'apprentissage automatique, et ses applications dans les réseaux sans-fil. Il détaille par la suite les différents algorithmes d'apprentissage existant et ceux qu'on va utiliser.

Le quatrième chapitre se concentre sur la présentation de l'état de l'art en relation avec notre problématique et sur l'analyse de notre contribution. On présente la formulation du problème, la solution proposée ($MARL$), et on discute les résultats obtenus.

Enfin, le dernier chapitre conclut ce mémoire et expose nos travaux futurs.

CHAPITRE II

LES COMMUNICATIONS D2D ET LA MISE EN CACHE

2.1 Introduction

Afin de mieux cerner l'étendu de notre problématique, nous présentons dans ce premier chapitre les concepts liés à la communication D2D, et la mise en cache du contenu dans les réseaux D2D. L'objectif est de les décrire et expliquer les liens qui existent entre eux, et l'importance de mettre en cache du contenu dans les réseaux D2D.

2.2 Les communications D2D

2.2.1 Définition

Une communication *D2D* permet à deux appareils proches de communiquer directement entre eux sans passer par une station de base (*BS*). Elle joue un rôle essentiel dans la réalisation des normes de communication établies par la *5G* en raison de sa capacité de partage des ressources. Les communications D2D promettent de fournir un débit de données élevé, de garantir une qualité de service, et contrôler la congestion du réseau.(Song *et al.*, 2016)

Bien que les communications *D2D* ont été absents dans les trois premières généra-

tions de communication sans fil (1 G, 2 G, 3 G), les opérateurs de réseau sont de nos jours attirés par la technologie *D2D*, en raison des tendances évolutives sur le marché du sans fil. Elle fournit également une solution de déchargement efficace aux opérateurs de réseaux mobiles (MNOs).(Astely *et al.*, 2013; Tehrani *et al.*, 2014)

2.2.2 Architecture

L'architecture d'un réseau D2D de base est divisée en trois parties : réseau local, gestion de réseau et applications D2D.(Alkurd *et al.*, 2014)

L'architecture fondamentale est représentée dans la figure 2.1. On retrouve un grand nombre de dispositifs, communiquant entre eux via des liaisons directes, et qui font partie du réseau de zone D2D. L'agrégateur est une entité qui regroupe plusieurs grandeurs ou flux en un seul (Wikipédia, 2020a). Il est disponible dans l'architecture réseau, collecter les données de tous les appareils D2D et, après agrégation, se connecte au réseau d'accès. Par la suite, les données sont envoyées à la passerelle, qui se connecte au réseau d'accès. Le réseau d'accès peut être câblé ou sans fil. Les appareils sont connectés aux fournisseurs de services par le biais du réseau d'infrastructure.(Gandotra *et al.*, 2017)

2.2.3 Types de communications D2D

Les réseaux cellulaires de 5G, qui englobent les communications D2D, sont considérés comme des réseaux à deux niveaux : un niveau de macrocellules et un niveau d'appareils. Le niveau des macrocellules implique des communications de la BS à l'appareil comme dans un système cellulaire conventionnel. Le niveau d'appareils implique les communications D2D. Si un appareil se connecte au réseau cellulaire via une BS, cet appareil est censé fonctionner au niveau des macrocellules. Si un

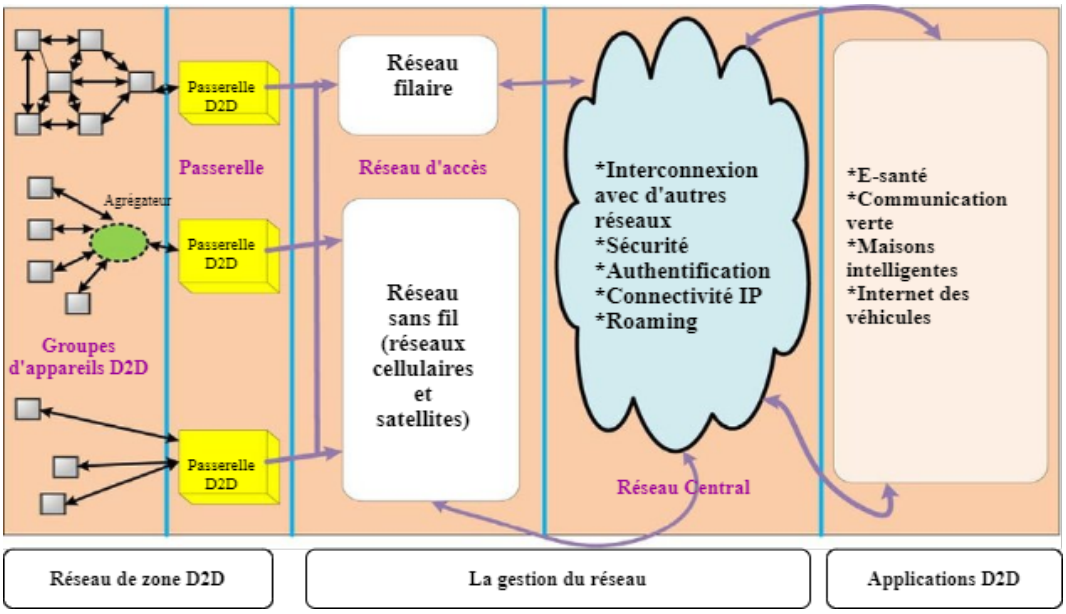

Figure 2.1 Architecture fondamentale de réseau pour les communications D2D (Gandotra *et al.*, 2017)

appareil se connecte directement à un autre appareil ou réalise sa transmission via l'assistance d'autres appareils, ces appareils sont censés être dans le niveau de l'appareil. Dans ce dernier, la station de base peut avoir différents niveaux de contrôle dans le réseau, qui peuvent être complets ou partiels, ou sans contrôle. Par conséquent, la communication D2D peut être classée en quatre types, selon le degré d'implication de la station de base, et sont discutées ci-dessous : (Tehrani *et al.*, 2014; Gandotra *et al.*, 2017)

1. Utilisation de relais avec établissement de liaison contrôlée par l'opérateur : Un appareil au bord d'une cellule ou dans une zone de couverture médiocre peut communiquer avec la BS en relayant ses informations via d'autres appareils comme démontré dans la figure 2.2. Cela permet à l'appareil d'atteindre une QoS plus élevée ou une plus grande autonomie de la batterie. L'opérateur communique avec les dispositifs de relais pour l'établissement

d'une liaison de contrôle partielle ou totale.

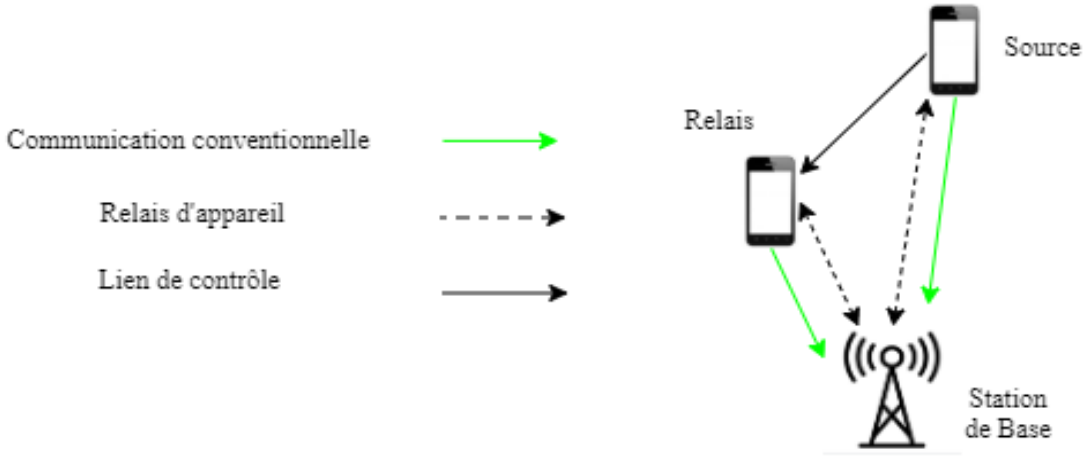

Figure 2.2 Relais d'appareil avec établissement de liaison contrôlée par l'opérateur (Tehrani *et al.*, 2014)

2. Communication D2D directe avec établissement de liaison contrôlée par l'opérateur : Les dispositifs source et de destination se parlent et s'échangent des données sans avoir besoin d'une BS, mais ils sont assistés par l'opérateur pour l'établissement de la liaison. Ceci est démontré dans la figure 2.3

3. Utilisation de relais avec établissement de liaison contrôlée par l'appareil : L'opérateur n'est pas impliqué dans le processus d'établissement de liaison. Par conséquent, les appareils source et de destination sont responsables de la coordination de la communication à l'aide de relais. Ceci est démontré dans la figure 2.4

4. Communication D2D directe avec établissement de liaison contrôlée par l'appareil : Les appareils source et de destination communiquent directement sans aucune intervention de l'opérateur comme montré dans la figure 2.5. Par conséquent, les appareils source et de destination doivent utiliser

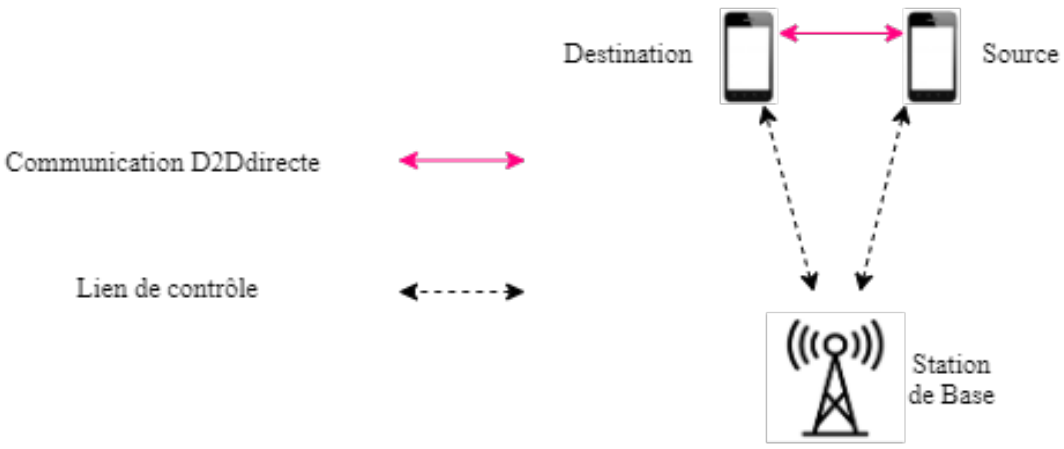

Figure 2.3 Communication D2D directe avec établissement de liaison contrôlée par l'opérateur (Tehrani *et al.*, 2014)

les ressources de manière à garantir une interférence limitée avec les autres appareils du même niveau et du niveau macrocellule.

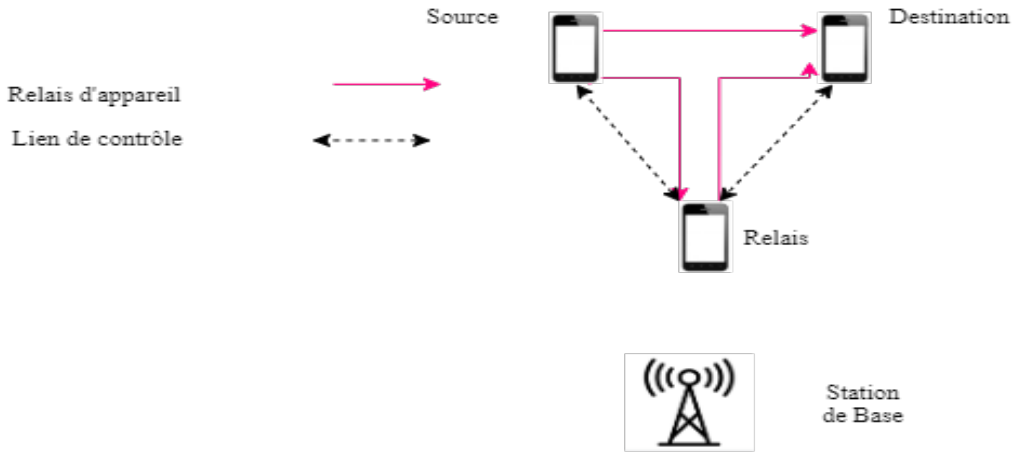

Figure 2.4 Relais d'appareil avec établissement de liaison contrôlée par l'appareil (Tehrani *et al.*, 2014)

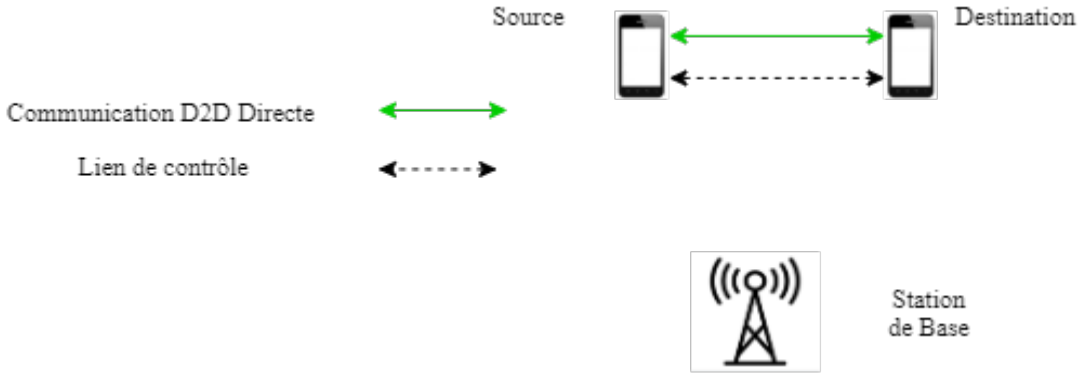

Figure 2.5 Communication D2D directe avec établissement de liaison contrôlée par l'appareil (Tehrani *et al.*, 2014)

2.2.4 Classification des communications D2D

On peut classer les communications *D2D* par rapport à leur utilisation du spectre de fréquence en deux catégories : *D2D Inband* (se produisant sur le spectre cellulaire) et *D2D Outband* (se produisant sur le spectre sans licence) (Jameel *et al.*, 2018).

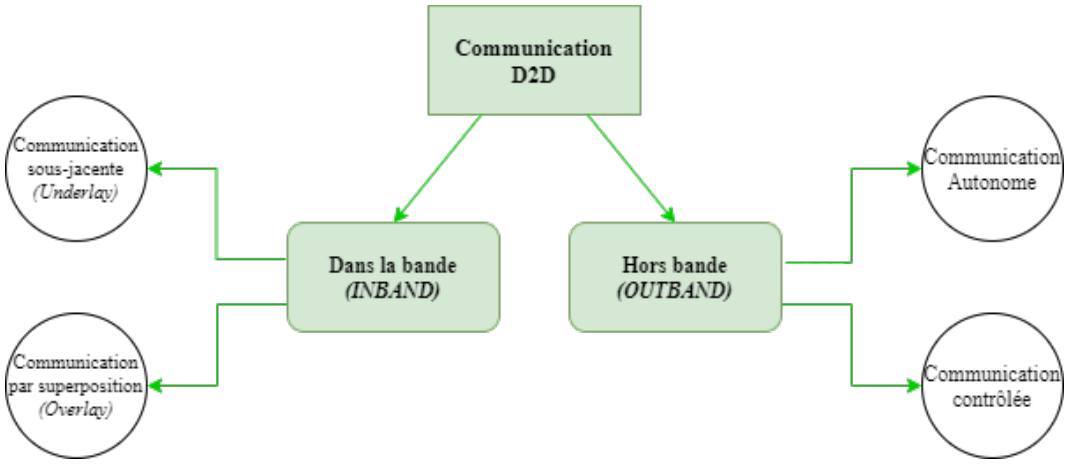

Figure 2.6 Types de communication D2D (basés sur l'accès au spectre). (Gandotra *et al.*, 2017)

Les communications *D2D Inband* peuvent être classées comme sous-jacentes (*Underlay*) ou par superposition (Overlay), tandis que les communications *D2D Outband* peuvent être autonomes ou contrôlées comme montré dans la figure 2.6 (Jameel *et al.*, 2018)

1. La communication *D2D Inband* : Le spectre cellulaire est partagé par les communications *D2D* et cellulaires. Cette communication est en outre classée en sous-jacente (Underlay) et par superposition (Overlay).
   (a) Communication sous-jacente (*Underlay*) : Dans ce cas, les équipements utilisateur D2D (DUE) accèdent de manière opportuniste aux ressources occupées par les utilisateurs cellulaires, ce qui améliore l'efficacité spectrale. Des blocs de ressources dédiés sont attribués aux utilisateurs cellulaires et l'émetteur D2D réutilise ces blocs de ressources pour une communication directe, mais il faudra faire attention à l'interférence. (Vitale *et al.*, 2015)
   (b) Communication par superposition (*Overlay*) : Dans ce cas, une partie du spectre cellulaire est dédiée à la communication D2D. Cela réduit

le problème d'interférence car les deux types de communications ont lieu dans des bandes spectrales distinctes. L'avantage de ce système est qu'il améliore l'ordonnancement et le contrôle de la puissance en communication directe D2D et qu'il offre une efficacité spectrale et une force de signal améliorées dans les réseaux assistés par relais.(Fodor *et al.*, 2012)

2. La communication *D2D Outband* : Les appareils cellulaires utilisent le spectre cellulaire sous licence pour la communication tandis que la communication D2D a lieu via un spectre sans licence, cette communication peut être autonome ou contrôlée.

   (a) Communication contrôlée : Dans ce type de communication, la coordination entre les interfaces radio telles que Bluetooth, ZigBee ou Wi-Fi Direct est contrôlée par le réseau cellulaire. Les ressources de spectre sont pré-allouées aux utilisateurs D2D afin qu'ils puissent lutter et utiliser équitablement les ressources de la bande ISM. [1] (Bin Zhou *et al.*, 2013)

   (b) Communication autonome : Les liaisons cellulaires sont contrôlées par la station de base tandis que les appareils communiquant en mode D2D sont responsables du contrôle de la communication D2D. Cette approche réduit considérablement la charge de travail du réseau cellulaire et comme aucun changement majeur n'est requis pendant le déploiement de BS, il s'agit également d'une solution intéressante pour les opérateurs et les fournisseurs de services mobiles.

---

1. Les bandes ISM (industriel, scientifique et médical) sont des bandes de fréquences qui peuvent être utilisées dans un espace réduit pour des applications industrielles, scientifiques, médicales, domestiques ou similaires.(Wikipédia, 2020b)

Tableau 2.1 Comparaison entre les communications D2D *Inband* et *Outband*.

Communication Inband	**Communication Outband**
Efficacité spectrale améliorée	Pas d'amélioration de l'efficacité spectrale
Contrôle par la station de Base	Pas de contrôle de la station de base
Niveau élevé d'interférence entre les DUEs et les CUEs	Pas d'interférence entre les DUEs et les CUEs
Aucune interface supplémentaire requise par l'appareil	Deux interfaces nécessaires à un appareil pour son bon fonctionnement
Possibilité élevée de gaspillage de ressources	Allocation plus facile des ressources et aucun gaspillage
Aucun codage / décodage impliqué, car une seule interface est utilisée	Le codage et le décodage des paquets sont essentiels

Le tableau 2.1 résume les principales différences entre les communications *Inband* et *Outband* (Gandotra *et al.*, 2017; Hayat *et al.*, 2019).

2.2.5 Métriques de performance

Dans cette section, on présente un ensemble de mesures de performance couramment utilisées pour évaluer les performances d'un réseau utilisant les communications D2D. La majorité de ces métriques sont principalement liés au rapport signal-sur-interférence-plus-bruit (SINR). (Gandotra *et al.*, 2017)

- Débit système : Il est défini comme le débit global de toutes les paires D2D et des utilisateurs cellulaires d'un système cellulaire. Une valeur de débit plus élevée signifie de meilleures performances. Il est mesuré en bits / sec.

C'est un indicateur d'un transfert d'informations réussi entre les paires. C'est la métrique sur laquelle on a travaillé dans (Moussaid *et al.*, 2018) et qu'on va présenter au quatrième chapitre.

- Capacité d'utilisateurs D2D : C'est le nombre de DUE qui peuvent être pris en charge pour un ensemble donné d'utilisateurs cellulaires dans le réseau, sous réserve de la contrainte de débit de données maximale. Une valeur de capacité utilisateur plus élevée est toujours souhaitable.
- L'équité (*Fairness*) : C'est un indicateur très critique pour évaluer l'équité d'un système donné (généralement l'allocation des ressources), pour la communication D2D.
- Taux de confidentialité *(Secrecy rate)* : Il s'agit d'un paramètre important pour évaluer les performances de confidentialité d'un réseau. Elle peut être considérée comme analogue à la capacité de canal traditionnelle, sous réserve de contraintes de confidentialité. La maximisation du taux de confidentialité est une priorité dans les réseaux cellulaires
- Score d'opinion moyen (MOS) : Il est utilisé pour évaluer les performances d'un réseau cellulaire, en termes de qualité d'expérience. La plage de MOS se situe entre 1 et 5. Des valeurs proches de 5 sont souhaitables. Cela signifie une excellente qualité d'information.
- Latence : C'est un indicateur du délai entre la transmission et la réception des informations. La communication D2D entraîne généralement une latence plus faible, en raison de la transmission sur une petite distance.
- Efficacité énergétique : il s'agit du rapport entre le débit du système et la consommation électrique par unité de surface. Il s'agit d'un indicateur de l'utilisation efficace de l'énergie (alimentation par batterie) au sein d'un réseau cellulaire.
- Efficacité spectrale : Le nombre de bits transmis par unité de bande passante indique l'efficacité spectrale. Il est mesuré en Bits/secondes/Hz .Il

est essentiel pour la quantification des performances du réseau D2D

2.3 La mise en cache du contenu dans les réseaux D2D

Les principaux défis de la 5G sont d'assurer une faible latence, d'atteindre des débits de données élevés, et d'assurer une connectivité à un nombre très élevé d'appareils. Les communications D2D permettent aux utilisateurs de se connecter directement entre eux sans avoir de connectivité avec la station de base. D2D traite les équipements utilisateur (UE) comme des concentrateurs de données pour le partage de contenu. Pour réduire la pression sur le réseau principal, la mise en cache sur l'appareil de l'utilisateur pour le partage de contenu via la communication D2D devient inévitable. (Shen, 2015)

2.3.1 Définition

La mise en cache des contenus et services populaires sur les appareils rapproche le contenu de son utilisateur final. Comme les appareils des utilisateurs disposeent de nos jours d'assez grandes capacités de stockage, ils peuvent partager le contenu populaire avec leurs pairs via des communications D2D.(Bastug *et al.*, 2014)

Lorsqu'un utilisateur demande un contenu déjà mis en cache sur son stockage local, le contenu peut être récupéré à partir du cache, sans délai et sans générer d'interférence avec d'autres utilisateurs. Sinon, le contenu peut être transmis à l'utilisateur via un lien unique qui est établi entre l'appareil qui demande du contenu et l'appareil qui a ce contenu en cache. De cette façon, la mise en cache préalable peut améliorer la qualité de service de tous les utilisateurs, directement ou indirectement, et améliorer le débit sans fil par déchargement.(Liu *et al.*, 2016)

2.3.2 Architecture de mise en cache de contenu D2D

Une architecture typique des techniques de mise en cache de contenu pour les communications D2D en 5G comprend des utilisateurs intelligents et des appareils IoT tels que smartphones, ordinateurs portables, PC, maisons intelligentes, véhicules, etc. À chaque fois qu'une demande de contenu est lancée par des utilisateurs, les données sont extraites des caches de dispositifs proches, puis sur les stations de base situées sur la périphérie du réseau.

La figure 2.7 illustre l'architecture des emplacements de mise en cache pour D2D dans la 5G. Elle montre les fonctionnalités de base du système de cache, ayant une architecture à quatre couches. La couche inférieure est la couche de base qui contient différents types d'utilisateurs et d'appareils IoT qui communiquent via des liens D2D pour le partage de contenu. Lorsqu'un échec de cache (Un échec de cache se produit lorsque les données extraites ne sont pas présentes dans le cache) se produit lors de la récupération du contenu demandé dans des appareils à proximité, les stations de base présentes au niveau de la couche de périphérie sont sollicitées et le partage de contenu a lieu à l'aide des routeurs Wi-Fi. La couche Middleware agit comme un fournisseur de services. Elle se compose de divers services cloud comme Amazon Web Service (AWS), Microsoft Azure, Google Cloud Platform, Jasper, IBM Watson, etc. La couche supérieure est appelée couche cloud et utilisée pour fournir les données demandées lorsque les stations de base ne peuvent pas trouver le contenu requis. La communication avec la couche cloud est établie à l'aide des connexions à large bande.(Prerna *et al.*, 2020)

2.3.3 Méthodes de mise en cache

On a dit auparavant que la mise en cache a été proposée pour minimiser la latence et alléger la charge sur le réseau. Il existe différentes méthodes pour effectuer la

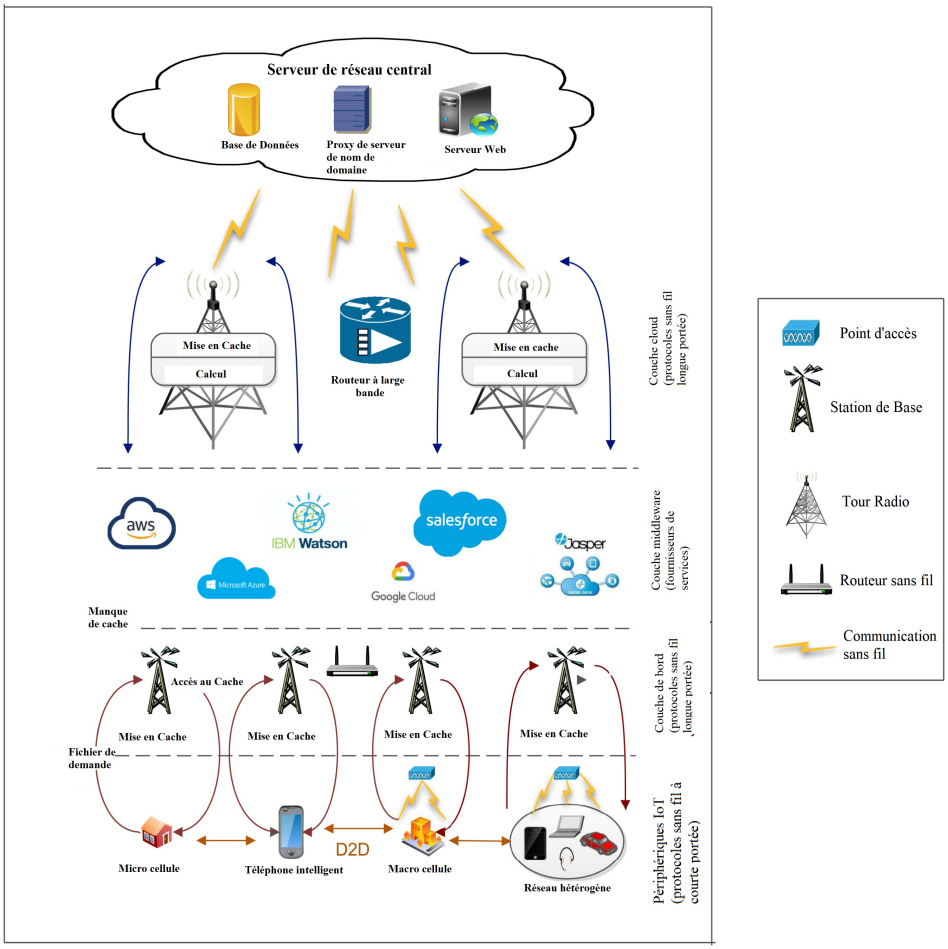

Figure 2.7 Architecture de mise en cache de contenu D2D en 5G. (Prerna *et al.*, 2020)

mise en cache dans un réseau 5G qui sont présentées comme suit : (Prerna *et al.*, 2020)

- Mise en cache basée sur la popularité : La mise en cache des fichiers sur la base de la popularité se réfère d'abord à la mise en cache des contenus les plus populaires, suivis des moins populaires, etc. La popularité des fichiers est calculée avec des fichiers ayant une mesure de similitude du contenu mis en cache
- Mise en cache coopérative : C'est un processus dans lequel plusieurs ap-

pareils partagent et coordonnent les données de cache entre eux sans se rendre à la BS. Elle peut jouer un rôle important pour gérer les demandes (Zhang *et al.*, 2018).

- Mise en cache hiérarchique : Il existe différents niveaux dans le cache en fonction de la proximité du contenu à placer.
- Mise en cache socialement consciente : La mise en cache peut être effectuée sur la base de la liaison sociale et du comportement de l'utilisateur. La popularité du contenu et le comportement des utilisateurs sont analysés pour la mise en cache et l'accès au contenu à partir de différents emplacements (Ma *et al.*, 2018).
- Mise en cache compétitive : Les contenus sont stockés sur la base d'une politique tarifaire utilisée par l'opérateur.

2.3.4 Types de mise en cache

La mise en cache à l'aide de la communication D2D en 5G est généralement classée en trois types : Mise en cache synchrone, mise en cache asynchrone et mise en cache hybride.(Prerna *et al.*, 2020)

- Mise en cache synchrone : Dans ce type, la BS conserve les informations de toutes les unités mobiles connectées. De plus, elle diffuse des messages sur le réseau pour chaque entrée de données mise en cache, et aucune unité mobile n'est autorisée à demander le cache avant le message de diffusion.
- Mise en cache asynchrone : La mise en cache asynchrone diffuse uniquement des messages pour la saisie de données et ne conserve aucune information sur les unités mobiles connectées.
- Mise en cache hybride : L'approche hybride conserve les informations des unités mobiles dans sa cellule et ne permet à aucune unité mobile d'accéder au cache jusqu'à ce que l'accusé de réception suivant soit reçu.

Tableau 2.2 Comparaison entre les types de mise en cache (Prerna *et al.*, 2020)

Synchrone	Asynchrone	Hybride
Le serveur conserve toutes les informations des unités mobiles	Ne conserve pas d'informations des unités mobiles	Le serveur ne conserve aucune information sur les unités mobiles
Le retard moyen est plus élevé mais utilise efficacement la liaison descendante	L'unité mobile ne peut pas envoyer de demande d'accès au cache tant que le prochain accusé de réception n'a pas été reçu	Le retard moyen est moins mais plus de trafic sur la liaison descendante

Le tableau 2.2 présente une comparaison entre les trois types de mise en cache en terme de l'emplacement de l'information et du délai.

2.3.5 Contraintes

Comme la 5G exige des débits de transmission élevés, on a vu que le fait d'amener le contenu proche des utilisateurs, leur permet de partager du contenu via la communication D2D. De plus, il réduit l'encombrement du réseau et réduit la latence. En outre, il existe diverses contraintes lors de la mise en cache résumées comme suit (Prerna *et al.*, 2020) :

- ⊗ Mobilité des utilisateurs : Comme les utilisateurs continuent de se déplacer, la vitesse détermine le temps de contact entre deux utilisateurs et a un impact sur le mécanisme de mise en cache.
- ⊗ Conscience sociale : Les utilisateurs préfèrent uniquement se connecter avec des utilisateurs qu'ils connaissent ou en qui ils ont confiance. Ainsi, la rela-

tion entre la mise en cache et la conscience sociale joue un rôle important.

- ⊗ Stockage : La capacité de cache de l'appareil de l'utilisateur joue un rôle majeur lors de la mise en cache en 5G. Bien que de grands espaces de stockage soient désormais disponibles à bas prix, ils restent limités par rapport à la demande. Par conséquent, la quantité de contenu mise en cache dépend de la capacité de cache de l'équipement.
- ⊗ Consommation d'énergie : Il existe beaucoup de partage de contenu dans la 5G, qui maintient les appareils mobiles connectés et vide la batterie de l'équipement qui met en cache. Des algorithmes de mise en cache écoénergétiques sont nécessaires pour limiter la consommation d'énergie.

2.4 Conclusion

Dans ce chapitre, nous avons présenté les notions de base associées aux communications D2D et la mise en cache du contenu dans les réseaux D2D. On a présenté une architecture globale d'un réseau D2D, ensuite on a donné un aperçu sur les types de communication D2D avec relais de l'appareil ou sans relais, et avec contrôle de l'opérateur ou sans contrôle. On a donné une classification des communications D2D en se basant sur le spectre, comme la communication D2D dans la bande, qui peut être classée comme sous-jacente ou par superposition, et la communication D2D hors bande, qui est soit autonome ou contrôlée. Ensuite on a présenté un ensemble de métriques de performance utilisées pour la communication D2D. Dans un second lieu, on a détaillé la mise en cache du contenu dans les réseaux D2D, en présentant une architecture à quatre niveaux, les méthodes de mise en cache, les types de mise en cache, et on a conclu avec des contraintes que présente la mise en cache au niveau des réseaux D2D.

CHAPITRE III

L'APPRENTISSAGE AUTOMATIQUE ET SES APPLICATIONS DANS LES RÉSEAUX SANS-FIL

3.1 Introduction

Dans ce chapitre, on définit l'apprentissage automatique, ses différents types, ensuite on parcourt la littérature dans le domaine des réseaux sans-fil pour tirer quelques applications dans lesquelles l'apprentissage automatique a été utlisé. Finalement on se concentre sur l'apprentissage par renforcement et quelques algorithmes de ce type d'apprentissage automatique.

3.2 L'apprentissage automatique

3.2.1 Définition

L'apprentissage automatique (ML) est défini par le magazine *LeBigData* (Bastien, 2018) comme une technologie d'intelligence artificielle permettant aux ordinateurs d'apprendre sans avoir été programmés explicitement à cet effet. Pour apprendre et se développer, les ordinateurs ont toutefois besoin de données à analyser et sur lesquelles s'entraîner. L'intelligence artificielle (IA) est un terme large qui fait référence aux systèmes ou machines qui imitent l'intelligence humaine. L'apprentissage automatique et l'IA sont souvent discutés ensemble, et les termes sont

parfois utilisés de manière interchangeable, mais ils ne signifient pas la même chose. (Oracle, 2020)

L'apprentissage automatique est récemment devenu une technologie populaire. Dans les années 90 et 2000, Internet a transformé notre façon de vivre et de faire des affaires et, ce faisant, a généré de nombreux pétaoctets de données. L'apprentissage automatique et l'analyse prédictive révolutionnent à nouveau notre société en transformant ces données en prévisions utiles. Un certain nombre d'applications commerciales importantes sont déjà apparues, notamment les moteurs de recommandation, les systèmes de reconnaissance de la parole et de l'écriture manuscrite, l'identification du contenu, la classification / récupération des images, le sous-titrage automatique, les filtres anti-spam et la prévision de la demande.(Lee *et al.*, 2018)

Le ML est idéal pour les problèmes complexes où les solutions existantes nécessitent beaucoup de réglages manuels, ou pour les problèmes pour lesquels il n'existe aucune solution en utilisant une approche traditionnelle. Ces problèmes peuvent être résolus en apprenant des données, en remplaçant les logiciels conventionnels contenant de longues listes de règles, par des routines ML qui apprennent automatiquement des données précédentes. Une différence importante du ML par rapport aux algorithmes cognitifs traditionnels est l'extraction automatique des fonctionnalités, grâce à laquelle une ingénierie de fonctionnalités artisanale coûteuse peut être supprimée. D'une manière générale, une tâche ML peut détecter des anomalies, prédire des scénarios futurs, s'adapter à des environnements fluctuants, avoir un aperçu des problèmes complexes avec de grandes quantités de données et, en général, découvrir les modèles qu'un humain peut manquer. (Morocho-Cayamcela *et al.*, 2019)

3.2.2 Types d'algorithmes d'apprentissage automatique

L'apprentissage automatique peut être classé en apprentissage supervisé, apprentissage non supervisé et apprentissage par renforcement. On présente dans ce qui suit la différence entre les trois types (Morocho-Cayamcela *et al.*, 2019) :

- Apprentissage supervisé : L'apprentissage supervisé utilise des ensembles de données de formation étiquetés pour créer des modèles. Il existe différentes méthodes d'étiquetage des ensembles de données connues sous le nom de vérité au sol. Cette technique d'apprentissage est utilisée pour «apprendre» à identifier des modèles ou des comportements dans les ensembles de données d'apprentissage «connus». En règle générale, cette approche est utilisée pour résoudre les problèmes de classification et de régression qui se rapportent à la prédiction de résultats valorisés discrets ou continus, respectivement. En revanche, il est possible d'employer des techniques de ML semi-supervisées face à des connaissances partielles. C-à-d, avoir des étiquettes incomplètes pour les données de formation ou des étiquettes manquantes.
- Apprentissage non-supervisé : L'apprentissage non supervisé utilise des ensembles de données d'apprentissage non étiquetés pour créer des motifs qui peuvent distinguer les modèles dans les données. Cette approche est particulièrement adaptée aux problèmes de clustering. Par exemple, les problèmes de détection de valeurs aberrantes et d'estimation de densité dans les réseaux peuvent concerner le regroupement de différentes instances d'attaques en fonction de leurs similitudes.
- Apprentissage par renforcement (RL) : Il diffère de l'apprentissage supervisé en ce qu'il utilise des commentaires évaluatifs de l'environnement pour estimer les récompenses ou les coûts réels par rapport aux commentaires instructifs utilisés dans l'apprentissage supervisé par des erreurs de clas-

sification. Dans l'apprentissage par renforcement, les événements d'entrée affectent les décisions prises à des moments ultérieurs et les événements de sortie qui en résultent, ce qui est inhérent à la nature d'un système dynamique. Une étape itérative de l'apprentissage par renforcement est généralement composée de deux étapes : l'évaluation des politiques, dans laquelle les conséquences à long terme des décisions actuelles sont caractérisées par un critique, suivi par l'amélioration des politiques, qui tente de modifier la politique actuelle en fonction des commentaires d'évaluation du critique

3.3 Application de l'apprentissage automatique dans les réseaux sans-fil

Dans cette section, on présente les applications de chaque type d'apprentissage automatique dans les réseaux sans-fil, ainsi que les modèles utilisés par chacun de ces types.

3.3.1 L'apprentissage supervisé

L'apprentissage supervisé avec ses différents modèles d'apprentissage a été appliqué plusieurs fois dans les communications mobiles et sans fil, on présente ci-dessous un exemple d'application de chaque modèle d'apprentissage :

- Régression logistique statistique : ce modèle a été appliqué dans (Azmat *et al.*, 2016) pour l'allocation dynamique de fréquence et de bande passante dans les déploiements de petites cellules LTE (*Long Term Evolution*).
- Les machines à vecteurs de support : Dans (Sanchez-Fernandez *et al.*, 2004), ce modèle a été utilisé pour classifier les informations d'état du canal pour sélectionner les indices d'antenne optimaux dans la technologie MIMO (*Multiple Input, Multiple Output*)

- Réseaux de neurones profonds : il a été utilisé dans (Alkhateeb *et al.*, 2018) pour la prédiction et la coordination des vecteurs de formation de faisceau à la BS.

3.3.2 L'apprentissage non-supervisé

Les modèles d'apprentissage de ce type ont été appliqué dans la littérature comme suit :

- Classification hiérarchique : ce modèle a été appliqué dans (Parwez *et al.*, 2017) pour la détection d'anomalies, de pannes et d'intrusions dans les réseaux sans fil mobiles.
- Apprentissage bayésien non paramétrique : ce modèle a été appliqué dans (Bastug *et al.*, 2014) pour la réduction du trafic dans un réseau sans fil en répondant de manière proactive aux demandes prévisibles des utilisateurs via la mise en cache sur la BS.
- Apprentissage de carte auto-organisé : ce modèle a été utilisé dans (Gazda *et al.*, 2018) pour planifier la couverture de réseaux hétérogènes avec des clusters dynamiques.

3.3.3 L'apprentissage par renforcement

De même pour ce type d'apprentissage automatique, il a été appliqué plusieurs fois dans les réseaux sans-fil comme suit :

- Q-learning : Dans (Sadeghi *et al.*, 2018), il a été utilisé pour permettre aux utilisateurs de sélectionner une BS qui va les servir en exploitant ses données locales et les résultats d'apprentissage des utilisateurs voisins.
- Apprentissage profond par renforcement : Il a été utilisé dans (Han *et al.*, 2017) pour que les utilisateurs secondaires décident du canal de communi-

cation et de la mobilité.

- L'apprentissage par renforcement multi-agent MARL : Il a été utilisé dans (Gengtian *et al.*, 2020) pour contrôler la puissance dans un réseau D2D.

Dans le cadre de ce projet, nous allons utiliser et comparer ces trois modèles d'apprentissage par renforcement. La section suivante présente plus de détails sur l'apprentissage par renforcement et les trois modèles cités ci-dessus.

3.4 Modèle d'apprentissage par renforcement

Le RL comporte quatre éléments essentiels :

- Agent : On définit l'agent intelligent comme étant une entité autonome capable d'apprendre ou utiliser des connaissances pour pouvoir réaliser ses objectifs.
- Environnement : Le monde, réel ou virtuel, dans lequel l'agent effectue des actions.
- Action : Un mouvement effectué par l'agent, qui provoque un changement d'état dans l'environnement.
- Récompense : L'évaluation d'une action, qui peut être positive ou négative.

Le modèle de base du RL est illustré dans la figure 3.1. L'Agent peut percevoir l'environnement et choisir une action pour obtenir la plus grande valeur de récompense en interagissant en continu avec l'environnement. L'interface interactive de l'agent intelligent et de l'environnement comprend l'action, la récompense et l'état (Qiang et Zhongli, 2011).

À chaque fois que le système d'apprentissage par renforcement interagit avec l'environnement, il accepte d'abord l'entrée de l'état d'environnement s, puis la sortie de l'action a agit sur l'environnement selon le mécanisme d'inférence interne. Enfin, l'environnement passe au nouvel état s' après avoir accepté l'action. Le système

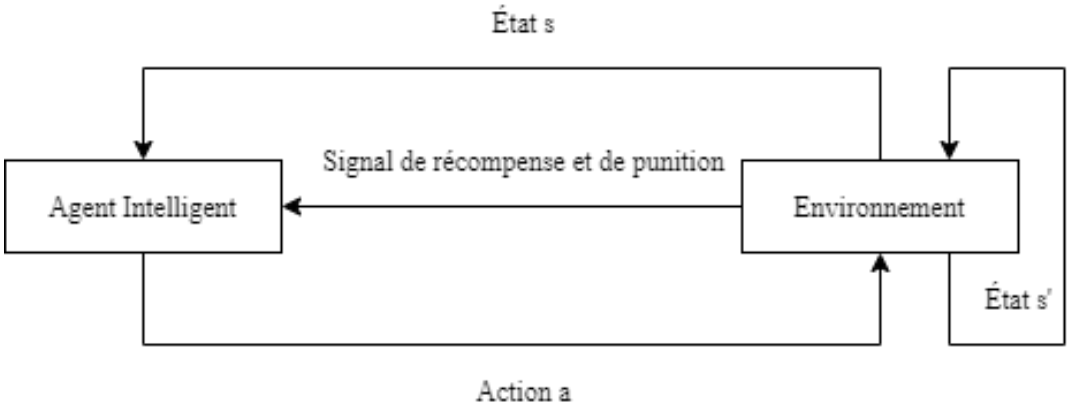

Figure 3.1 Le modèle de base du RL

accepte l'entrée du nouvel état s' et obtient la récompense ou la punition r à partir de l'environnement du système.

Le but du système d'apprentissage par renforcement est d'apprendre une stratégie d'action nommée $\pi : S \rightarrow A$, où S représente l'ensemble des états, et A l'ensemble des actions. La stratégie π permet d'obtenir la plus grande valeur de récompense cumulée de l'environnement. Elle peut être représentée par la formule suivante : $\sum_{i=0}^{\infty} \gamma^i r_{i+1}$, γ est appelé le facteur d'actualisation (*Discount factor*). L'idée de base derrière l'apprentissage par renforcement est la suivante : si l'action d'un certain système provoque la récompense positive de l'environnement, le système générant cette action récemment renforcera la tendance, il s'agit d'un processus de rétroaction positive ; sinon, le système générant cette action atténuera cette tendance (Qiang et Zhongli, 2011).

Généralement, le RL est utilisé pour résoudre les problèmes de décision de Markov (MDP) qu'on présente dans la prochaine sous-section.

3.4.1 Processus de décision de Markov (MDP)

MDP est un cadre qui peut résoudre la plupart des problèmes d'apprentissage par renforcement à des actions discrètes. Avec le processus de décision Markov, un agent peut arriver à une politique optimale pour des récompenses maximales au fil du temps (dan lee, 2019).

On peut représenter un MDP par un tuple $<S, A, T, R>$ où :

- S : Un ensemble fini d'états $\{s^1, \cdots, s^N\}$ où la taille de l'espace d'état est N. Un état $s \in S$ est une caractérisation unique de tout ce qui est important dans un état du problème modélisé (Otterlo et Wiering, 2012).
- A : Un ensemble fini d'actions $\{a^1, \cdots, a^K\}$ où la taille de l'espace d'actions est K. Les actions peuvent être utilisées pour contrôler l'état du système. L'ensemble des actions qui peuvent être appliquées dans un état particulier $s \in S$, est noté $A(s)$, où $A(s) \subseteq A$. Une action $a \in A$ est applicable dans un état $s \in S$ (Otterlo et Wiering, 2012).
- T : En appliquant l'action $a \in A$ dans un état $s \in S$, le système effectue une transition de s vers un nouvel état $s' \in S$, basé sur une distribution de probabilité sur l'ensemble des transitions possibles. La fonction de transition T est définie comme $T : S \times A \times S \to [0, 1]$, c'est-à-dire que la probabilité de se retrouver dans l'état s' après avoir fait l'action a dans l'état s est notée $T(s, a, s')$. Il est nécessaire que pour toute action a, et tout état s et s', $T(s, a, s') \geq 0$ et $T(s, a, s') \leq 1$. Bien évidemment, pour tout les états et actions, on doit avoir $\sum_{s' \in S} T(s, a, s') = 1$, et donc T définit une distribution de probabilité appropriée sur les prochains états possibles (Otterlo et Wiering, 2012).
- R : C'est la fonction de récompense qui spécifie les récompenses pour être dans un état ou pour effectuer une action dans un état. La fonction de

récompense d'état est définie comme $R : S \rightarrow R$, et elle spécifie la récompense obtenue dans les états. Cependant, il existe deux autres définitions. On peut définir soit $R : S \times A \rightarrow R$ qui donne des récompenses pour effectuer une action dans un état, ou bien $R : S \times A \times S \rightarrow R$ qui donne des récompenses pour des transitions particulières entre les états (Otterlo et Wiering, 2012).

La fonction de transition T et la fonction de récompense R définissent ensemble le modèle du MDP. Souvent, les MDPs sont représentés comme un graphe de transition d'états où les nœuds correspondent aux états et les arêtes (orientés) indiquent les transitions.

3.4.2 La politique d'apprentissage

Dans un MDP $<S, A, T, R>$, on définit une politique comme étant une fonction qui est générée pour chaque état $s \in S$ et action $a \in A$. Formellement, on définit la politique π par la fonction $\pi : S \times A \rightarrow [0, 1]$.

L'application d'une politique à un MDP se fait de la manière suivante. Tout d'abord, un état de départ s_0 est généré à partir de la distribution d'états initiale I. Ensuite, la politique π suggère l'action $a_0 = \pi(s_0)$ et cette action est effectuée. Sur la base de la fonction de transition T et de la fonction de récompense R, une transition est effectuée vers l'état s_1, avec une probabilité $T(s_0, a, s_1)$ et une récompense $r_0 = R(s_0, a_0, s_1)$ est reçue. Ce processus continue, produisant $s_0, a_0, r_0, s_1, a_1, r_1, s_2, a_2, ...$, et il se termine quand il atteint l'objectif.

La politique fait partie de l'agent et son objectif est de contrôler l'environnement modélisé comme un MDP (Otterlo et Wiering, 2012).

3.4.3 Critères d'optimisation et actualisation

Le but de l'apprentissage dans un MDP est de récolter des récompenses. Si l'agent ne se préoccupait que de la récompense immédiate, un critère d'optimalité simple serait d'optimiser $E[r_t]$. Cependant, il existe plusieurs façons de prendre en compte l'avenir pour savoir comment se comporter dans l'instant présent. Il existe essentiellement trois modèles d'optimalité dans le MDP, qui sont suffisants pour couvrir la plupart des approches de la littérature (Otterlo et Wiering, 2012) :

- Le modèle d'horizon fini : Il prend un horizon fini de longueur h et indique que l'agent doit optimiser sa récompense attendue sur cet horizon, on peut le modéliser par $E[\sum_{t=0}^{h} r_t]$. Cependant, le problème avec ce modèle, est que le choix optimal pour la longueur d'horizon h n'est pas toujours connu.
- Le modèle à horizon infini : Dans ce modèle, la récompense à long terme est prise en compte, mais les récompenses reçues à l'avenir sont actualisées en fonction du temps dont lequel elles seront reçues, on le représente par $E[\sum_{t=1}^{\infty} \gamma^t r_t]$, γ est le facteur d'actualisation avec $0 \leq \gamma < 1$. Ce modèle est mathématiquement plus pratique, mais conceptuellement similaire au modèle à horizon fini.
- Le modèle de récompense moyenne : Le troisième modèle d'optimalité est le modèle de récompense moyenne, maximisant la récompense moyenne à long terme, on le représente par $\lim_{h \to +\infty} \frac{1}{h} E[\sum_{t=0}^{h} r_t]$. Parfois, cela s'appelle la politique de gain optimal et dans la limite, lorsque le facteur d'actualisation approche 1, il est égal au modèle actualisé à horizon infini.

Le choix entre ces critères d'optimalité peut être lié au problème d'apprentissage. Si la durée de l'épisode est connue, le modèle à horizon fini représente le meilleur choix. Cependant, souvent cela n'est pas connu, ou la tâche se poursuit, le modèle à horizon infini est plus approprié.

3.4.4 Fonctions de valeur et équations de Bellman

Dans les sections précédentes, nous avons défini le MDP et les critères d'optimalité qui peuvent être utilsés pour apprendre les politiques optimales. Dans cette section, nous définissons des fonctions de valeur, qui sont un moyen de lier les critères d'optimalité aux politiques. La plupart des algorithmes d'apprentissage pour les MDP calculent des politiques optimales en apprenant une fonction de valeur. Cette dernière représente une estimation de la qualité de l'agent dans un certain état : $V^{\pi}(s) = E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k} | s_t = s\}$ ou de la qualité de l'exécution d'une certaine action dans cet état : $Q^{\pi}(s,a) = E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k} | s_t = s, a_t = a\}$. On utilise E_{π} pour l'espérance sous la politique π (Otterlo et Wiering, 2012).

L'objectif d'un MDP est de trouver la meilleure politique, c'est-à-dire la politique qui reçoit la récompense la plus élevée. Cela revient à maximiser la fonction de valeur $V^{\pi}(s)$ pour tous les états $s \in S$. Une politique optimale π^*, est telle que $V^{\pi^*}(s) > V^{\pi}(s)$ pour tout $s \in S$ et pour toute politique π. On définit dans ce contexte l'*équation d'optimalité de Bellman* comme suit :

$$V^{\pi^*}(s) = \max_{a \in A} \sum_{s' \in S} T(s,a,s') \Big(R(s,a,s') + \gamma V^{\pi^*}(s') \Big) \tag{3.1}$$

l'équation (3.1) déclare que la valeur d'un état sous une politique optimale doit être égale au rendement attendu pour la meilleure action dans cet état.

On définit aussi la fonction de valeur action optimale comme suit :

$$Q^*(s,a) = \sum_{s' \in S} T(s,a,s') \Big(R(s,a,s') + \gamma \max_{a'} Q^*(s',a') \Big) \tag{3.2}$$

La relation entre Q^* et V^* est : $V^*(s) = \max_a Q^*(s,a)$. Autrement dit, la meilleure action est l'action qui a l'utilité attendue la plus élevée en fonction des prochains états possibles résultant de cette action (Otterlo et Wiering, 2012).

3.5 Algorithmes d'apprentissage par renforcement

Maintenant que nous avons défini les MDP, les politiques, les critères d'optimalité et les fonctions de valeur, il est temps de savoir comment calculer les politiques optimales. Résoudre un MDP donné revient à déterminer une politique optimale π^*. On présente dans cette section trois algorithmes qui ont été conçus pour résoudre des MDPs et qu'on utilise dans le chapitre suivant lors de la résolution de la problématique de ce projet.

3.5.1 *Q-Learning*

L'apprentissage Q (en anglais *Q-learning*) est un algorithme d'apprentissage par renforcement hors politique qui cherche à trouver la meilleure action à entreprendre compte tenu de l'état actuel. Il est considérée comme non conforme à la politique, car la fonction de *Q-learning* apprend des actions qui sont en dehors de la politique actuelle, comme prendre des mesures aléatoires, et par conséquent, une politique n'est pas nécessaire. Plus précisément, le *Q-learning* cherche à apprendre une politique qui maximise la récompense totale (Violante, 2019).

L'idée de base du *Q-learning* est d'estimer de manière incrémentielle les fonctions de valeurs Q pour les actions, en fonction des récompenses et de la fonction de valeur Q de l'agent. La règle de mise à jour est une variante du thème de l'apprentissage , utilisant des valeurs Q et un opérateur max intégré sur les valeurs Q de l'état suivant afin de mettre à jour Q_t en Q_{t+1} :

$$Q_{k+1}(s_t, a_t) = Q_k(s_t, a_t) + \alpha\Big(r_t + \gamma \max_a Q_k(s_{t+1}, a_t) - Q_k(s_t, a_t)\Big) \tag{3.3}$$

avec $\alpha \in [0, 1]$ qui peut simplement être défini comme le degré d'acceptation de la nouvelle valeur par rapport à l'ancienne.

L'agent fait un pas dans l'environnement de l'état s_t à s_{t+1} en utilisant l'action a_t

tout en recevant la récompense r_t. La mise à jour a lieu sur la valeur Q de l'action a_t dans l'état s_t à partir duquel cette action a été exécutée. Pour cela, Q-Table est la structure de données utilisée pour calculer les récompenses futures maximales attendues pour l'action à chaque état, ce tableau nous guidera vers la meilleure action dans chaque état. Le processus d'algorithme *Q-learning* est présenté dans la figure 3.2 :

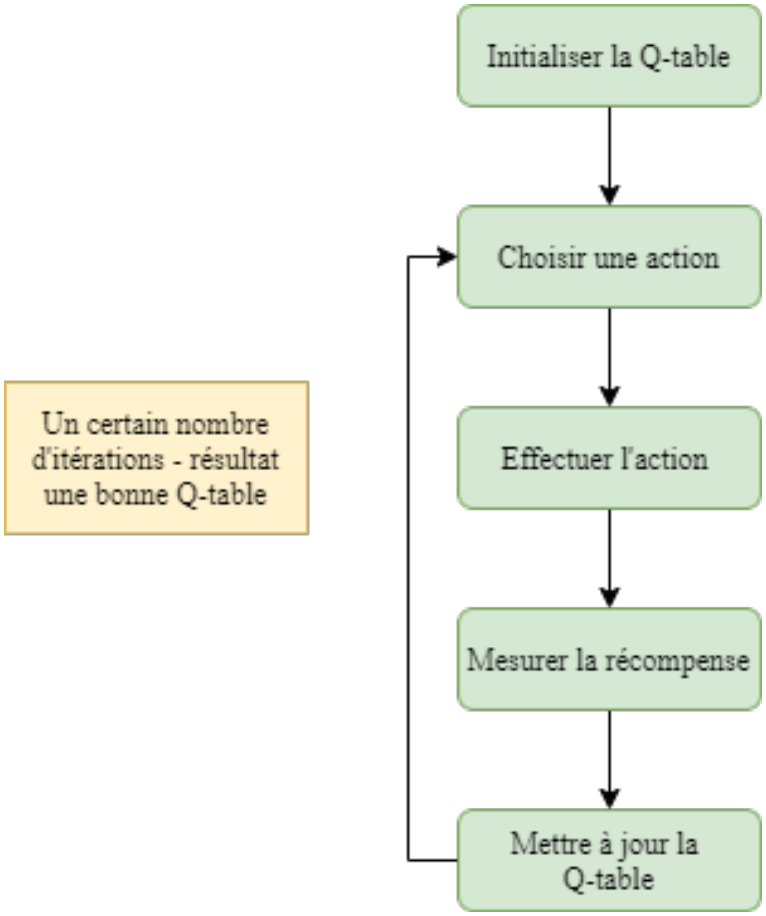

Figure 3.2 Algorithme Q-learning

Cet algorithme n'est pas appliqué dans la littérature, surtout quand il s'agit d'un problème d'une grande dimension. Par exemple, si l'algorithme apprend à jouer à des jeux basés sur l'entrée d'image, le nombre de pixels d'une image peut être

des dizaines de milliers, et les dimensions de l'état sont trop grandes pour être comptées en raison de la différence des composantes de couleur (Li *et al.*, 2019).

3.5.2 *Deep Q-Network*

Le DQN est un algorithme qui est une extension du *Q-learning* et qu'on utilise lorsque la taille de l'espace des états et actions est très grande (Li *et al.*, 2019). Cet algorithme utilise un réseau neuronal (qui est un système dont la conception est à l'origine schématiquement inspirée du fonctionnement des neurones biologiques (Wikipédia, 2020c)) pour avoir une approximation de la fonction de valeur Q. L'état est donné comme entrée et la valeur Q de toutes les actions possibles est générée comme sortie comme présenté dans la figure 3.3 :

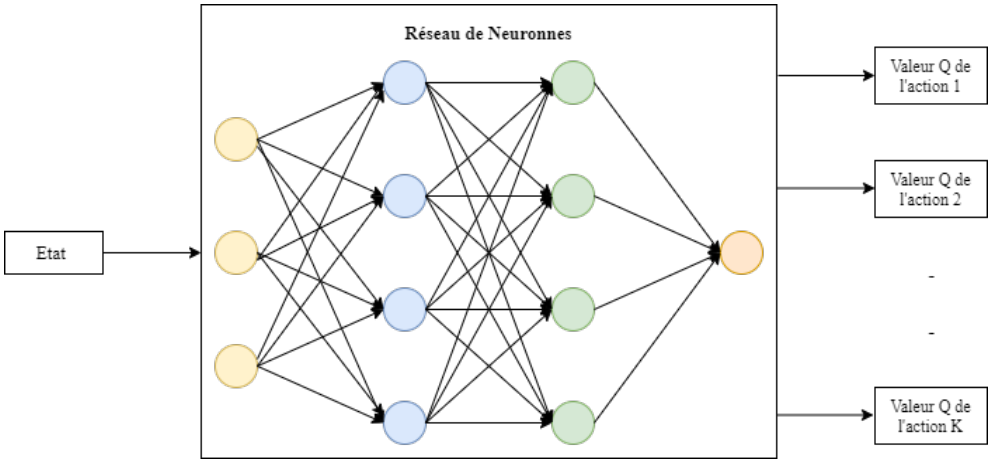

Figure 3.3 DQN

Les étapes de l'apprentissage par renforcement à l'aide de l'algorithme DQN sont :

- Toute l'expérience passée est stockée dans une mémoire qu'on appelle *Replay Memory*.
- L'action suivante est déterminée par la sortie maximale du réseau Q.
- La fonction de perte est ici l'erreur quadratique moyenne de la valeur Q

prédite et de la valeur Q cible qu'on a représenté par Q^*. Il s'agit essentiellement d'un problème de régression. Cependant, on ne connaît pas l'objectif car nous avons affaire à un problème d'apprentissage par renforcement. Dans l'équation 3.3 de mise à jour de la valeur Q dérivée de l'équation de *Bellman*, l'expression $r_t + \gamma \max_a Q_k(s_{t+1}, a_t)$ représente l'objectif, on peut dire qu'il prédit sa propre valeur, mais puisque r est la vraie récompense impartiale, le réseau va mettre à jour son gradient en utilisant la rétropropagation pour finalement converger (Choudhary, 2019).

Pour mieux comprendre le fonctionnement de cet algorithme, on énumère les étapes qui sont impliquées pour son déroulement et qui sont tirées de (Choudhary, 2019) :

1. Prétraiter les états qui seront les entrées du réseau profond Q.
2. Sélectionner une action à l'aide de la politique epsilon-greedy. Avec la probabilité ϵ, une action aléatoire a est sélectionnée et avec la probabilité 1-epsilon, une action qui a une valeur Q maximale est sélectionnée, telle que $a = argmax(Q(s, a))$
3. Effectuer cette action dans un état s et passer à un nouvel état s' pour recevoir une récompense. La transition est stockée dans la mémoire (*Replay Memory*) comme $< s, a, r, s' >$.
4. Ensuite, l'algorithme échantillonne des lots aléatoires de transitions à partir de la mémoire et calcule la perte en utilisant l'équation suivante :

$$Loss = (r + \gamma \max_a Q(s', a'; \theta') - Q(s, a; \theta))^2 \tag{3.4}$$

   L'équation 3.4 représente la différence au carré entre la cible Q et la prédiction de Q, θ est le poids du réseau neuronal.

5. Effectuer une descente de gradient par rapport aux paramètres du réseau réel afin de minimiser cette perte.

6. Après chaque itération, l'algorithme copie les poids du réseau réel sur les poids du réseau cible.
7. Répéter ces étapes pour un certain nombre d'épisodes.

On peut conclure cette section en résumant la définition du DQN comme suit : DQN = Q-Learning + Fonction d'approximation + Réseau profond.

3.5.3 MARL

Les deux algorithmes qu'on a présenté précédemment utilisent la notion d'un seul agent, ça veut dire qu'il y a un seul agent centralisé qui effectue toutes les taches. Or dans l'apprentissage automatique et plus précisément dans l'apprentissage par renforcement, il existe des algorithmes qui utilisent plusieurs agents pour leurs fonctionnement, ces algorithmes rentrent dans la catégorie de l'apprentissage par renforcement multi-agents (MARL).

Les algorithmes MARL utilisent plusieurs agents qui travaillent d'une manière soit coopérative ou compétitive, on présente la différence entre les deux catégories dans ce qui suit.

Dans les systèmes multi-agent (MAS) coopératifs, les agents ont un objectif commun, ils se caractérisent par la liberté de concevoir les agents. Les agents peuvent être construits pour apprendre avec une connaissance approfondie du système et ils peuvent s'attendre à des intentions bienveillantes d'autres agents. Contrairement aux MAS coopératives, les agents dans des MAS compétitifs ont des objectifs non alignés, ils sont individualistes et cherchent uniquement à maximiser leurs propres gains (Hoen *et al.*, 2005). Le tableau 3.1 présente les sujets qu'on peut traiter avec les MAS coopératifs et compétitifs :

Tableau 3.1 Sujets dans les systèmes multi-agent (DesJardins, 2005)

MAS coopératifs	MAS compétitifs
Résolution de problèmes distribuées : moins d'autonomie	Rationalité distribuée : vote, enchères
Planification distribuée : modèles de planification et de travail d'équipe	Négociation

Dans notre projet, on a conçu des agents qui travaillent d'une manière coopérative, pour celà, on présente dans ce qui suit des notations et formulations pour un MAS coopératif.

Un MAS est représenté par un tuple $< \mathcal{K}, \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \mathcal{O}, \gamma >$, dans lequel $\mathcal{K}$ est le nombre d'agents, $\mathcal{S}$ est l'espace d'états, $\mathcal{A} = \{\mathcal{A}_1, \cdots, \mathcal{A}_\mathcal{K}\}$ est l'ensemble des actions pour tous les agents, $\mathcal{P}$ est la probabilité de transition entre les états, $\mathcal{R}$ est la fonction de récompense, et $\mathcal{O} = \{\mathcal{O}_1, \cdots, \mathcal{O}_\mathcal{K}\}$ est l'ensemble des observations pour tous les agents, et $\gamma \in [0, 1]$ est le facteur d'actualisation.

Dans un problème coopératif avec $\mathcal{K}$ agents avec une observabilité totale de l'environnement, chaque agent $i \in \mathcal{K} = \{1, ..., n\}$ à l'instant t observe l'état global $s_t \in \mathcal{S}$ et utilise la politique stochastique locale π_i pour agir $a_i^t \in \mathcal{A}$ et reçoit ensuite la récompense r_i^t. Le système passe ensuite à l'état s_{t+1} et récompense chaque agent i par $R^i(s_t, a_t, s_{t+1})$. L'objectif de l'agent i est d'optimiser sa propre récompense à long terme, en trouvant la politique π_i

Si l'environnement est pleinement coopératif, à chaque pas de temps, tous les agents observent une valeur de récompense commune $r_t = r_1^t = \cdots = r_K^t$. Si les agents ne sont pas en mesure d'observer pleinement l'état du système, chaque agent accède uniquement à sa propre observation locale o_d^t.

Similaire à un système à un seul agent, dans une MAS chaque agent est capable

d'apprendre la valeur Q optimale ou la politique stochastique optimale. Cependant, puisque la politique de chaque agent change au fur et à mesure que le déroulement de l'algorithme progresse, par conséquent, l'environnement devient non stationnaire du point de vue de tout agent individuel. Par conséquent, l'équation de Bellman adoptée pour MARL (en supposant l'observabilité totale) est la suivante :

$$Q^*(s, a_i|\pi_{-i}) = \sum_{a_{-i}} \pi_{-i}(a_{-i}, s)\Big[r(s, a_i, a_{-i}) + \gamma \sum_{s'} P(s'|s, a_i, a_{-i}) \max_{a'_i} Q^*_i(s, a'_i)\Big] \tag{3.5}$$

En raison du fait que π_{-i} qui représente la politique de tous les agents sauf l'agent i change avec le temps à mesure que la politique des autres agents change, on ne peut pas obtenir la valeur Q optimale en utilisant l'équation classique de *Bellman*.

Dans la plupart des problèmes de MARL, les agents ne sont pas en mesure d'observer l'état complet du système, qui sont classés comme processus décisionnel de Markov décentralisé partiellement observable (Dec-POMDP). En raison de l'observabilité partielle et de la non stationnarité des observations locales, les Dec-POMDP sont des problèmes encore plus difficiles à résoudre. Une équation similaire à l'équation 3.5 peut également être obtenue pour l'environnement partiellement observable.

Dans le MARL, le bruit et la variance des récompenses augmentent ce qui se traduit par l'instabilité de l'entraînement. La raison est que la récompense d'un agent dépend des actions des autres agents, et la récompense conditionnée par l'action d'un seul agent peut présenter beaucoup plus de bruit et de variabilité que la récompense d'un seul agent. Par conséquent, la formation d'un algorithme de gradient de politique ne serait pas non plus efficace en général.

Enfin, on définit la notation suivante qui est utilisée dans quelques articles en

équilibre de Nash[1]. Une politique conjointe π^* définit un équilibre de Nash si et seulement si :

$$\forall \pi_i \in \Pi_i, \forall s \in \mathcal{S}, J_i(\pi_i^*, \pi_{-i}^*) \geq J_i(\pi_i, \pi_{-i}^*); \forall i \in \{1, \cdots, \mathcal{K}\} \tag{3.6}$$

où $J_i(s)$ est le retour attendu à long terme de l'agent i dans l'état s et Π_i est l'ensemble de toutes les politiques possibles pour l'agent i. Fondamentalement, cela signifie que chaque agent préfère ne pas changer sa politique, s'il veut atteindre la récompense à long terme (Oroojlooy jadid et Hajinezhad, 2019).

3.6 Conclusion

Dans ce chapitre, on a donné un aperçu général de l'apprentissage automatique, ses différents types, des exemples d'applications dans le domaine des réseaux sans-fil tirés de la littérature, ensuite on s'est concentré sur un type d'apprentissage automatique qui est l'apprentissage par renforcement, on a présenté différentes notions de ce type, et quelques algorithmes qu'on a utilisé dans notre projet, on va détaillé ces algorithmes techniquement dans le chapitre suivant.

---

1. C'est un type de solution d'un jeu non coopératif impliquant deux joueurs ou plus dans lequel chaque joueur est censé connaître les stratégies d'équilibre des autres joueurs, et aucun joueur n'a rien à gagner en changeant seulement leur propre stratégie (Wikipédia, 2020d).

CHAPITRE IV

APPARIEMENT INTELLIGENT DES APPAREILS D2D SOUS-JACENTS AUX RÉSEAUX CELLULAIRES ET ACTIVÉS POUR LE CACHE

Dans ce chapitre, nous étudions le problème d'appariement des appareils D2D dans les réseaux cellulaires sous-jacents activés pour le cache. Nous commençons par parcourir la littérature et donner l'état de l'art en relation avec notre problème étudié. Ensuite, nous décrivons la motivation derrière l'étude du problème d'appariement D2D. Par la suite, on présente le modèle de notre système, une formulation (bien détaillée) du problème et de la solution MARL qu'on a adopté. Et nous finissons par évaluer les performances de notre solution en la comparant par rapport à celles du DQN et *Q-Learning*.

4.1 État de l'art

Les communications D2D se sont d'abord révélées avantageuses pour accroître l'efficacité spectrale et la réutilisation de la fréquence, tout en réduisant les délais de communication. Toutefois, ce mode de communication a introduit des interférences supplémentaires en raison de multiples communications simultanées sur les mêmes bandes de fréquences, ce qui est difficile à contrôler et à atténuer. Cela est encore plus évident lorsque les communications D2D se produisent dans des scénarios Inband et/ou Underlay, c'est-à-dire dans le même spectre que le réseau

cellulaire, et/ou en même temps que les communications cellulaires (Asadi *et al.*, 2014).

Dans ce contexte, plusieurs chercheurs ont étudié des problématiques d'allocation des ressources, p. ex., répartition de la puissance et de la fréquence, et établissement de liens pour les communications D2D. Dans (Cheng *et al.*, 2016), les auteurs ont étudié l'impact de l'exigence de la QoS en terme du délai sur le rendement des communications D2D et cellulaires dans un réseau sans fil sous-jacent. Compte tenu de l'approvisionnement en QoS statistique, ils ont dérivé des solutions optimales d'allocation de puissance visant à maximiser le débit du réseau en respectant les contraintes de QoS.

Dans (Zhang *et al.*, 2015), les auteurs ont présenté un réseau sous-jacent basé sur D2D, où les utilisateurs D2D peuvent agir comme relais pour les communications cellulaires, afin de maximiser les taux de transfert réalisables des utilisateurs D2D tout en satisfaisant les exigences minimales de la QOS des utilisateurs cellulaires. Sous contrainte de puissance globale, la stratégie optimale d'allocation de puissance à la (BS) et à l'émetteur D2D est obtenue en mode fermé. De plus, les résultats de la simulation démontrent la supériorité de l'utilisation des dispositifs D2D comme relais complets par rapport aux communications cellulaires classiques.

Les auteurs de (Girmay *et al.*, 2019) ont étudié le problème du canal commun et de l'allocation de puissance, visant à améliorer le débit total des utilisateurs cellulaires et D2D, tout en respectant la QoS minimale des utilisateurs WiFi. Ils ont proposé un algorithme d'optimisation PSO (*particles warm optimization*), qui réduit considérablement l'interférence dans les bandes autorisées et non autorisées et améliore les performances de débit. Ces travaux se sont concentrés sur la dérivation de solutions optimales ou heuristiques qui sont appliquées hors ligne, c.-à-d.,

avec la connaissance à priori de l'état de l'ensemble du système. Cependant, ces méthodes seraient inadéquates pour les déploiements en ligne lorsque l'emplacement des dispositifs et les conditions des canaux changent continuellement. Afin de relever ces défis, des approches d'apprentissage renforcées ont été proposées dans la littérature.

Dans (Luo *et al.*, 2014), les auteurs ont proposé une méthode d'apprentissage Q qui assigne conjointement des canaux et des niveaux de puissance aux appareils D2D afin de maximiser le flux de trafic de bout en bout dans un scénario sous-jacent D2D. Ils ont montré que leur méthode surpasse les indices de référence en termes de capacité moyenne du système. Les auteurs de (Huang *et al.*, 2018) ont employé du Q-learning avec regret logarithmique afin de minimiser la puissance de transmission totale et les associations à la BS. Ensuite, ils ont étendu leur proposition pour distribuer *Q-learning* afin de réduire la complexité. Les résultats obtenus ont démontré la supériorité de leurs approches par rapport aux repères, en termes de réduction de la consommation d'énergie de transmission. En outre, dans (Xu *et al.*, 2018) le problème d'allocation de puissance dans les réseaux sous-jacents D2D a été étudié. Les auteurs ont proposé un contrôle de puissance D2D basé sur une machine d'apprentissage extrême hiérarchique, qui surpasse à la fois le *Q-learning* distribué et l'arbre de décision CART, en termes de débit de communication et d'efficacité énergétique. Enfin, (Moussaid *et al.*, 2018) ont proposé une approche d'apprentissage par renforcement profond pour l'ordonnancement des liaisons D2D dans un réseau de sous-couche cellulaire, visant à maximiser le débit de somme D2D. Les résultats de la simulation ont démontré que la méthode proposée surpasse les approches de base.

Dans les systèmes cellulaires à mémoire de mise en cache, il est clair que les communications D2D contribueraient de façon importante à améliorer les performances du système, en déchargeant le contenu et le trafic du réseau cellulaire,

en réduisant les délais de récupération et de livraison du contenu. et améliorer l'efficacité énergétique du réseau cellulaire. Ces avantages ont été démontrés dans plusieurs travaux récents.

Dans (Gregori *et al.*, 2016), les auteurs ont proposé de mettre en cache du contenu soit dans la BS cellulaire, soit dans les dispositifs D2D. Ensuite, ils ont étudié la conception conjointe de la politique de mise en cache et de livraison du contenu, compte tenu de leur connaissance préalable des demandes des utilisateurs. Les auteurs ont décrit dans (Yang *et al.*, 2016b) les communications cognitives D2D sous-jacentes aux transmissions cellulaires afin de mettre en cache les fichiers entre les appareils de manière transparente. Ensuite, ils ont évalué le délai et la longueur de la file d'attente sur les appareils BS et D2D. En outre, les auteurs ont proposé dans (Li *et al.*, 2018b) un algorithme de mise en cache qui minimise le délai moyen de livraison de contenu dans un réseau cellulaire assisté par D2D. Ils ont démontré la supériorité de leur algorithme glouton proposé sur la politique de mise en cache naïve basée sur la popularité. Dans (Wang *et al.*, 2017), les auteurs ont exploité la coopération entre la mise en cache BS et la mise en cache D2D dans le but de maximiser la probabilité de transmission réussie, qui mesure la proportion d'utilisateurs satisfaisant leurs exigences de QoS de retard. Ils ont montré qu'en utilisant l'algorithme de descente de coordonnées de bloc proposé, ils peuvent obtenir des gains de performances significatifs par rapport aux méthodes de mise en cache conventionnelles.

En revanche, les auteurs de (Li *et al.*, 2018b) ont utilisé un algorithme d'apprentissage du comportement des utilisateurs qui prédit les demandes des utilisateurs et estime la popularité des fichiers. Ainsi, ces informations sont utilisées pour ajuster la politique de mise en cache entre les appareils D2D, dans le but de minimiser les délais de transmission pour la livraison des fichiers. A travers des simulations, il est montré que l'approche proposée surpasse à la fois la mise en cache naïve et

la mise en cache probabiliste. Dans (Jaafar *et al.*, 2018), les auteurs ont étudié le problème de placement de cache conjoint et d'allocation des ressources dans les réseaux hétérogènes assistés par D2D. Ils ont proposé deux solutions heuristiques sous-optimales qui surpassent les systèmes qui ne prennent pas en charge D2D, en termes de délais moyens de livraison de contenu. Dans (Jaafar *et al.*, 2019), les auteurs ont étudié le placement de contenu visant à minimiser le délai moyen de livraison de contenu dans le contexte d'un Hetnet assisté par D2D. Pour une bibliothèque de contenu unique, la solution optimale est obtenue, puis utilisée pour concevoir une heuristique peu complexe pour une bibliothèque de contenu plus grande. Grâce à des simulations, ils ont montré les avantages de la mise en cache de fichiers populaires dans des appareils D2D avec les meilleurs liens vers d'autres appareils D2D, au sein du système cellulaire.

Les auteurs de (Yin *et al.*, 2018) ont étudié la mise en cache et la livraison de contenu dans un réseau D2D. Ils ont exploité les réseaux à état d'écho pour la popularité du contenu et la prédiction de la mobilité des utilisateurs, tandis qu'un algorithme de réseau Q profond DQN pour optimiser la livraison du contenu en ce qui concerne les retards et les contraintes énergétiques. Dans (Jiang *et al.*, 2018), (Jiang *et al.*, 2019), les auteurs ont formulé un problème de mise en cache D2D d'apprentissage par renforcement multi-agents (MARL), visant à améliorer la latence moyenne de livraison de contenu et le taux de succès du cache, sans connaissance préalable de la popularité du contenu. Pour le résoudre, ils ont proposé le Q-learning pour les dispositifs d'apprentissage indépendants et les dispositifs d'apprentissage par action conjointe. Grâce à des simulations, ils ont constaté que les algorithmes d'apprentissage par action conjointe obtiennent de meilleures performances que les algorithmes d'apprentissage individuels et d'autres approches de base.

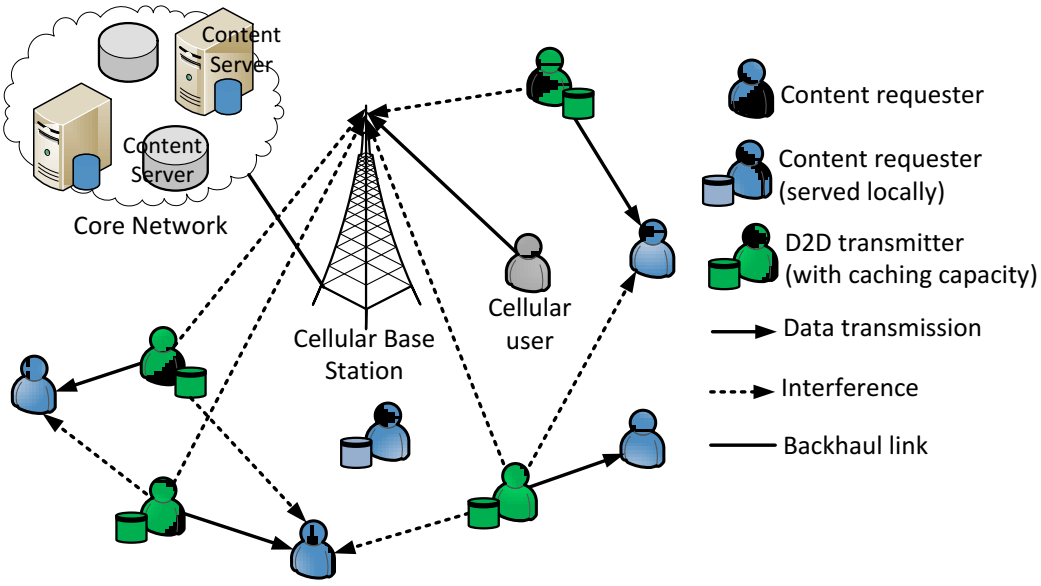

Figure 4.1 Modèle du système.

4.2 Modèle du système

4.2.1 Modèle du réseau

Nous supposons une cellule de couverture d'une BS, où N utilisateurs cellulaires communiquent avec la BS en utilisant des ressources de fréquences orthogonales, coexistant avec un certain nombre d'utilisateurs D2D. Nous supposons que les utilisateurs D2D peuvent être regroupés de telle sorte que chaque cluster utilise l'une des ressources de fréquence disponibles. Par souci de simplicité, nous supposons que les utilisateurs D2D sont déjà regroupés et nous nous concentrons sur les communications se produisant dans un seul cluster, comme illustré sur la figure 4.1

Soit D le nombre de nœuds D2D au sein de ce cluster, indexé à partir de l'ensemble $\mathcal{D} = \{d_1, \ldots, d_D\}$. Aussi, nous supposons que le temps est également divisé

en (TS), où t désigne un intervalle de temps (TS).

4.2.2 Modèle de mise en cache et de déchargement des données

Dans notre réseau, nous supposons que les fichiers F (ou segments) de même taille appartiennent à une bibliothèque $\mathcal{F} \equiv \{1, \ldots, F\}$ du cœur de réseau (Yang *et al.*, 2016a).

Sans perte de généralité, la même taille de fichiers peut être obtenue en divisant de gros fichiers en segments de taille égale. Nous supposons que tout utilisateur D2D d a une capacité de mise en cache $C_d \ll F$. Nous supposons qu'initialement l'utilisateur D2D d met en cache les fichiers C_d de manière aléatoire (ou selon une distribution probabiliste) sélectionnés dans la bibliothèque disponible. On désigne l'état de la mémoire cache d'un utilisateur d par le vecteur $\mathbf{c}_d = [c_{d,f}]_{1\times F}$, tel que $c_{d,f} = 1$ si l'utilisateur d met en cache le fichier f, sinon $c_{d,f} = 0$, et $\sum_{f=1}^{F} c_{d,f} \leq C_d$, $\forall d \in \mathcal{D}$.

A n'importe quel intervalle de temps, l'utilisateur D2D d peut demander un fichier avec une probabilité $p_d \in [0, 1]$, appelé «demandeur». Ainsi, soit $m_{d,f}$ la variable binaire indiquant que l'utilisateur D2D d a demandé le fichier f au début d'un intervalle de temps ou non. De plus, le fichier f est sélectionné par l'utilisateur D2D d selon la distribution de probabilité uniforme, où $q_f = \frac{1}{F}$ désignant la probabilité de sélectionner le fichier f.

Lorsque le fichier f est demandé par l'utilisateur D2D d, ce dernier commence par vérifier sa propre mémoire cache. Si le contenu est disponible localement, il est obtenu directement sans délai. Sinon, il diffuse sa demande à ses nœuds voisins, qui sont disponibles pour aider. Un voisin ayant le fichier demandé dans sa mémoire peut potentiellement établir une communication avec l'utilisateur d et le servir. En cas d'absence de réponse des voisins, la demande sera servie par la BS en tant

que communication cellulaire.

4.2.3 Modèles de canal et de communication

Nous considérons un réseau cellulaire sous-jacent D2D, où les communications D2D d'utilisateurs groupés se produisent à la même fréquence de liaison montante qu'un utilisateur cellulaire. L'avantage d'utiliser la bande de fréquences montante réside dans la nature sous-chargée des ressources montantes et dans la force de la BS pour atténuer les interférences entrantes (Lin *et al.*, 2014). Selon le modèle de déchargement de données, les communications D2D sont impliquées dans le réseau lorsque les fichiers demandés ne sont pas dans les mémoires cache locales, mais sont disponibles sur un ou plusieurs appareils D2D voisins. Soit $\psi : s \mapsto d$ la fonction qui associe l'utilisateur s à l'utilisateur d, c'est-à-dire que l'utilisateur s est disponible pour aider, possède le fichier f demandé par l'utilisateur d et est un «voisin» de d. A noter qu'un voisin de d peut être défini en fonction de la distance ou de la qualité de la liaison de communication entre les utilisateurs s et d. Par conséquent, le signal reçu au niveau du demandeur D2D d, lorsqu'il est servi par des émetteurs D2D, peut être exprimé par

$$\begin{aligned} y_d &= \sqrt{P_s} l_{sd}^{-\frac{\beta}{2}} h_{sd} x_d + \sqrt{P_0} l_{0d}^{-\frac{\beta}{2}} h_{0d} x_0 + n_d \\ &+ \sum_{i \in \mathcal{S},\, i \neq s} \delta_i \sqrt{P_i} l_{id}^{-\frac{\beta}{2}} h_{id} x_{\psi(i)},\ \forall d \in \mathcal{D}', \end{aligned} \tag{4.1}$$

où P_s (resp. P_0 et P_i) est la puissance d'émission de l'utilisateur s (resp. utilisateur cellulaire et utilisateur D2D i), l_{sd} (resp. l_{0d} et l_{id}) et h_{sd} (resp. h_{0d} et h_{id}) sont la distance et le coefficient de canal à petite échelle entre les utilisateurs d'une seule antenne s (respectivement utilisateur cellulaire et utilisateur D2D i) et d, respectivement, β est l'exposant de perte de trajet, x_d (resp. x_0 and $x_{\psi(i)}$) est le signal destiné à l'utilisateur d (resp. à BS et l'utilisateur i) avec une énergie unitaire, n_d est le bruit blanc Gaussien additif de puissance σ^2, $\mathcal{S} = \{1, \ldots, S\}$ est

l'ensemble des utilisateurs disponibles pour aider, et $\delta_i \in \{0,1\}$ est une variable binaire indiquant si l'utilisateur i assiste activement un utilisateur (différent) ou non. Enfin, l'ensemble $\mathcal{D}' \subset \mathcal{D}$ inclut les demandeurs qui n'ont pas mis en cache leur fichier demandé, mais peuvent l'obtenir d'un utilisateur voisin, c'est-à-dire,

$$\mathcal{D}' = \left\{ d \mid \exists s \in \mathcal{D}, \sum_{f=1}^{F} m_{s,f} = 0, c_{d,f_0} = 0, m_{d,f_0} = 1, c_{s,f_0} = 1 \right\}, \tag{4.2}$$

où f_0 est le fichier demandé par l'utilisateur d. Le premier terme dans (4.1) est le signal souhaité, tandis que le deuxième et le dernier terme sont les composants d'interférence de l'utilisateur cellulaire et d'autres émetteurs D2D, respectivement.

De même, le signal reçu à la BS peut être écrit comme

$$y_0 = \sqrt{P_0} l_{00}^{-\frac{\beta}{2}} h_{00} x_0 + \sum_{i \in \mathcal{S}} \delta_i \sqrt{P_i} l_{i0}^{-\frac{\beta}{2}} h_{i0} x_{\psi(i)} + n_0, \tag{4.3}$$

où l_{00}, h_{00}, h_{i0} et n_0 sont définis comme dans (4.1). De (4.1) - (4.3), le rapport signal sur interférence plus bruit (SINR) correspondant peut être donné par

$$\gamma_d = \frac{P_s l_{sd}^{-\beta} |h_{sd}|^2}{P_0 l_{0d}^{-\beta} |h_{0d}|^2 + \sum_{i \in \mathcal{S}, i \neq s} \delta_i P_i l_{id}^{-\beta} |h_{id}|^2 + \sigma^2} \; \forall d \in \mathcal{D}' \tag{4.4}$$

et

$$\gamma_0 = \frac{P_0 l_{00}^{-\beta} |h_{00}|^2}{\underbrace{\sum_{i \in \mathcal{S}} \delta_i P_i l_{i0}^{-\beta} |h_{i0}|^2 + \sigma^2}_{=I_0}}, \tag{4.5}$$

où γ_d et γ_0 sont respectivement le SINR reçu à l'utilisateur $d \in \mathcal{D}'$ et la BS. Nous supposons que les transmissions de l'utilisateur cellulaire sont continues dans le temps et que par conséquent l'agrégation des communications D2D doit respecter la puissance d'interférence maximale tolérée au niveau de la BS, notée Q, c'est-à-dire $I_0 \leq Q$ (Jaafar *et al.*, 2016). Étant donné que la bande passante de la porteuse est W, la performance de débit somme du réseau D2D est exprimée par

$$R_{\text{D2D}} = \sum_{d \in \mathcal{D}'} R_d = \sum_{d \in \mathcal{D}'} W \log_2 (1 + \gamma_d), \tag{4.6}$$

où $R_d = W \log_2 (1 + \gamma_d)$ est le débit de données atteint au demandeur d.

4.3 Formulation du problème

Pour le scénario représenté sur la figure 4.1, nous cherchons à obtenir la configuration d'appariement entre les demandeurs D2D et les émetteurs D2D disponibles, de sorte que R_{D2D} soit maximisé. Étant donné que les utilisateurs de D2D sont des dispositifs à antenne unique, nous supposons que chaque émetteur disponible peut servir au plus un demandeur à la fois, et qu'un demandeur peut recevoir d'un émetteur à la fois. Lorsqu'un émetteur est appairé à un récepteur, une certaine exigence de QoS doit être garantie, c'est-à-dire $\gamma_d \geq \gamma_{\text{th}}$, où γ_{th} est un arbitraire seuil de décodage SINR. Soit $\mathbf{\Lambda} = [\lambda_{ij}]_{D\times D}$ la matrice binaire des décisions d'appariement des utilisateurs, telles que $\lambda_{ij} = 1$ si la transmission de l'utilisateur i à l'utilisateur j est activée, et $\lambda_{ij} = 0$ sinon, $\forall (i,j) \in \mathcal{D}^2$. Par conséquent, le problème peut être formulé comme suit

$$\max_{\mathbf{\Lambda}} \quad R_{\text{D2D}} \tag{P1}$$

$$\text{s.t.} \quad \gamma_d \geq \gamma_{\text{th}} - \left(1 - \sum_{i=1}^{D} \lambda_{id}\right) \cdot M, \ \forall d \in \mathcal{D} \tag{P1.a}$$

$$I_0 \leq Q \tag{P1.b}$$

$$\delta_i = \sum_{d=1}^{D} \lambda_{id} \leq 1, \forall i \in \mathcal{D} \tag{P1.c}$$

$$\sum_{i=1}^{D} \lambda_{id} \leq 1, \forall d \in \mathcal{D} \tag{P1.d}$$

$$\lambda_{ij} \in \{0,1\}, \ \forall (i,j) \in \mathcal{D}^2, \tag{P1.e}$$

où $M \gg 1$ permet la validité de la première contrainte. (P1.a) garantit que lorsqu'un demandeur d est servi, un débit minimal est garanti, et (P1.b) garantit le respect du seuil d'interférence au niveau de la BS. D'autre part, (P1.c) - (P1.d) valide le fait qu'un émetteur D2D ne dessert qu'un seul demandeur et qu'un de-

mandeur est servi uniquement par un utilisateur. Enfin, (P1.e) met l'accent sur la nature binaire des variables d'optimisation.

Ce problème peut être vu comme un problème de correspondance un-à-un, qui peut être résolu de manière centralisée si on a la connaissance de toutes les informations des utilisateurs, c'est-à-dire de tous les canaux (entre tous les utilisateurs et vers la BS), l'état de la mémoire cache et les demandes, sont disponible sur un contrôleur central, par exemple la BS (Song *et al.*, 2018). Bien que la solution optimale puisse être obtenue avec ces hypothèses, une telle méthode n'est pas pratique car elle générerait un surcharge d'informations important entre la BS et les utilisateurs, outre le fait que les informations collectées peuvent être rapidement obsolètes. Par conséquent, nous promouvons dans ce mémoire une approche distribuée intelligente, où aucun échange d'informations n'est requis, tandis que les décisions d'appariement sont gérées localement (au niveau des demandeurs), évitant ainsi les transmissions de surcharge vers et depuis la BS.

4.4 Solution proposée : MARL

La résolution du problème (P1) est difficile en raison de sa dureté NP issue de sa nature combinatoire, en plus du manque d'informations globales pour décider de la stratégie d'appariement des utilisateurs D2D. En optant pour une méthode distribuée pour résoudre (P1), nous exploitons un agent intelligent au niveau de chaque demandeur D2D afin de se coupler avec l'émetteur D2D le plus adéquat. Pour ce faire, l'apprentissage par renforcement est adopté, en raison de son efficacité avérée à résoudre de tels problèmes. Plus précisément, nous commençons dans cette section en présentant une description de base de l'apprentissage par renforcement et de l'algorithme QMIX. Ensuite, nous formulons le cadre MARL distribué, qui sera utilisé pour résoudre le problème d'appariement des utilisateurs

D2D.

4.4.1 Système multi-agents entièrement coopératif

Avant de présenter notre approche basée sur QMIX, nous fournissons une introduction de fond au système multi-agents entièrement coopératif.

Un système multi-agents entièrement coopératif peut être décrit comme un processus de décision de Markov décentralisé partiellement observable (Dec-POMDP) (Oliehoek et Amato, 2016), consistant en un tuple $G = \langle \mathcal{X}, \mathcal{A}, P, r, Z, n, O, \eta \rangle$, où $\mathcal{X}$ est l'ensemble (fini) des états, $\mathcal{A}$ est l'ensemble des actions conjointes, P est la fonction de probabilité de transition, r est la fonction de récompense immédiate, Z est l'ensemble des observations conjointes, n est le nombre d'agents, O est la fonction de probabilité d'observation et η est le facteur de remise. $x \in \mathcal{X}$ décrit le véritable état de l'environnement. A chaque pas de temps, chaque agent $d \in \mathcal{K} = \{1, \ldots, n\}$ sélectionne une action $a^d \in \mathcal{A}^d$, formant ainsi une action commune $\mathbf{a} \in \mathcal{A} = \mathcal{A}^1 \times \ldots \times \mathcal{A}^n$. Cela provoque une transition sur l'environnement selon la fonction de transition d'état $P(x'|x, \mathbf{a}) : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to [0, 1]$. Ensuite, tous les agents partagent la même récompense immédiate selon la fonction $r(x, \mathbf{a}) : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$.

Dans un environnement partiellement observable, chaque agent d dessine des observations individuelles $z \in Z$ selon la fonction $O(x, d) : \mathcal{X} \times \mathcal{K} \to Z$. L'agent d a un historique d'observation des actions $\tau^d \in T \equiv (Z, \mathcal{A})$, sur lequel il conditionne une politique stochastique $\pi^d(a^d, \tau^d) : T \times \mathcal{A} \to [0, 1]$. La politique conjointe de tous les agents π a une fonction valeur-action conjointe définie comme

$$Q^\pi(x_t, \mathbf{a}_t) = \mathbb{E}_{x_{t+1:\infty}, \mathbf{a}_{t+1:\infty}}[R_t | x_t, \mathbf{a}_t] \tag{4.8}$$

où $\mathbb{E}(\cdot)$ est la fonction d'espérance et $R_t = \sum_{i=0}^{\infty} \eta^i r_{t+i}$ est le *rendement réduit*.

On note que dans Dec-POMDP, la formation est centralisée, mais la prise de décision est décentralisée. En effet, l'algorithme d'apprentissage a accès à tout l'histoire locale des observations d'actions $\boldsymbol{\tau} = [\tau^1, \ldots, \tau^n]$ et à l'état global x, cependant, la politique apprise de chaque agent ne peut conditionner que son propre historique action-observation τ^d.

4.4.2 L'approche *QMIX*

Comme pour Dec-POMDP, l'approche QMIX est une méthode hybride qui unifie deux méthodes extrêmes, à savoir l'apprentissage Q indépendant (IQL) (Tan, 1993) et les réseaux de décomposition de valeur (VDN) (Sunehag *et al.*, 2018). En effet, IQL traite un problème multi-agents en le divisant en un ensemble de problèmes mono-agents simultanés partageant le même environnement. IQL ne traite pas de la non stationnarité de l'environnement en raison de l'évolution des politiques des agents. Par conséquent, il ne garantit pas la convergence comme Q-learning, alors que VDN apprend une fonction action-valeur conjointe, dénotée $Q_{\text{tot}}(\boldsymbol{\tau}, \mathbf{a})$. Q_{tot} est défini comme la somme des fonctions de valeurs individuelles, dénoté $Q_d(\tau^d, a^d)$ pour l'agent d, qui est conditionné uniquement sur l'historique action-observation individuel.

L'approche clé de QMIX est d'utiliser la factorisation partielle du VDN pour dériver des politiques décentralisées (Rashid *et al.*, 2018). Il suffit de s'assurer qu'un *argmax* global, exécuté sur Q_{tot}, donne le même résultat qu'un ensemble d'opérations *argmax* exécutées sur chaque fonction Q_d, ce qui signifie,

$$\arg\max_{\mathbf{a}} Q_{\text{tot}}(\boldsymbol{\tau}, \mathbf{a}) = \begin{pmatrix} \arg\max_{a^1} Q_1(\tau^1, a^1) \\ \vdots \\ \arg\max_{a^D} Q_D(\tau^D, a^D) \end{pmatrix}. \tag{4.9}$$

Par conséquent, une contrainte de monotonie est appliquée sur la relation entre

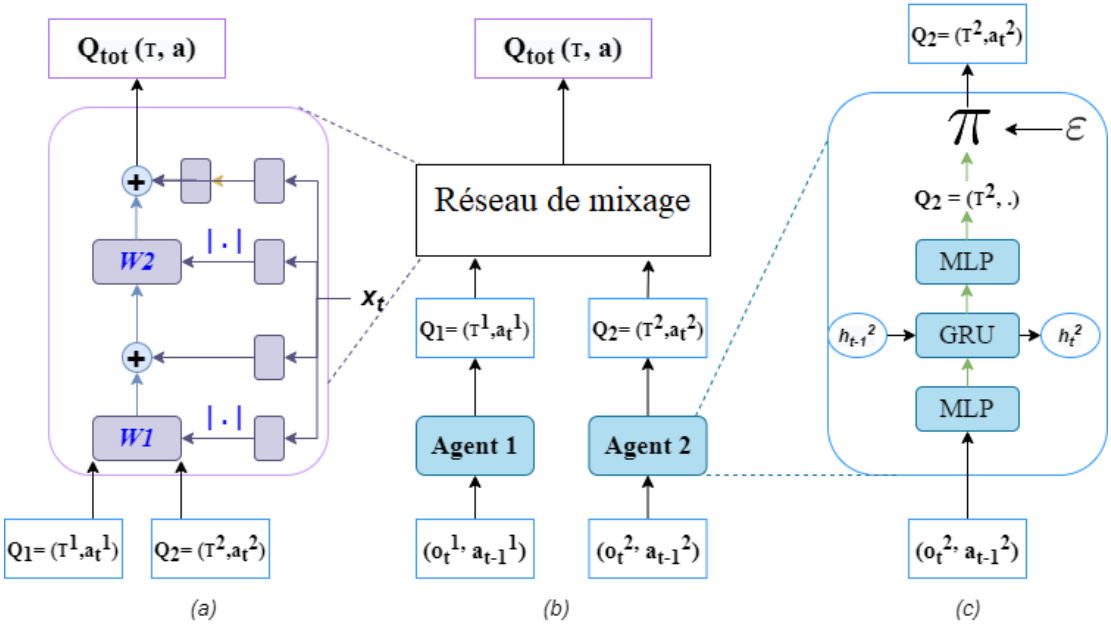

Figure 4.2 Structure QMIX pour 2 agents : (a) Réseau de mixage ; (b) Architecture QMIX globale ; (c) Réseau d'agent. (Rashid *et al.*, 2018)

Q_{tot} et chaque fonction Q_d comme suit :

$$\frac{\partial Q_{\text{tot}}}{\partial Q_d} \geq 0, \forall d = 1, \ldots, n. \tag{4.10}$$

Comme illustré sur la figure 4.2, la structure QMIX est composée de réseaux d'agents, d'un réseau de mixage et d'un ensemble d'hyper-réseaux. Le réseau d'agents représente chaque Q_d, tandis que le réseau de mixage les combine en Q_{tot} d'une manière complexe et non linéaire, assurant ainsi la cohérence entre les politiques centralisées et décentralisées. Simultanément, il applique la contrainte (4.10) en restreignant le réseau de mélange à des poids positifs. Ces poids sont produits par des hyper réseaux séparés, c'est-à-dire d'autres réseaux de neurones.

Le QMIX est formé de bout en bout en minimisant la fonction de perte $\mathcal{L}(\boldsymbol{\theta})$,

exprimée par

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{b} \left[(y_i^{\text{tot}} - Q_{\text{tot}}(\boldsymbol{\tau}, \mathbf{a}, x; \boldsymbol{\theta}))^2 \right], \tag{4.11}$$

où b est la taille du lot des transitions échantillonnées à partir du tampon de relecture, $y^{\text{tot}} = r + \gamma \max_{\mathbf{a}'} Q_{\text{tot}}(\boldsymbol{\tau}', \mathbf{a}', x'; \boldsymbol{\theta}^-)$ et $\boldsymbol{\theta}^-$ sont les paramètres du réseau cible. Puisque (4.9) tient, la maximisation de Q_{tot} peut être effectuée dans une durée linéaire avec le nombre d'agents (Rashid *et al.*, 2018).

4.4.3 Appariement des appareils D2D basé sur *QMIX*

Dans cette sous-section, nous présentons la méthode d'appariement D2D basée sur *QMIX* proposée pour résoudre le problème (P1) de manière distribuée. Le problème d'appariement D2D multi-agents est formulé comme un Dec-POMDP, où chaque demandeur D2D peut être vu comme un agent ayant des connaissances spécifiques et des actions stratégiques associées à sa propre perception de l'environnement.

Notre problème sera considéré comme un jeu, où les demandeurs $D' = |\mathcal{D}'|$ D2D évaluent leur environnement visible et décident avec quel utilisateur s'associer pour obtenir les fichiers demandés. Ce jeu de Markov peut être défini par un ensemble d'espaces d'état $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_{D'}$, où l'espace d'état $\mathcal{X}_d$ correspond à l'environnement vu par le demandeur D2D d. On note x_d un état dans $\mathcal{X}_d$, défini par

$$x_d = \{\mathbf{m}_d, \mathbf{c}_d, \mathcal{S}_d\}, \tag{4.12}$$

où $\mathbf{m}_d = [m_{d,1}, \ldots, m_{d,F}]$ indique quel fichier est demandé, et $\mathcal{S}_d$ est l'ensemble des émetteurs potentiels vers le demandeur D2D d, défini comme

$$\mathcal{S}_d = \left\{ s \in \mathcal{S} \mid \sum_{f=1}^{F} m_{s,f} = 0,\ \exists f_0,\ c_{s,f_0} = m_{d,f_0} = 1,\ \bar{\gamma}_{sd} > \gamma_{\text{th}} \right\}, \tag{4.13}$$

$\bar{\gamma}_{sd} = P_s l_{sd}^{-\beta} |h_{sd}|^2$. Nous supposons ici que l'utilisateur D2D d ne peut communiquer avec l'utilisateur s que lorsque $\bar{\gamma}_{sd}$ est strictement supérieur au seuil de décodage γ_{th}, cette condition stricte garantit la réservation d'une marge au cas où cette communication serait perturbée par d'autres transmissions.

L'ensemble des espaces d'action des agents peut être défini comme $\mathcal{A} = \mathcal{A}_1 \times \ldots \times \mathcal{A}_{D'}$, où l'espace d'action de l'agent d est noté $\mathcal{A}_d$. A chaque instant t, l'agent D2D d utilise une politique $\pi : \mathcal{X}_d \mapsto \mathcal{A}_d$ pour déterminer l'action $a^d \in \mathcal{A}_d$. Cette action correspond à la décision d'appariement du demandeur D2D d avec un émetteur D2D s. Il peut être défini par le vecteur binaire

$$a_d = [\lambda_{d1}, \ldots, \lambda_{dD}], \tag{4.14}$$

où a_d est la d^{th} ligne dans $\mathbf{\Lambda}$ Suite aux actions entreprises par tous les agents D, une récompense commune est calculée par un contrôleur central et diffusée à tous les agents. La fonction de récompense est donnée par R_{D2D} in (4.6), car nous supposons que la récompense de tout demandeur D2D servi par lui-même ou par la BS est nulle. Enfin, cette récompense immédiate servira à ajuster les paramètres *QMIX* de tous les agents.

4.5 Résultats de la simulation

Dans cette section, nous décrivons la configuration de la simulation, puis nous évaluons les performances de notre solution MARL pour différents scénarios et paramètres.

4.5.1 Configuration de la simulation

Des simulations ont été menées sur le framework PyMARL pour l'apprentissage par renforcement multi-agents (Samvelyan *et al.*, 2019), en utilisant un serveur

équipé du *GPU AMD Radeon VII* (16 Go de mémoire RAM).

Les architectures de tous les réseaux d'agents sont unifiées aux réseaux Q profonds avec une couche réseau neuronale récurrente composée d'une unité récurrente synchronisée (GRU) avec deux couches entièrement connectées, séparées par une couche cachée en 64 dimensions. Le réseau de mixage comprend une seule couche cachée de 32 unités qui intègre la fonction d'activation ELU. L'hyperréseau produisant la polarisation finale du réseau de mixage a une couche cachée unique de 32 unités avec une fonction d'activation ReLU.

L'hyperréseau produisant le biais final du réseau de mixage a une couche cachée unique de 32 unités avec une fonction d'activation ReLU. L'exploration est effectuée pendant le déroulement de l'algorithme en utilisant ϵ *-greedy* d'action indépendante par chaque agent d sur sa propre valeur-fonction Q_d. Tout au long de la formation, nous fixons ϵ linéairement de $1,0$ à $0,05$ sur 50 000 épisodes, puis le maintenons constant pour les épisodes suivants. Pour nos simulations, nous supposons un réseau D2D sous-jacent à un réseau cellulaire, comme présenté dans la Fig. 4.1. Sauf indication contraire, nous supposons que le nombre d'appareils D2D est de $D = 15$, répartis aléatoirement dans une zone de 1 km ^{2} autour d'une BS cellulaire et de son utilisateur cellulaire associé.

Chaque périphérique D2D a une capacité de mise en cache de fichiers de 5, tandis que la bibliothèque entière est de $F = 10$ de fichiers, disponibles dans le réseau central. Par souci de simplicité, nous suivons une politique de mise en cache aléatoire sur les appareils D2D. Enfin, les paramètres restants sont présentés dans le tableau 4.1 (Li et Guo, 2020; Zhao *et al.*, 2018; Rashid *et al.*, 2018).

Tableau 4.1 Les paramètres de la simulation

Parameter	Symbol	Value
Nombre des utilisateurs D2D	D	15, 30, 45
Nombre de fichiers	F	10, 50, 100
Puissance de transmission de l'utilisateur cellulaire	P_0	13 dBm
Puissance de transmission de l'utilisateur D2D	P_s	13 dBm
Exposant d'affaiblissement	β	2.7
Largeur de bande de la porteuse	W	20 MHz
Puissance d'interférence maximale tolérée par la BS	Q	5 dB
Seuil de décodage SINR	γ_{th}	0
Probabilité de demander un fichier ou non	p_d	0.5
Facteur d'actualisation	η	0.99
La taille du buffer	–	32
Le taux d'apprentissage	ξ	5×10^{-4}

4.5.2 Évaluation des performances

Sur la figure 4.3, nous comparons les performances de notre solution basée sur QMIX proposée à des schémas de référence, à savoir le *Q-Learning* distribué (Nie *et al.*, 2016) et le DQN (Moussaid *et al.*, 2018) , en termes de débit de données total. Dans nos simulations, nous supposons que le nombre d'agents, c'est-à-dire de demandeurs D2D, est de $D' = 5$, tandis que les 10 autres appareils D2D peuvent servir comme émetteurs de contenu.

Comme on peut le voir, les performances de toutes les solutions algorithmiques augmentent avec le nombre d'épisodes jusqu'à la convergence vers la politique optimale, qui réalise les meilleurs débits de données totaux. Bien que *Q-learning* et DQN convergent environ après 200 000 épisodes, *QMIX* nécessite plus de for-

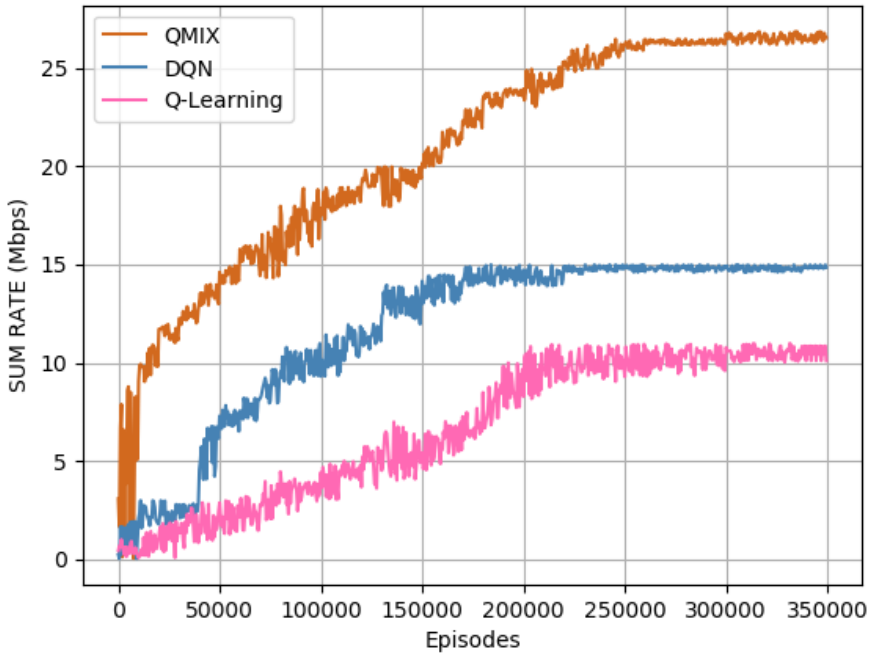

Figure 4.3 Débit total en fonction des nombre d'épisodes (différents schémas)

mation pour converger (environ 250 000 épisodes), cependant, ce dernier surpasse les deux schémas de référence. En effet, *QMIX* permet l'apprentissage d'une riche fonction commune action-valeur, qui admet une décomposition traitable des fonctions action-valeur par agent. Ceci est réalisé en imposant la contrainte de monotonie sur le réseau de mixage. Enfin, DQN est supérieur au *Q-learning* en raison du réseau neuronal profond qui permet une meilleure sélection de politique.

La figure 4.4 représente la somme des débit de données en fonction du seuil de décodage γ_{th} pour les trois schémas étudiés. Au fur et à mesure que γ_{th} devient strict, le nombre d'émetteurs voisins pour chaque demandeur D2D diminue, poussant davantage de demandeurs à être servis par la BS, et par conséquent, les performances de la somme de débit de données des utilisateurs D2D se dégradent.

Dans les résultats restants, nous étudions l'impact de plusieurs paramètres sur les

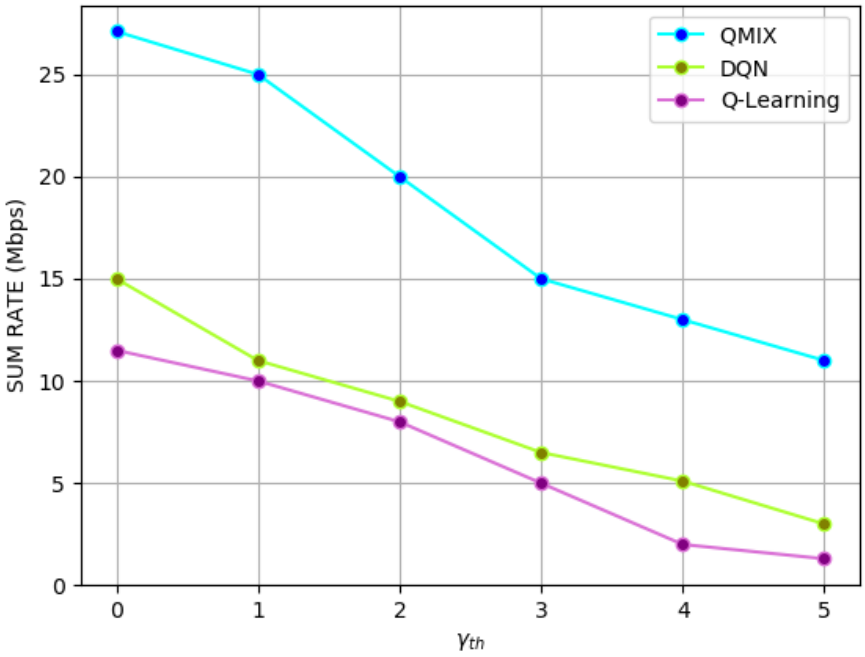

Figure 4.4 Débit total versus γ_{th} (differents schémas).

performances de débit de données de la solution QMIX proposée.

Dans la Fig. 4.5, nous illustrons l'impact de la taille du réseau D2D sur la somme des performances de débit de données de la solution *QMIX* proposée. En effet, on fait varier $D \in \{15, 30, 45\}$. Nous remarquons que le débit de données global s'améliore avec le nombre d'appareils D2D D. Cela est dû à la probabilité accrue qu'un fichier demandé soit présent dans un nœud D2D voisin. De plus, à mesure que D augmente, la convergence de la solution basée sur *QMIX* devient plus lente. Par exemple, il est d'environ 250 000 épisodes pour $D = 15$ mais passe à 320 000 épisodes pour $D = 45$. Cela s'explique par une plus grande complexité du système, due principalement à un espace d'action conjoint plus élevé.

Dans la figure 4.6, nous étudions l'impact de la taille de la bibliothèque F sur les performances du débit de données total, étant donné $D = 15$ appareils. À mesure que F augmente, le débit de données global se dégrade. Ceci est attendu car une

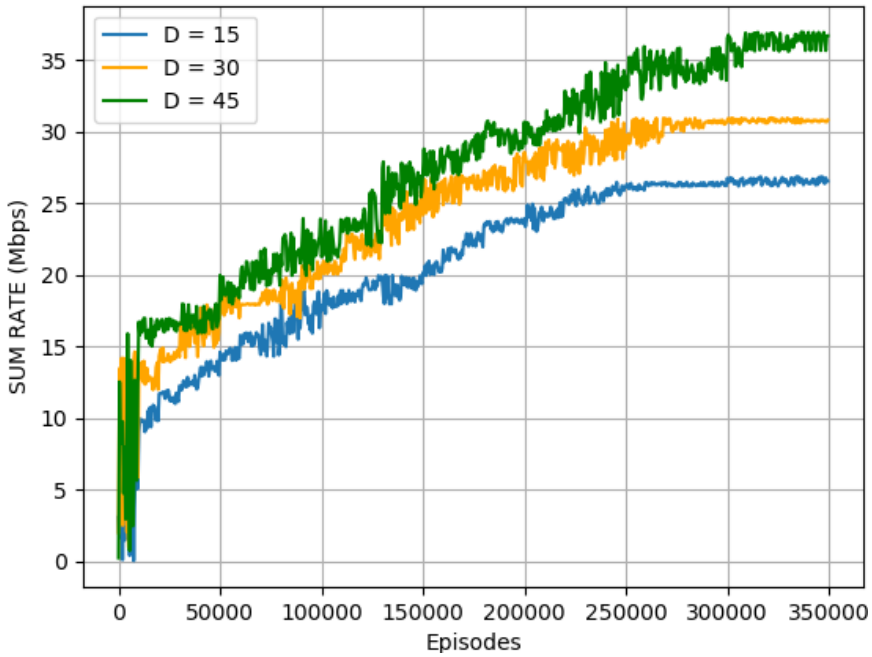

Figure 4.5 Débit total par rapport au nombre d'épisodes (nombre différent d'utilisateurs D2D).

bibliothèque de grande taille implique que la proportion de fichiers mis en cache dans les périphériques D2D est très petite, ainsi plusieurs demandes de fichiers sont redirigées vers la BS, ce qui réduit la récompense globale D2D. De plus, la convergence du *QMIX* devient plus lente avec F en raison de la dimension plus élevée de l'espace d'états.

La figure 4.7 évalue l'impact de la capacité de mise en cache des appareils D2D, C_d, sur les performances du système, étant donné $D = 30$ et $F = 100$. $C_d = 5$ présente la pire performance de débit de données. En augmentant à $C_d = 10$, un gain significatif est obtenu (environ 30 %). En effet, dans cette configuration réseau, plus de communications D2D sont favorisées, au détriment de la redirection des requêtes vers la BS, en raison de la disponibilité de plus de fichiers dans les appareils D2D. Cependant, lorsque $C_d = 15$, le gain de débit devient faiblepar

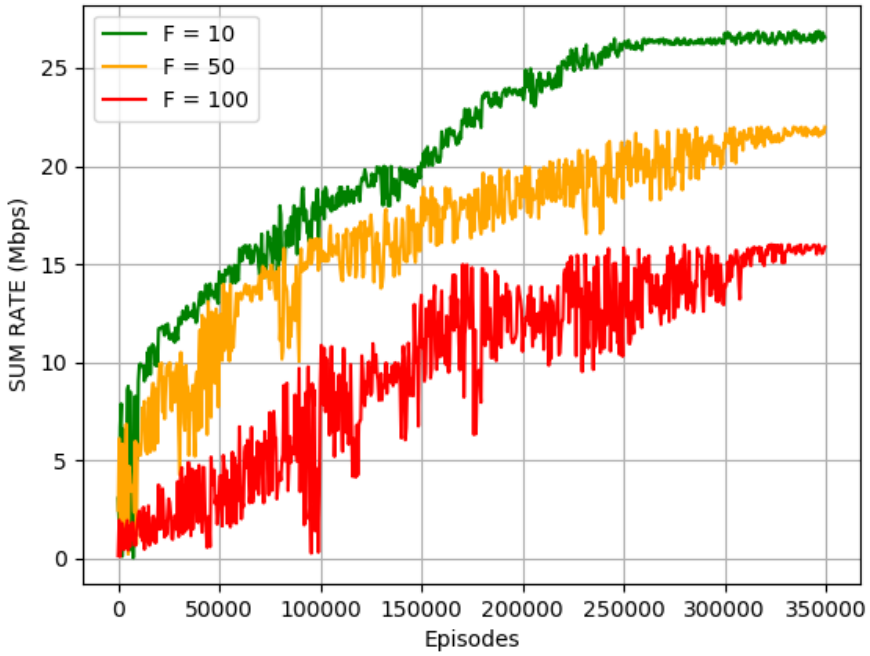

Figure 4.6 Débit total par rapport au nombre d'épisodes (différentes tailles de la bibliothèque).

rapport à $C_d = 10$ (moins de 10 %). En fait, une capacité de cache plus grande favorise la satisfaction des requêtes localement, d'où la nécessité d'établir des communications D2D simultanées moins nombreuses mais plus fortes.

Enfin, nous étudions dans la figure 4.8 l'impact du taux d'apprentissage, ξ, sur les performances de la méthode *QMIX*. En supposant le même système que sur la figure 4.3, l'augmentation du taux d'apprentissage a un impact positif sur le comportement d'apprentissage du *QMIX*, car il permet une performance de débit de données plus élevée à un petit nombre d'épisodes avant la convergence. Néanmoins, un taux d'apprentissage plus élevé tend à atteindre un optimum local plutôt que global, lorsque le taux d'apprentissage est élevé, ce qui explique la forte variance de performance (à la convergence) de la courbe bleue ($\xi = 5 \times 10^{-3}$), comparé au rouge ($\xi = 5 \times 10^{-4}$).

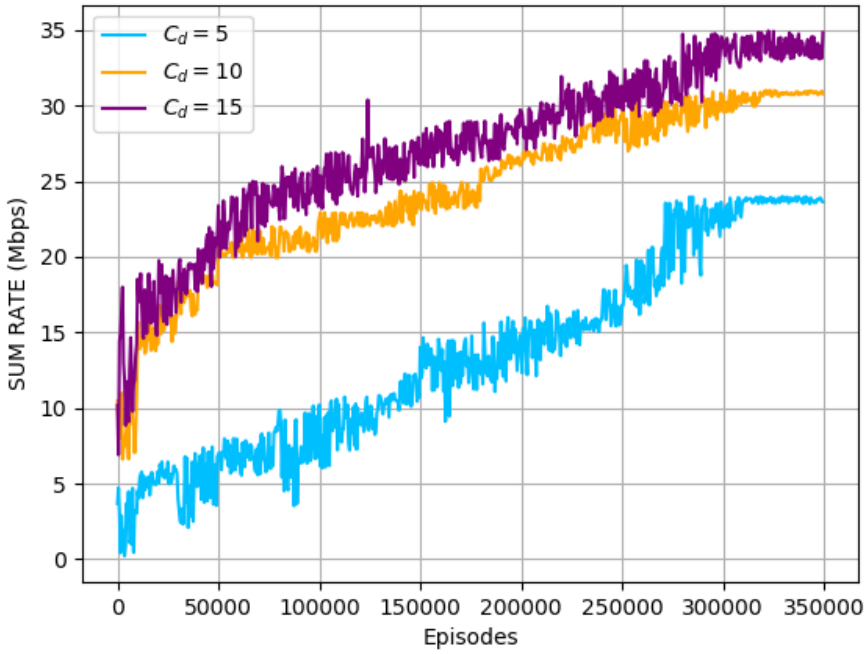

Figure 4.7 Débit total par rapport au nombre d'épisodes (différentes capacités de cache des appareils D2D).

4.6 Conclusion

Dans ce travail, nous avons étudié le problème d'appariement des appareils D2D dans les réseaux cellulaires sous-jacents activés pour le cache. En raison de la grande complexité du problème, nous avons proposé une approche d'apprentissage par renforcement multi-agents intelligente et distribuée basée sur *QMIX*. Les résultats obtenus démontrent la supériorité de notre approche par rapport aux algorithmes de base d'apprentissage par renforcement. Enfin, notre méthode est pratique pour un déploiement dans des systèmes à grande échelle, avec un grand nombre d'appareils D2D et de fichiers de contenu

Figure 4.8 Débit total par rapport au nombre d'épisodes (differents taux d'apprentissage).

CONCLUSION

La croissance explosive du nombre d'appareils mobiles et des applications Internet mobiles apporte un grand confort à la société. Il crée également une énorme demande de trafic qui remet en question la conception des systèmes de communication mobile de prochaine génération. Une partie importante de l'augmentation du trafic provient du téléchargement en double de fichiers vidéo populaires à partir de serveurs distants. Par conséquent, la mise en cache et le partage de contenu local peuvent réduire l'encombrement des réseaux mobiles.

Dans ce mémoire, nous avons étudié le problème d'appariement D2D dans les réseaux cellulaires sous-jacents activés pour le cache. En raison de la grande complexité du problème, nous avons proposé une approche d'apprentissage par renforcement multi-agents intelligente et distribuée basée sur QMIX. Les résultats obtenus démontrent la supériorité de notre approche par rapport aux méthodes basiques d'apprentissage par renforcement. Enfin, notre méthode est pratique pour un déploiement dans des systèmes à grande échelle, avec un grand nombre d'appareils D2D et de fichiers de contenu.

Nous envisageons dans nos travaux futurs de mettre en œuvre un scénario plus réaliste tenant compte de la mobilité des appareils *D2D*. Nous comptons aussi améliorer la complexité du système, en mettant en oeuvre plusieurs CUEs au lieu d'un seul.

RÉFÉRENCES

Alkhateeb, A., Alex, S., Varkey, P., Li, Y., Qu, Q. et Tujkovic, D. (2018). Deep learning coordinated beamforming for highly-mobile millimeter wave systems. *IEEE Access*, *6*, 37328–37348.

Alkurd, R., Shubair, R. M. et Abualhaol, I. (2014). Survey on device-to-device communications : Challenges and design issues. Dans *2014 IEEE 12th International New Circuits and Systems Conference (NEWCAS)*, 361–364.

Andreev, S., Pyattaev, A., Johnsson, K., Galinina, O. et Koucheryavy, Y. (2014). Cellular traffic offloading onto network-assisted device-to-device connections. *IEEE Communications Magazine*, *52*(4), 20–31.

Ansari, R. I., Chrysostomou, C., Hassan, S. A., Guizani, M., Mumtaz, S., Rodriguez, J. et Rodrigues, J. J. P. C. (2018). 5g d2d networks : Techniques, challenges, and future prospects. *IEEE Systems Journal*, *12*(4), 3970–3984.

Asadi, A., Wang, Q. et Mancuso, V. (2014). A survey on device-to-device communication in cellular networks. *IEEE Communications Surveys Tutorials*, *16*(4), 1801–1819.

Astely, D., Dahlman, E., Fodor, G., Parkvall, S. et Sachs, J. (2013). Lte release 12 and beyond [accepted from open call]. *IEEE Communications Magazine*, *51*(7), 154–160.

Azmat, F., Chen, Y. et Stocks, N. (2016). Predictive modelling of rf energy for wireless powered communications. *IEEE Communications Letters*, *20*(1), 173–176.

Bastien, L. (2018). Machine learning et big data : définition et explications. Récupéré le 2020-02-01 de `https://www.lebigdata.fr/machine-learning-et-big-data`

Bastug, E., Bennis, M. et Debbah, M. (2014). Living on the edge : The role of proactive caching in 5g wireless networks. *IEEE Communications Magazine*, *52*(8), 82–89.

Bin Zhou, Saisai Ma, Jing Xu et Zhenhong Li (2013). Group-wise channel sensing and resource pre-allocation for lte d2d on ism band. Dans *2013 IEEE Wireless Communications and Networking Conference (WCNC)*, 118–122.

Cheng, W., Zhang, X. et Zhang, H. (2016). Optimal power allocation with statistical qos provisioning for d2d and cellular communications over underlaying wireless networks. *IEEE Journal on Selected Areas in Communications*, *34*(1), 151–162.

Choudhary, A. (2019). A hands-on introduction to deep q-learning using openai gym in python. Récupéré le 2019-04-18 de `https://www.analyticsvidhya.com/blog/2019/04/introduction-deep-q-learning-python/`

Cisco (2020). Cisco annual internet report (2018–2023) white paper. Récupéré le 2020-03-09 de `https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html`

dan lee (2019). Reinforcement learning, part 3 : The markov decision process. Récupéré le 2019-10-30 de `https://medium.com/ai\%C2\%B3-theory-practice-business/reinforcement-learning-part-3-the-markov-decision-process-9f5066e073a2`

DesJardins, M. (2005). Multi-agent systems : Overview and research directions. Récupéré le 2005-03-15 de `https://slideplayer.com/slide/8480095/`

Fodor, G., Dahlman, E., Mildh, G., Parkvall, S., Reider, N., Miklós, G. et Turányi, Z. (2012). Design aspects of network assisted device-to-device communications. *IEEE Communications Magazine*, *50*(3), 170–177.

Gandotra, P., Jha], R. K. et Jain, S. (2017). A survey on device-to-device (d2d) communication : Architecture and security issues. *Journal of Network and Computer Applications*, *78*, 9 – 29. `http://dx.doi.org/https://doi.org/10.1016/j.jnca.2016.11.002`. Récupéré de `http://www.sciencedirect.com/science/article/pii/S1084804516302727`

Gazda, J., ŠLapak, E., Bugár, G., Horváth, D., Maksymyuk, T. et Jo, M. (2018). Unsupervised learning algorithm for intelligent coverage planning and performance optimization of multitier heterogeneous network. *IEEE Access*, *6*, 39807–39819.

Gengtian, S., Koshimizu, T., Saito, M., Zhenni, P., Jiang, L. et Shimamoto,

S. (2020). Power control based on multi-agent deep q network for d2d communication. Dans *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 257–261.

Girmay, G. G., Pham, Q. et Hwang, W. (2019). Joint channel and power allocation for device-to-device communication on licensed and unlicensed band. *IEEE Access*, *7*, 22196–22205.

Gregori, M., Gómez-Vilardebó, J., Matamoros, J. et Gündüz, D. (2016). Wireless content caching for small cell and d2d networks. *IEEE Journal on Selected Areas in Communications*, *34*(5), 1222–1234.

Han, G., Xiao, L. et Poor, H. V. (2017). Two-dimensional anti-jamming communication based on deep reinforcement learning. Dans *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2087–2091.

Hayat, O., Ngah, R. et Zahedi, Y. (2019). In-band device to device (d2d) communication and device discovery : A survey. *Wireless Personal Communications*, *106*(2), 451 – 472. Récupéré de `http://search.ebscohost.com.proxy.bibliotheques.uqam.ca/login.aspx?direct=true&db=iih&AN=136015793&lang=fr&site=ehost-live`

Hoen, P., Tuyls, K., Panait, L., Luke, S. et Poutré, J. (2005). An overview of cooperative and competitive multiagent learning. 1–46.

Huang, Y., Tan, T., Wang, N., Chen, Y. et Li, Y. (2018). Resource allocation for d2d communications with a novel distributed q-learning algorithm in heterogeneous networks. Dans *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 2, 533–537.

Jaafar, W., Ajib, W. et Elbiaze, H. (2018). Joint caching and resource allocation in d2d-assisted heterogeneous networks. Dans *2018 14th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, 1–8.

Jaafar, W., Ajib, W. et Elbiaze, H. (2019). Caching optimization for d2d-assisted heterogeneous wireless networks. Dans *2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 1–6.

Jaafar, W., Ohtsuki, T., Ajib, W. et Haccoun, D. (2016). Impact of the csi on the performance of cognitive relay networks with partial relay selection. *IEEE Transactions on Vehicular Technology*, *65*(2), 673–684.

Jameel, F., Hamid, Z., Jabeen, F., Zeadally, S. et Javed, M. A. (2018). A survey of device-to-device communications : Research issues and challenges. *IEEE Communications Surveys Tutorials*, *20*(3), 2133–2168.

Jiang, W., Feng, G., Qin, S. et Yum, T. S. P. (2018). Efficient d2d content caching using multi-agent reinforcement learning. Dans *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 511–516.

Jiang, W., Feng, G., Qin, S., Yum, T. S. P. et Cao, G. (2019). Multi-agent reinforcement learning for efficient content caching in mobile d2d networks. *IEEE Transactions on Wireless Communications*, *18*(3), 1610–1622.

Klügel, M. et Kellerer, W. (2020). Optimal mode selection by cross-layer decomposition in d2d cellular networks. *IEEE Transactions on Wireless Communications*, *19*(4), 2528–2542.

Lee, J. H., Shin, J. et Realff, M. J. (2018). Machine learning : Overview of the recent progresses and implications for the process systems engineering field. *Computers Chemical Engineering*, *114*, 111 – 121. FOCAPO/CPC 2017, `http://dx.doi.org/https://doi.org/10.1016/j.compchemeng.2017.10.008`. Récupéré de `http://www.sciencedirect.com/science/article/pii/S0098135417303538`

Li, L., Zhao, G. et Blum, R. S. (2018a). A survey of caching techniques in cellular networks : Research issues and challenges in content placement and delivery strategies. *IEEE Communications Surveys Tutorials*, *20*(3), 1710–1732.

Li, X., Liu, H. et Wang, X. (2019). Solve the inverted pendulum problem base on dqn algorithm. Dans *2019 Chinese Control And Decision Conference (CCDC)*, 5115–5120.

Li, Y., Zhong, C., Gursoy, M. C. et Velipasalar, S. (2018b). Learning-based delay-aware caching in wireless d2d caching networks. *IEEE Access*, *6*, 77250–77264.

Li, Z. et Guo, C. (2020). Multi-agent deep reinforcement learning based spectrum allocation for d2d underlay communications. *IEEE Transactions on Vehicular Technology*, *69*(2), 1828–1840.

Lin, X., Andrews, J. G., Ghosh, A. et Ratasuk, R. (2014). An overview of 3gpp device-to-device proximity services. *IEEE Communications Magazine*, *52*(4), 40–48.

Liu, D., Chen, B., Yang, C. et Molisch, A. F. (2016). Caching at the wireless edge : design aspects, challenges, and future directions. *IEEE Communications Magazine*, *54*(9), 22–28.

Luo, Y., Shi, Z., Zhou, X., Liu, Q. et Yi, Q. (2014). Dynamic resource allocations based on q-learning for d2d communication in cellular networks. Dans *2014 11th International Computer Conference on Wavelet Actiev Media Technology and Information Processing(ICCWAMTIP)*, 385–388.

Ma, C., Ding, M., Chen, H., Lin, Z., Mao, G., Liang, Y. et Vucetic, B. (2018). Socially aware caching strategy in device-to-device communication networks. *IEEE Transactions on Vehicular Technology*, *67*(5), 4615–4629.

Morocho-Cayamcela, M. E., Lee, H. et Lim, W. (2019). Machine learning for 5g/b5g mobile and wireless communications : Potential, limitations, and future directions. *IEEE Access*, *7*, 137184–137206.

Moussaid, A., Jaafar, W., Ajib, W. et Elbiaze, H. (2018). Deep reinforcement learning-based data transmission for d2d communications. Dans *2018 14th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, 1–7.

Nie, S., Fan, Z., Zhao, M., Gu, X. et Zhang, L. (2016). Q-learning based power control algorithm for D2D communication. Dans *Proc. IEEE 27th Ann. Int. Symp. Pers., Ind., Mob. Radio Commun. (PIMRC)*, 1–6.

Noura, M. et Nordin, R. (2016). A survey on interference management for device-to-device (d2d) communication and its challenges in 5g networks. *Journal of Network and Computer Applications*, *71*. `http://dx.doi.org/10.1016/j.jnca.2016.04.021`

Oliehoek, F. et Amato, C. (2016). A concise introduction to decentralized POMDPs. *Springer*. `http://dx.doi.org/10.1007/978-3-319-28929-8`

Oracle (2020). Oracle artificial intelligence (ai)—what is machine learning ? Récupéré le 2020-02-01 de `https://www.oracle.com/artificial-intelligence/what-is-machine-learning.html`

Oroojlooy jadid, A. et Hajinezhad, D. (2019). A review of cooperative multi-agent deep reinforcement learning.

Otterlo, M. et Wiering, M. (2012). Reinforcement learning and markov decision processes. *Reinforcement Learning : State of the Art*, 3–42. `http://dx.doi.org/10.1007/978-3-642-27645-3_1`

Parwez, M. S., Rawat, D. B. et Garuba, M. (2017). Big data analytics for

user-activity analysis and user-anomaly detection in mobile wireless network. *IEEE Transactions on Industrial Informatics*, *13*(4), 2058–2065.

Prerna, D., Tekchandani, R. et Kumar, N. (2020). Device-to-device content caching techniques in 5g : A taxonomy, solutions, and challenges. *Computer Communications*, *153*, 48 – 84. http://dx.doi.org/https://doi.org/10.1016/j.comcom.2020.01.057. Récupéré de http://www.sciencedirect.com/science/article/pii/S0140366419318225

Qiang, W. et Zhongli, Z. (2011). Reinforcement learning model, algorithms and its application. Dans *2011 International Conference on Mechatronic Science, Electric Engineering and Computer (MEC)*, 1143–1146.

Rashid, T., Samvelyan, M., de Witt, C. S., Farquhar, G., Foerster, J. et Whiteson, S. (2018). QMIX : Monotonic value function factorisation for deep multi-agent reinforcement learning. Récupéré de arxiv.org/abs/1803.11485

Sadeghi, A., Sheikholeslami, F. et Giannakis, G. B. (2018). Optimal and scalable caching for 5g using reinforcement learning of space-time popularities. *IEEE Journal of Selected Topics in Signal Processing*, *12*(1), 180–190.

Samvelyan, M., Rashid, T., de Witt, C. S., Farquhar, G., Nardelli, N., Rudner, T. G. J., Hung, C.-M., Torr, P. H. S., Foerster, J. et Whiteson, S. (2019). The StarCraft Multi-Agent Challenge. Récupéré de arxiv.org/abs/1902.04043

Sanchez-Fernandez, M., de-Prado-Cumplido, M., Arenas-Garcia, J. et Perez-Cruz, F. (2004). Svm multiregression for nonlinear channel estimation in multiple-input multiple-output systems. *IEEE Transactions on Signal Processing*, *52*(8), 2298–2307.

Shen, X. (2015). Device-to-device communication in 5g cellular networks. *IEEE Network*, *29*(2), 2–3.

Song, L., Cheng, X., Chen, M., Zhang, S. et Zhang, Y. (2016). Coordinated device-to-device local area networks : the approach of the china 973 project d2d-lan. *IEEE Network*, *30*(1), 92–99.

Song, W., Zhao, Y. et Zhuang, W. (2018). Stable device pairing for collaborative data dissemination with device-to-device communications. *IEEE Internet of Things Journal*, *5*(2), 1251–1264.

Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V., Jaderberg, M., Lanctot, M., Sonnerat, N., Leibo, J. Z., Tuyls, K. et Graepel,

T. (2018). Value-decomposition networks for cooperative multi-agent learning based on team reward. Dans *Proc. 17th Int. Conf. Auton. Agents and MultiAgent Syst.*, AAMAS '18, p. 2085–2087., Richland, SC. Int. Foundation for Auton. Agents and Multiagent Syst.

Tan, M. (1993). Multi-agent reinforcement learning : Independent vs. cooperative agents. Dans *Proc. 10th Int. Conf. Mach. Learn.*, 330–337. Morgan Kaufmann.

Tehrani, M. N., Uysal, M. et Yanikomeroglu, H. (2014). Device-to-device communication in 5g cellular networks : challenges, solutions, and future directions. *IEEE Communications Magazine*, *52*(5), 86–92.

Violante, A. (2019). Simple reinforcement learning : Q-learning. Récupéré le 2019-03-18 de `https://towardsdatascience.com/simple-reinforcement-learning-q-learning-fcddc4b6fe56`

Vitale, C., Mancuso, V. et Rizzo, G. (2015). Modelling d2d communications in cellular access networks via coupled processors. Dans *2015 7th International Conference on Communication Systems and Networks (COMSNETS)*, 1–8.

Wang, Y., Tao, X., Zhang, X. et Gu, Y. (2017). Cooperative caching placement in cache-enabled d2d underlaid cellular network. *IEEE Communications Letters*, *21*(5), 1151–1154.

Wikipédia (2020a). Agrégateur — wikipédia, l'encyclopédie libre. Récupéré le 2020-02-01 de `http://fr.wikipedia.org/w/index.php?title=Agr\%C3\%A9gateur&oldid=166967965`

Wikipédia (2020b). Bande industrielle, scientifique et médicale. Récupéré le 2020-04-10 de `https://en.wikipedia.org/wiki/ISM_band`

Wikipédia (2020c). Réseau de neurones artificiels. Récupéré le 2020-06-23 de `https://fr.wikipedia.org/wiki/R%C3%A9seau_de_neurones_artif`

Wikipédia (2020d). Équilibre de nash. Récupéré le 2020-05-19 de `https://fr.wikipedia.org/wiki/%C3%89quilibre_de_Nash`

Wu, D., Zhou, L., Cai, Y. et Qian, Y. (2018). Collaborative caching and matching for d2d content sharing. *IEEE Wireless Communications*, *25*(3), 43–49.

Xu, J., Gu, X. et Fan, Z. (2018). D2d power control based on hierarchical extreme learning machine. Dans *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications*

(PIMRC), 1–7.

Yang, C., Yao, Y., Chen, Z. et Xia, B. (2016a). Analysis on cache-enabled wireless heterogeneous networks. *IEEE Transactions on Wireless Communications*, *15*(1), 131–145.

Yang, C., Zhao, X., Yao, Y. et Xia, B. (2016b). Modeling and analysis for cache-enabled cognitive d2d communications in cellular networks. Dans *2016 IEEE Global Communications Conference (GLOBECOM)*, 1–6.

Yin, J., Li, L., Xu, Y., Liang, W., Zhang, H. et Han, Z. (2018). Joint content popularity prediction and content delivery policy for cache-enabled d2d networks : A deep reinforcement learning approach. Dans *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 609–613.

Zhang, G., Yang, K., Liu, P. et Wei, J. (2015). Power allocation for full-duplex relaying-based d2d communication underlaying cellular networks. *IEEE Transactions on Vehicular Technology*, *64*(10), 4911–4916.

Zhang, K., Leng, S., He, Y., Maharjan, S. et Zhang, Y. (2018). Cooperative content caching in 5g networks with mobile edge computing. *IEEE Wireless Communications*, *25*(3), 80–87.

Zhao, M., Wei, Y., Song, M. et Da, G. (2018). Power control for d2d communication using multi-agent reinforcement learning. Dans *2018 IEEE/CIC International Conference on Communications in China (ICCC)*, 563–567.